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

The Mathematical Centre, founded 11th 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 148

LINEAR PROGRAMMING AND FINITE MARKOVIAN CONTROL PROBLEMS

L.C.M. KALLENBERG

MATHEMATISCH CENTRUM AMSTERDAM 1983

1980 Mathematics subject classification: 90C40

ISBN 90 6196 236 6

ACKNOWLEDGMENTS

This monograph is a slightly revised version of my thesis which was written under the supervision of professor A. Hordijk. I am highly grateful to him not only for introducing me in the field of Markovian decision theory but especially for the guidance and stimulating discussions during the course of the investigations. Without his cooperation this book would not have been written.

Also, I am independent to professor G. Zoutendijk, who aroused my interest for linear programming, and to professor J. Wessels. They read the manuscript and made a number of helpful comments. Furthermore, I would like to thank P.J. van den Berg for the computer programming.

I thank the Mathematical Centre for the opportunity to publish this monograph in the series Mathematical Centre Tracts and all those at the Mathematical Centre who have contributed to its technical realization.

CONTENTS

INTRODUCTION

In this monograph we study *Markovian control problems*. These problems concern the control of systems which have a dynamic structure, i.e. decisions have to be made at different points in time. If a decision is made, then the behaviour of the system is uncertain, i.e. the state of the system at the next decison time point is not deterministic, but is given by a probability distribution on the state space.

An example of such a control problem is the following (cf. ROSS [1970] pp.138-139). Suppose a person wants to sell his house and an offer is made every week. The seller has two possible decisions: to reject or to accept the offer. If he rejects the offer, then the offer is no longer available and the offer of the next week is uncertain. Furthermore, a maintenance cost is incurred for each week that the house remains unsold. Which policy has to be chosen to obtain the maximum expected profit?

In this monograph, we shall pay special attention to the construction of *algorithms*, based on *linear programming*, to compute optimal policies for several optimality criteria. We will discuss *finite* Markov decision problems, semi-Markov problems and stochastic games.

Markov decision problems can be characterized by a state space, an action space, transition probabilities, rewards and a utility function. The system is observed at discrete time points to be in one state of the state space. Then the decision maker chooses an action from the action space and two things occur:

(1) a reward is earned,

(2) the next state of the system is chosen according to a probability distribution on the state space.

If the decision maker uses a stationary policy, i.e. the chosen action only depends on the state of the system, then the sequence of states form a Markov chain. For this reason the problem is called a Markov decision problem. Markov decision models were introduced by BELLMAN [1957] and HOWARD [1960]. At this moment, there is an extensive literature on this subject and there are several books which deal with Markov decision problems, e.g. DERMAN [1970], ROSS [1970], MINE & OSAKI [1970], HINDERER [1970] and HORDIJK [1974].

The *semi-Markov decision models* differ from the (discrete) Markov decision models by the fact that the times between the several decision points are random variables. Hence, if the decision maker uses a stationary policy, then the process $\{X(t), t \geq 0\}$, where $X(t)$ describes the state at time t, is a semi-Markov process. Semi-Markov decision models were introduced by DE CANI [1964], HOWARD [1963], JEWELL [1963] and SCHWEITZER [1965].

The third class of models that are studied, are the *stochastic games*. In a stochastic game several players control the system simultaneously. At any decision time point all players independently choose an action from their own action space. These choices produce a reward for every player, and the next state of the system is determined by a probability distribution which depends on the present state and the chosen actions. Stochastic games were introduced by SHAPLEY [1953], thus before the Markov decision model. If all players except one have only one action available in each state, then the stochastic game reduces to a Markov decision problem.

Methods to solve finite Markovian control problems are based on techniques such as policy improvement, successive approximation or linear programming.

The policy improvement method is an iterative procedure that computes a sequence of so-called pure and stationary policies such that subsequent policies give a higher value of the utility function. Since there exists a pure and stationary optimal policy and since the set of pure and stationary policies is finite, the procedure terminates after a finite number of iterations with an optimal pure and stationary policy.

The maximum value of the utility function satisfies a functional equation. By the method of successive approximation the solution of this equation is approximated, and corresponding policies are computed, using the well-known techniques on contraction mappings.

In this thesis we will discuss linear programming methods for the solution of several Markovian control problems.

The fact that linear programming can be used is based on the property that the maximal value of the utility function is the smallest so-called *superharmonic* vector. Since the superharmonic property is a condition formulated in terms of linear inequalities, we have to find the smallest element which satisfies a system of linear inequalities. Therefore, this maximal value can be found as the optimal solution of a linear program and an optimal policy may be obtained from the optimal solution of the dual program. It will be shown that the complementary slackness property plays an important role in proving the optimality properties. The concept of superharmonicity was introduced by HORDIJK [1974].

Already in 1960, linear programming formulations were known for some Markov decision models (cf. DE GHELLINCK [1960], D'EPENOUX [1960] and MANNE [1960]). We will prove similar results to several other Markovian control problems. For a short review we refer to HORDIJK & KALLENBERG [1981d]. The linear programming approach has some advantages in comparison with other techniques, e.g.

(1) In many industrial environments linear programming computer codes are available. Hence, linear programming algorithms can be made operational very easily.

(2) If we use linear programming, then we have the opportunity to apply sensitivity analysis on the optimal solution. Therefore, the decision maker may obtain information about the behaviour of the optimal policy when the data are changed.

(3) By linear programming we can solve Markovian control problems with additional constraints. As far as we know, linear programming is the only technique for the solution of this kind of problems.

In this thesis, we only discuss models with a finite state space and a finite action space. If we drop the finiteness, then linear programming formulations also may be obtained (e.g. HEILMANN [1977]). Since the emphasis of our work is on the construction of finite algorithms for the solution of Markovian control problems, we restrict ourselves to finite models.

The scope of the monograph is as follows. In the first two chapters we survey some basic results from the *theory of linear programming* (chapter 1) and from the *theory of Markov decision processes* (chapter 2).

In chapter 3 we consider Markov decision problems with the expected *total reward* as utility function. We introduce the concept of superharmonicity and we prove that the optimal utility vector - when we restrict the policies to the class of transient policies - is the smallest superharmonic vector. Hence, the linear programming approach is applicable. We present a linear programming formulation which gives a pure and stationary policy that is optimal in the class of transient policies. Also, the relation between stationary transient policies and the feasible solutions of the linear program is analysed. These results generalize the well-known linear programming method for *discounted* dynamic programming. Moreover, we discuss the Markov decision problem with *additional constraints* and we show that a stationary optimal transient policy can be found by the solution of a linear program. As special cases, we present the *optimal stopping* problem and the *contracting dynamic programming* problem. For the latter model, we prove that the linear programming method and the policy improvement method are equivalent and that the elimination of *suboptimal actions* can be implemented in the algorithm. In this chapter we also treat the *positive* and the *negative* dynamic programming models and, for both models, finite algorithms are derived for the determination of a pure and stationary optimal policy.

Chapter 4 deals with the expected *average reward* as utility function. Although we can present an approach similar to the previous chapter, the analysis of this model is more complex and we have to perform more calculations to obtain optimal policies. The concept of a superharmonic vector is introduced such that the optimal utility vector for the present criterion is the smallest superharmonic vector. A pure and stationary optimal policy can be obtained directly from an extreme optimal solution of a linear program. If we consider special models for which the Markov chains induced by stationary (optimal) policies are unichained, then the linear programs may be simplified considerably. It will be shown that there is a close relationship between the linear programming method and the policy improvement method. The determination of an optimal policy for the Markov decision model with *additional constraints* is complicated. We will construct an algorithm for the computation of a memoryless optimal policy. Although there exists no stationary optimal policy in general, fortunately, in many cases a stationary optimal policy may be found. We give sufficient conditions for its existence, and we present an algorithm for the computation of a stationary policy which is optimal when these conditions are satisfied. In the unichain case, a stationary optimal policy always exists and a simplified algorithm may be used.

Sometimes, a criterion that is more selective than the average reward criterion is preferable. In chapter 5, we discuss such a criterion. An optimal policy with respect to this criterion is a so-called *bias-optimal* policy. We present two algorithms for its computation. The first algorithm, which will be favourable when the number of average optimal pure and stationary policies is small, enumerates the extreme optimal solutions of the linear program used in chapter 4. For any optimal solution we have to perform additional computations to obtain the so-called bias-value. A policy which maximizes this bias-value is a bias-optimal policy. In the second algorithm, which is a modification of DENARDO [1970a], a pure and stationary bias-optimal policy is obtained by the solution of three linear programs and one search procedure, in the worst case. We also present a simplified algorithm for the unichain case.

In chapter 6 we consider a *two-person zero-sum stochastic game*. We only consider models in which the transition probabilities are controlled by one player (otherwise the linear programming approach is not possible). The total reward criterion (under a contraction assumption) and the average reward criterion will be treated analogously. We show that the value of the game is the smallest superharmonic vector which can be found as the optimal solution of a linear program. Stationary optimal policies for both players can be obtained from the optimal solution of the dual program. Moreover, the linear programming approach provides a new proof of the existence of the value of the game.

In the final chapter, the *semi-Markov decision model* is studied. Also for these models we can introduce a concept of superharmonicity which leads to a linear programming formulation. In the discounted reward case as well as in the average reward case we obtain pure and stationary optimal policies from the linear programming solution. We also show the equivalence with certain discrete Markov decision models. Hence, the results of the chapters 3 and 4 may also be applied on the semi-Markov decision model.

In this monograph Markovian control problems over an infinite horizon are considered. The linear programming approach is also applicable for finite horizon models (cf. KALLENBERG [1981a]).

CHAPTER 1

LINEAR PROGRAMMING

1.1. INTRODUCTION AND SUMMARY

In this chapter we shall present a survey of some basic results in the theory of linear programming. In the sequel of this monograph it will be shown that linear programming is a useful approach to derive finite algorithms for a number of Markovian control problems.

In section 1.2 we mention some properties of convex polyhedra. Convex polyhedra play an important role in the theory of linear programming. We present a theorem on separating hyperplanes and we give a characterization of the set of extreme points of a convex polyhedron.

Then, in section 1.3, the linear program is introduced and the well-known optimality and duality theorems are summarized, including the complementary slackness property. Optimality and duality properties will be a useful instrument for the proofs of the theorems in the following chapters.

Section 1.4 deals with the simplex method, developed by G.B. Dantzig in 1947. The simplex tableau is presented. Moreover, we derive an algorithm to compute all extreme optimal solutions of a linear program.

The theory of linear programming can be found in many text books. For the proofs of the theorems we refer to these books.

NOTATIONS 1.1.1.

(i) A (column) vector x with n components is denoted by $x = (x_1,x_2,\dots,x_n)^T$ or by $x = (x_i)$; a matrix A with (i,j)-th element a_{ij} is denoted by $A = (a_{ij})$; the k-th column and the i-th row of A are denoted by $a_{\cdot k}$ respectively $a_{i\cdot}$.

(ii) Let x and y be n-component vectors. Then $x \geq y$ denotes that $x_i \geq y_i$ for all i, $x > y$ means $x \geq y$ and $x \neq y$, $x >> y$ signifies that

$x_i > y_i$ for all i; we denote $x < \infty$ if and only if $x_i < \infty$ $i = 1,2,\dots,n$.

(iii) When the range of a variable is unspecified, then its entire range is intended, e.g. $\Sigma_i x_i = \Sigma_{i=1}^n x_i$ if $x = (x_1,x_2,\dots,x_n)^T$; the dimension of vectors and matrices is not always explicitly mentioned, but this dimension will be clear from the context.

(iv) $\mathbb{N}$ is the set of positive integers: $\mathbb{N} = \{1,2,\dots\}$; $\mathbb{N}_0$ is the set of nonnegative integers: $\mathbb{N}_0 = \{0,1,\dots\}$.

(v) $\mathbb{R}^n$ is the set of all real n-component vectors; $\mathbb{R}^+ = \{a \in \mathbb{R}^1 | a > 0\}$.

(vi) By $|E|$ we denote the cardinality of a set E.

(vii) $E \backslash F$ is the set of all elements of E which do not belong to F.

(viii) The notation $x := y$ will be used to indicate that the variable x gets the value y.

(ix) The symbol $\square$ indicates the end of a proof.

(x) For $a \in \mathbb{R}^1$ we denote by $\lfloor a \rfloor$ the largest integer not greater than a.

DEFINITIONS 1.1.1.

(i) The *null-vector*, denoted by 0, has all components zero; the *null matrix*, also denoted by 0, has all elements zero; the *identity matrix*, denoted by I, has elements (δ_{ij}), where δ_{ij} is Kronecker's delta; the j-th *unit vector*, notated by e_j, is the vector with all entries zero except entry j, which is a one; $e := (1,1,\dots,1)^T$ is the *sum vector*.

(ii) The *inner product* of two real n-component vectors x and y is denoted by $x^T y$ and defined by $x^T y := \Sigma_i x_i \cdot y_i$.

(iii) The (supremum) *norm* of $x \in \mathbb{R}^n$ is defined by $\|x\| := \max_{1 \le i \le n} |x_i|$; the (supremum) norm of a matrix A is defined by $\|A\| := \sup_{\|x\|=1} \|Ax\|$. (It can easily be verified that $\|A\| = \max_i \Sigma_j |a_{ij}|$).

1.2. CONVEX POLYHEDRA

In this section we review some results about convex polyhedra that are fundamental for the theory of linear programming.

DEFINITIONS 1.2.1.

(i) $S \subset \mathbb{R}^n$ is a *convex set* if for any two vectors $x,y \in S$ and any $\lambda \in (0,1)$ $\lambda x + (1-\lambda)\, y \in S$; the *convex hull* of a set $S \subset \mathbb{R}^n$ is the intersection of all convex sets containing S as subset; the *closed convex hull* of S is the smallest closed convex set containing S as subset (this closed convex hull will be denoted by $\bar{S}$).

(ii) A *face* of a convex set S is a convex subset S' of S such that every closed line segment in S with a relative interior point in S' has both endpoints in S'. The zero-dimensional faces of S are called the *extreme points* of S. (Then $x \in S$ is an extreme point of S if and only if there do not exist points $y,z \in S$ distinct from x for which $x = \lambda y+(1-\lambda)z$ for some $\lambda \in (0,1)$). If S' is a half-line face of S, then we call the direction of S' an *extreme direction* of S.

(iii) A *convex polyhedron* R is the intersection of a finite number of closed half-spaces, i.e. $R = \{x|Ax \le b\}$ for some m×n matrix A and some vector $b \in \mathbb{R}^m$. A bounded convex polyhedron is a *polytope*.

(iv) $C \subset \mathbb{R}^n$ is a *cone* if for any $x \in C$, $\lambda x \in C$ for every $\lambda \ge 0$; a *convex polyhedral cone* generated by the m×n matrix A is the set $\{y|y = A^T u;\ u \ge 0\}$. The vectors $(a_{i\cdot})^T$ are the *extreme rays* of the cone; the *dual cone* C^* is defined by $C^* := \{y|y^T x \le 0$ for every $x \in C\}$.

THEOREM 1.2.1. *Let* $S \subset \mathbb{R}^n$ *be any closed convex set and suppose that* $x \notin S$. *Then there exists a vector* $r \in \mathbb{R}^n$ *and a real number* r_0 *such that*

$$r^T x > r_0 > r^T y \quad \text{for every } y \in S.$$

PROOF. See KARLIN [1959] pp.397-398. □

Consider the set $R := \{x|Ax = b;\ x \ge 0\}$, where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and A an m×n matrix. Since each equality may be replaced by two inequalities, R is a convex polyhedron.

THEOREM 1.2.2. $x \in R$ *is an extreme point if and only if* $\{a_{\cdot k}|x_k > 0\}$ *is a linearly independent set of vectors.*

PROOF. See COLLATZ & WETTERLING [1966] pp.9-10. □

LEMMA 1.2.1. *The number of extreme points and extreme directions of* R *is finite.*

PROOF. See ROCKAFELLAR [1970] pp. 170-172. □

LEMMA 1.2.2. *If* R *is non-empty, then also the set of extreme points of* R *is non-empty.*

PROOF. See COLLATZ & WETTERLING [1966] pp.10-11. □

THEOREM 1.2.3. *If* R *is non-empty with extreme points* $\{x^k\}_{k=1}^K$ *and extreme directions* $\{s^\ell\}_{\ell=1}^L$, *then any* $x \in R$ *can be written as*

$$x = \sum_{k=1}^K \lambda_k x^k + \sum_{\ell=1}^L \mu_\ell s^\ell,$$

where $\lambda_k \geq 0$ $k = 1,2,\ldots,K$, $\Sigma_{k=1}^K \lambda_k = 1$ *and* $\mu_\ell \geq 0$ $\ell = 1,2,\ldots,L$.

PROOF. See ROCKAFELLAR [1970] pp.170-172. □

COROLLARY 1.2.2. *Any polytope is the convex hull of its extreme points.*

1.3. OPTIMALITY AND DUALITY

DEFINITIONS 1.3.1. The *linear programming problem* is the problem of finding a vector $x \in \mathbb{R}^n$ which maximizes a linear form p^Tx (called the *objective function*), subject to the linear *constraints* $Ax \leq b$, $x \geq 0$, where $b \in \mathbb{R}^m$ and A is an m×n matrix. This problem is usually notated by

$$(1.3.1) \qquad max\{p^Tx \mid Ax \leq b;\ x \geq 0\}.$$

A linear programming problem is also called a *linear program*. The convex polyhedron $R := \{x|Ax \leq b;\ x \geq 0\}$ is said to be the *feasible region*. Any $x \in R$ is called a *feasible solution*. For any $x \in R$ we define $y := b-Ax$; then $y \in \mathbb{R}^m$ and $y \geq 0$. Furthermore, we introduce

$$\bar{A} := (A,I), \quad \bar{x} := \binom{x}{y} \quad \text{and} \quad \bar{p} := \binom{p}{0}.$$

Then we can write the linear program (1.3.1) as

$$(1.3.2) \qquad max\{\bar{p}^T\bar{x} \mid \bar{A}\bar{x} = b;\ \bar{x} \geq 0\}$$

with feasible region $\bar{R} := \{\bar{x}|\bar{A}\bar{x} = b;\ \bar{x} \geq 0\}$.

A similar formulation is

$$(1.3.3) \qquad max\{p^Tx \mid Ax + y = b;\ x \geq 0,\ y \geq 0\}.$$

THEOREM 1.3.1. *x is an extreme point of R if and only if $\bar{x}$ is an extreme point of $\bar{R}$.*

PROOF. The proof is straightforward. □

DEFINITIONS 1.3.2. Given a linear programming problem, there are three possibilities:

1. There is no feasible solution. In this case the problem is said to be *infeasible*.
2. There is a feasible solution x° with $p^T x^{\circ} \geq p^T x$ for every $x \in R$. Then x° is called an *optimal solution* and we say that the linear program has a *finite solution*.
3. There is a feasible solution $x^{\circ} \in R$ and a vector $s^{\circ} \in \mathbb{R}^n$ such that $p^T s^{\circ} > 0$ and $x^{\circ} + \lambda s^{\circ} \in R$ for all $\lambda \geq 0$. Then the objective function can be made arbitrarily large and the problem is said to be *unbounded* or has an *infinite solution*. The vector s° is called an *infinite direction in* x°.

THEOREM 1.3.2. *If the linear program has a finite solution, then it has an optimal extreme solution.*

PROOF. See COLLATZ & WETTERLING [1966] pp.12-13. □

LEMMA 1.3.1. *The set of optimal solutions is convex.*

PROOF. See COLLATZ & WETTERLING [1966] p.11. □

DEFINITIONS 1.3.3. A vector $s \in \mathbb{R}^n$ is said to be a *feasible direction* in a point $x \in R$ if there exists a $\lambda > 0$ such that $x + \lambda s \in R$. If, in addition, $p^T s > 0$ then s is said to be a *usable direction*. For any $x \in R$ we define $M(x) := \{i \mid a_{i\cdot}^T x = b_i\}$, $N(x) := \{j \mid (e_j)^T x = 0\}$ and

$$S(x) := \left\{ s \in \mathbb{R}^n \;\middle|\; \begin{array}{lll} a_{i\cdot}^T s & \leq 0, & i \in M(x) \\ & & \\ (-e_j)^T s & \leq 0, & j \in N(x) \end{array} \right\}.$$

$S(x)$ is the *cone of feasible directions* in x.

THEOREM 1.3.3. *(Optimality theorem) $x \in R$ is an optimal solution of the linear program* (1.3.1) *if and only if there exist vectors $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ such that $p = A^T u - v$, $u \geq 0$, $v \geq 0$ and $u^T(b-Ax) = v^T x = 0$.*

PROOF. See ZOUTENDIJK [1976] pp.23-24. □

REMARK 1.3.1. Suppose that x is an optimal solution of the linear program (1.3.1). Then from the convexity of the set of optimal solutions (see lemma 1.3.1) it follows that x is the unique optimal solution if and only if $p^T s < 0$ for all $s \in S(x)$. Hence, x is unique if and only if p is an interior point of the dual cone of cone S(x).

DEFINITIONS 1.3.4. We define for the linear programming problem (1.3.1) the *dual problem* by

$$(1.3.4) \qquad \min\{b^T u \mid A^T u \geq p,\ u \geq 0\}$$

with feasible region $D := \{u \mid A^T u \geq p;\ u \geq 0\}$. Defining the vector v by $v := A^T u - p$, the dual problem can also be written as

$$(1.3.5) \qquad \min\{b^T u \mid A^T u - v = p;\ u \geq 0;\ v \geq 0\}.$$

Problem (1.3.1) is said to be the *primal problem*.

THEOREM 1.3.4. *(Duality theorem)*

(i) *The dual problem of the dual problem is the primal problem.*

(ii) *If* $x \in R$ *and* $u \in D$, *then* $p^T x \leq b^T u$.

(iii) *If the primal problem has an optimal solution* x°, *then the dual problem has also a finite optimal solution, say* u°. *Moreover,*

$$p^T x^\circ = b^T u^\circ, \qquad (u^\circ)^T (b - Ax^\circ) = 0 \quad \textit{and} \quad (x^\circ)^T (A^T u^\circ - p) = 0.$$

(iv) *If* $x \in R$ *and* $u \in D$ *satisfy* $u^T(b-Ax) = x^T(A^T u - p) = 0$, *then* x *and* u *are optimal solutions of the primal and the dual problem respectively.*

(v) *If the primal problem has an infinite solution, then the dual problem is infeasible.*

(vi) *If the primal problem is infeasible, then the dual problem either has an infinite solution or it is infeasible.*

PROOF. See ZOUTENDIJK [1976] pp.24-26. □

COROLLARY 1.3.1. *(Complementary slackness property) Suppose that* (x,y) *and* (u,v) *are optimal solutions of the programs* (1.3.3) *and* (1.3.5) *respectively. Then*

(i) $x_j > 0 \Rightarrow v_j = 0$.
(ii) $y_i > 0 \Rightarrow u_i = 0$.
(iii) $u_i > 0 \Rightarrow y_i = 0$.
(iv) $v_j > 0 \Rightarrow x_j = 0$.

1.4. SIMPLEX METHOD

Consider the linear programming problem formulated as (1.3.2). Assume that the columns of $\bar{A}$ are rearranged such that $\bar{A} = (B,N)$, where B is an m×m nonsingular matrix. Let $\bar{x} = (x_B,x_N)^T$, where x_B is the vector of variables corresponding to the columns of B, and x_N is the vector of variables that correspond to the columns of N. Then, $Ax = b$ can be written as $Bx_B + Nx_N = b$. Since B is nonsingular, the inverse matrix B^{-1} exists and we obtain $x_B = B^{-1}b - B^{-1}Nx_N$. Assume, in addition, that $B^{-1}b \geq 0$. Then, by theorem 1.2.2, the solution $x_B = B^{-1}b$, $x_N = 0$ is an extreme point of the feasible region $\bar{R}$. We say that the matrix B is a *basis matrix* and that $(x_B,x_N)^T$ is a *basic solution,* where x_B are the *basic variables* and x_N the *nonbasic variables*. The corresponding value x_0 of the objective function satisfies

$$(1.4.1) \qquad x_0 = p^Tx = p_B^Tx_B + p_N^Tx_N = p_B^TB^{-1}b + (p_N^T - p_B^TB^{-1}N)x_N.$$

We define the (n+m)-component vector $d = (d_B,d_N)^T$ by $d_B := 0$ and $d_N := p_B^TB^{-1}N-p_N^T$. The vector d may also be partitioned into parts corresponding to the original vectors y and x: $d = (u,v)^T$, where $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$.

<u>THEOREM 1.4.1</u>. *The vectors* u *and* v, *defined above, satisfy*

$$A^Tu - v = p; \qquad u^Ty = v^Tx = 0.$$

<u>PROOF</u>. See ZOUTENDIJK [1976] pp.36-37. □

<u>REMARK 1.4.1</u>. Theorem 1.4.1 implies that if $d \geq 0$, then (u,v) is a feasible solution of the dual program (1.3.5). Therefore, d is called the *vector of dual variables*. Moreover, theorem 1.3.4(iv) implies that $\bar{x}$ and (u,v) are optimal solutions of the primal and dual linear program respectively.

REMARK 1.4.2. In the *simplex method* basic solutions are iteratively computed such that the value of the objective function in subsequent iterations never decreases. To be sure that the simplex method is finite, it is sufficient to prove that a basis matrix cannot return. If $B^{-1}b \gg 0$ for every basis matrix B, then the value of the objective function increases at each iteration. Problems which have this property are said to be *nondegenerated*. Hence, the simplex method is finite for nondegenerated problems. For degenerated problems we need sophisticated rules to determine different basis matrices in subsequent iterations. A very elegant rule has been developed by BLAND [1977]. For the details of the simplex method, including its numerical aspects, we refer the reader to the chapters 3 and 4 in ZOUTENDIJK [1976].

For the computation of an optimal solution by the simplex method we use the so-called *simplex tableau*. In this tableau we store the basic and nonbasic variables but also the dual variables. This tableau has the following form

(1.4.2)

		x_N
x_B	$B^{-1}b$	$B^{-1}N$
x_0	$p_B^TB^{-1}b$	$d_N^T = p_B^TB^{-1}N-p_N^T$

REMARK 1.4.3. We have assumed that the columns of $\bar{A}$ can be rearranged such that $\bar{A} = (B,N)$, where B is a nonsingular matrix satisfying $B^{-1}b \geq 0$. In general, such a partition is not possible; moreover, if a partition is possible, then we don't know which columns can be chosen to form a regular basis matrix. Fortunately, by adding some artificial variables, we can overcome this difficulty if we apply the so-called *phase I - phase II* simplex method. Therefore, we partition the contraints of the linear programming problem in three subsets:

$$\textstyle\sum_j a_{ij}x_j \leq b_i \quad \text{and} \quad b_i \geq 0\text{: } I_1,$$

$$\textstyle\sum_j a_{ij}x_j \leq b_i \quad \text{and} \quad b_i < 0\text{: } I_2,$$

$$\sum_j a_{ij}x_j = b_i \qquad\qquad : I_3,$$

(we may assume that $b_i \geq 0$, $i \in I_3$, because otherwise the equality can be multiplied by -1). Introducing nonnegative *slack variables* y_i, $i \in I_1 \cup I_2$, and *artificial variables* z_i, $i \in I_2 \cup I_3$, we consider the linear program

$$(1.4.3) \qquad max\left\{ -\sum_i z_i \left| \begin{array}{lll} \sum_j a_{ij}x_j + y_i & = b_i & i \in I_1;\ x_j \geq 0 \quad j = 1,2,\ldots,n \\ -\sum_j a_{ij}x_j - y_i + z_i & = -b_i & i \in I_2;\ y_i \geq 0 \quad i \in I_1 \cup I_2 \\ \sum_j a_{ij}x_j + z_i & = b_i & i \in I_3;\ z_i \geq 0 \quad i \in I_2 \cup I_3 \end{array} \right. \right\}.$$

Then, we can start taking as basis matrix the identity matrix corresponding to the columns of y_i, $i \in I_1$, and z_i, $i \in I_2 \cup I_3$. This matrix satisfies the assumptions and we can apply the simplex method in order to obtain an optimal solution of (1.4.3). This is called the phase I. Suppose that $(x^\circ, y^\circ, z^\circ)$ is an optimal solution of (1.4.3).

If $\Sigma_i z_i^\circ > 0$, then the original problem is infeasible.

If $\Sigma_i z_i^\circ = 0$, then we have a feasible solution (x°, y°).

In the latter case, we take as new objective function the original objective function $\Sigma_j p_j x_j$ and continue the simplex method, maintaining $\Sigma_i z_i = 0$, to obtain an optimal solution for the original problem. This is called the phase II.

It may occur that the linear programming problem has an infinite solution. Then, we shall obtain a simplex tableau with a nonpositive column corresponding to a nonbasic variable, say $(x_N)_k$, such that $(d_N)_k < 0$. Define the direction vector s by

$$(1.4.4) \qquad \begin{cases} s_B := (-B^{-1}N)_{\cdot k} \\ s_N := e_k. \end{cases}$$

Then, we have

$$s \geq 0$$

$$\bar{A}s = Bs_B + Ns_N = -N_{\cdot k} + N_{\cdot k} = 0$$

$$\bar{p}^T s = p_B^T s_B + p_N^T s_N = (-p_B^T B^{-1} N + p_N^T)_k = -(d_N)_k > 0.$$

Consequently, s is an infinite direction.

We close this section with a discussion about the problem of finding *all optimal basic solutions* of a linear program. Suppose that the optimal simplex tableau (we assume that the linear program has a finite solution) is given by

$$(1.4.5)\qquad \begin{cases} (x_B)_i = b_i^* - \sum_j a_{ij}^* (x_N)_j & i = 1,2,\dots,m \\ x_0 = b_0^* - \sum_j (d_N)_j (x_N)_j \end{cases}$$

Since b_0^* is the optimal value and all variables are nonnegative, it follows from (1.4.5) that any optimal solution x satisfies

(i) $(x_N)_k = 0$ if $(d_N)_k > 0$,

(ii) $(x_N)_k = 0$ if for some i $b_i^* = 0$, $a_{i\cdot}^* \geq 0$ and $a_{ik}^* > 0$.

If we know that $(x_N)_k = 0$ for any optimal solution, then we may remove the corresponding column from the tableau; after this reduction we have $(d_N)_k = 0$ for every k. Hence, we may apply the following rule:

(iii) Every variable $(x_N)_k$ may enter the basis to obtain an optimal solution with a new basis matrix.

If $b_i^* = 0$ and $a_{i\cdot}^* = 0$, then we can remove this row from the tableau. Hence, we obtain a tableau with $d_N = 0$ and with in any row i where $b_i^* = 0$ at least one negative coefficient.

The optimal simplex tableau may contain artificial variables as basis variables. These variables can be removed from the tableau in the following way. Suppose that $(x_B)_i$ is an artificial variable, say z_ℓ. Then $b_i^* = 0$ and consequently there exists an index k such that $a_{ik}^* < 0$. Exchange the variables $(x_N)_k$ and z_ℓ by pivoting with pivot element a_{ik}^*. The variable z_ℓ becomes nonbasic and the corresponding column can be removed.

Mostly, we can simplify the tableau considerably by the rules stated above. In the reduced tableau, we may apply rule (iii) and the following rule (iv) in order to determine all optimal extreme solutions.

(iv) If $b_i^* = 0$ and $a_{ik}^* \neq 0$, then the variables $(x_N)_k$ and $(x_B)_i$ can be exchanged and an optimal solution with a new basis matrix is obtained.

Since the set of optimal solutions is convex, we can compute all extreme optimal solutions by successive computation of all extreme optimal solutions that are adjacent to the present extreme optimal solution (cf. HADLEY [1962] pp.166-168). This computation is elaborated in the following algorithm:

Algorithm I for the computation of all extreme optimal solutions of a linear program.

<u>step 1</u>: Determine an optimal solution by the simplex method and denote the coefficients of the optimal simplex tableau by (b_i^*), (a_{ij}^*) and $((d_N)_j)$.

<u>step 2</u>: If $(d_N)_j > 0$ for all j, then the optimal solution is unique (STOP).

<u>step 3a</u>: For every k such that $(d_N)_k > 0$, remove the corresponding column from the tableau.

<u>step 3b</u>: For every k such that $a_{ik}^* > 0$ for some i which satisfies $b_i^* = 0$ and $a_{i\cdot}^* \geq 0$, remove the corresponding column from the tableau.

<u>step 3c</u>: For every i such that $b_i^* = 0$ and $a_{i\cdot}^* = 0$, remove the corresponding row from the tableau.

<u>step 3d</u>: For every i such that $(x_B)_i$ is an artificial variable, say $(x_B)_i = z_\ell$, execute one pivot step with pivot element $a_{ik}^* < 0$ and remove the column corresponding to z_ℓ from the tableau.

<u>step 4</u>: Put the basis matrix on the list L_1 (L_1 will contain all basis matrices corresponding to extreme optimal solutions; the basis matrices, for which the adjacent extreme optimal solutions already are determined, are marked); put the optimal solution x on the list L_2 (L_2 will contain all extreme optimal solutions); set $L_3 = \emptyset$ (L_3 will contain all extreme infinite directions).

<u>step 5</u>: If all elements of L_1 are marked, then all extreme optimal solutions are stored in L_2 (extreme solutions) and L_3 (extreme directions); STOP.

<u>step 6</u>: Take any unmarked basis from L_1, mark this basis and determine the corresponding simplex tableau (denote the coefficients again by (b_i^*), (a_{ij}^*) and $((d_N)_j)$).

<u>step 7</u>: For every i and k such that $b_i^* = 0$, $a_{ik}^* \neq 0$ and such that the basis where the variables $(x_N)_k$ and $(x_B)_i$ are exchanged is not in L_1:

put this new basis on L_1.

<u>step 8</u>: For every k such that $a_{\cdot k}^* \leq 0$ and such that the direction vector s, where

$$s_j := \begin{cases} -a^*_{ik} & \text{if } x_j = (x_B)_i \\ 1 & \text{if } x_j = (x_N)_k \\ 0 & \text{elsewhere} \end{cases}$$

does not belong to the list L_3:

put this direction s on L_3.

step 9: For every k such that

(i) $a^*_{ik} > 0$ for at least one i

(ii) $\min\{b^*_i/a^*_{ik} \mid a^*_{ik} > 0\} = b^*_r/a^*_{rk} > 0$

(iii) the basis matrix which is obtained after exchanging the variables $(x_N)_k$ and $(x_B)_r$ is not in L_1

do: a. put this new basis matrix on L_1,

b. if the solution corresponding to this new basis is not in L_2, then put this solution on the list L_2.

step 10: Go to step 5.

CHAPTER 2

MARKOV DECISION PROCESSES

2.1. INTRODUCTION AND SUMMARY

In this chapter we present a survey of some results about Markov chains and Markov decision processes. This survey is far from comprehensive. We only discuss the topics we need in the following chapters of this monograph.

In section 2.2 we introduce the Markov decision models with various optimality criteria such as discounted optimality, average optimality, bias optimality and Blackwell optimality. Furthermore, we give some notations and definitions.

Section 2.3 deals with the theory of Markov chains. We give a summary of some well-known results on the transition matrix and the stationary matrix. Also we present an algorithm for identifying the ergodic sets and the transient states of a stochastic matrix, and an algorithm for the computation of the stationary matrix.

In section 2.4 we review some results on (sub)stochastic matrices. We present some properties of the stationary, the fundamental and the deviation matrix.

In section 2.5 we mention results about the existence of optimal pure and stationary policies for the optimality criteria introduced in section 2.2. Also, we present a theorem, due to Derman and Strauch, which implies that restriction to Markov policies is allowed. Furthermore, we give a result, due to Blackwell, which relates discounted rewards to average rewards for discount factors near to 1.

2.2. MARKOV DECISION MODELS

Consider a dynamic system that is observed at discrete time points $t = 1,2,\dots$. We allow that with positive probability the system breaks down and then the process is terminated. If at any discrete time point t the system is in one of a finite number of states, then an action has to be chosen. The state space is denoted by $E = \{1,2,\dots,N\}$ and $A(i)$ is the finite set of possible actions in state i, $i \in E$. If the system is in state i and action $a \in A(i)$ is chosen, then the following happens, independently of the history of the process:

1. A reward r_{ia} is earned immediately.
2. The next state of the process is chosen according to the transition probabilities p_{iaj}, where $p_{iaj} \geq 0$ and $\Sigma_j \, p_{iaj} \leq 1$ for every $a \in A(i)$ and $i,j \in E$.

A (discrete) *Markov decision problem* is given by a four-tuple (E,A,p,r), where

- E is the *state space*,
- $A = \cup_{i \in E} A(i)$ is the *action space*,
- p is a *transition probability* from $E \times A$ to E,
- r is a real-valued *reward function* on $E \times A$,

($E \times A$ has to be interpreted as $\{(i,a) \mid i \in E, \, a \in A(i)\}$). A Markov decision problem is also called a *(stochastic) dynamic programming problem*.

Let H_t denote the set of possible *histories* of the system up to time t, i.e.

$$H_t := \{(i_1,a_1,\dots,i_{t-1},a_{t-1},i_t) \mid i_k \in E, a_k \in A(i_k), \; k=1,2,\dots,t-1; \; i_t \in E\}.$$

A *decision rule* π^t at time t is a nonnegative function on $H_t \times A$ such that for every $(i_1,a_1,\dots,i_t) \in H_t$

$$\pi^t_{i_1a_1\dots i_ta_t} = 0 \qquad \text{if } a_t \notin A(i_t)$$

and

$$\textstyle\sum_{a_t} \pi^t_{i_1a_1\dots i_ta_t} = 1.$$

A *policy* R is a sequence of decision rules: $R = (\pi^1,\pi^2,\dots,\pi^t,\dots)$. We let $\mathcal{C}$ denote the class of all policies. A policy $R = (\pi^1,\pi^2,\dots)$ is said to be *memoryless* if the decision rule π^t is independent of $(i_1,a_1,\dots,i_{t-1},a_{t-1})$ for every $t \in \mathbb{N}$. Memoryless policies are also called *Markov policies*. By $\mathcal{C}_M$ we denote the class of Markov policies. We let $\mathcal{C}_S$ denote the class of *stationary policies*, i.e. the Markov policies for which π^t is time invariant. Hence, a stationary policy is completely determined by a decision rule which depends only on the last state i. We will denote the stationary policy $R = (\pi,\pi,\dots)$ by π^∞. By $\mathcal{C}_D$ we denote the subclass of $\mathcal{C}_S$ consisting of the *pure and stationary policies*, i.e. stationary policies with nonrandomized decision rules. Therefore, a pure and stationary policy can be described by a function f defined on E such that $f(i) \in A(i)$, $i \in E$. We will denote this policy by f^∞.

For any $R = (\pi^1,\pi^2,\dots) \in \mathcal{C}$, we denote by p^t_{ija} the probability that the system is at time t in state j and then action a is chosen, given that the system is at time $t=1$ in state i. For $R \in \mathcal{C}_M$ the numbers $p^t_{ija}(R)$ can be computed iteratively:

$$p^1_{ija}(R) = \begin{cases} 0 & j \neq i, \quad a \in A(j) \\ \pi^1_{ja} & j = i, \quad a \in A(j) \end{cases}$$

$$p^{t+1}_{ija}(R) = \sum_{i_t,a_t} p^t_{ii_ta_t}(R)\cdot p_{i_ta_tj}\cdot\pi^{t+1}_{ja} \qquad , j\in E, a\in A(j),\ t\in \mathbb{N}.$$

Let us define the matrix $P(\pi^t)$ by $P(\pi^t) := (\Sigma_a p_{iaj}\pi^t_{ia})$, $\qquad t\in \mathbb{N}$.
Then, for $R \in \mathcal{C}_M$ we have

$$p^{t+1}_{ija}(R) = [P(\pi^1)P(\pi^2)\dots P(\pi^t)]_{ij}\cdot\pi^{t+1}_{ja},\ j\in E,\ a\in A(j),\ t \in \mathbb{N}_0,$$

where $P(\pi^1)P(\pi^2)\dots P(\pi^t) := I$ if $t = 0$.

Let $\{X_t,\ t = 1,2,\dots\}$ and $\{Y_t,\ t = 1,2,\dots\}$ be the sequences of random variables denoting the observed states and chosen actions respectively. Then, we can also write

$$p^t_{ija}(R) = \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i).$$

Furthermore, we denote by $p^t_{ij}(R)$ the probability that the system is at time t in state j, given that state i is the starting state. Hence, we obtain

$$p^t_{ij}(R) = \mathbb{P}_R(X_t = j \mid X_1 = i) = \Sigma_a\ \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i).$$

The matrix $P^t(R)$ is defined by $P^t(R) := (p^{t+1}_{ij}(R))$.

The expected reward in the t-th period, given initial state i and the use of policy R, is denoted by $v^t_i(R)$, i.e.

$$v_i^t(R) := \sum_j \sum_a \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja}.$$

The *expected total reward* over an infinite horizon, given initial state i and the use of policy R, where R is such that $\lim_{T\to\infty} \Sigma_{t=1}^T\ v_i^t(R)$ exists (possibly $+\infty$ or $-\infty$), is denoted by $v_i(R)$, i.e.

$$v_i(R) := \sum_{t=1}^{\infty} \sum_j \sum_a \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja}.$$

For a real number $\alpha \in [0,1)$ the *expected discounted reward,* given initial state i and the use of policy R, is denoted by $v_i^{\alpha}(R)$, i.e.

$$v_i^{\alpha}(R) := \sum_{t=1}^{\infty} \alpha^{t-1} \sum_j \sum_a \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja}.$$

α is called the *discount factor*. The *expected average reward* over an infinite horizon, given initial state i and the use of policy R, is denoted by $\phi_i(R)$ and defined by

$$\phi_i(R) := \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_j \sum_a \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i) \cdot r_{ja}.$$

For a Markov decision model with as utility function the total reward criterion we will use the name *TMD-model*. In a TMD-model we define the *TMD-value-vector* v by

$$v_i := \sup_R\ v_i(R), \quad i \in E.$$

A policy R^* is said to be *total optimal* if $v(R^*) = v$. A Markov decision model with the discounted reward criterion is called a *DMD-model*. The *DMD-value-vector* v^{α} is defined by

$$v_i^{\alpha} := \sup_R\ v_i^{\alpha}(R), \quad i \in E.$$

A policy R^* is *α-discounted optimal* if $v^{\alpha}(R^*) = v^{\alpha}$; a policy R^* is said to be *bias optimal* if $\lim_{\alpha\uparrow 1}\{v_i^{\alpha}(R^*)-v_i^{\alpha}\} = 0$, $i \in E$; a policy R^* is called *Blackwell optimal* if for some $\alpha_o \in [0,1)$ R^* is α-discounted optimal for every $\alpha \in [\alpha_o,1)$.

If we use as utility function the average reward criterion, then the name of the model will be abbreviated by *AMD-model*. The *AMD-value-vector*

ϕ is defined by

$$\phi_i := sup_R \ \phi_i(R), \quad i \in E.$$

The policy R^* is *average optimal* if $\phi(R^*) = \phi$.

The policy R is said to be a *transient policy* if $\Sigma_{t=1}^{\infty} p_{ij}^t(R) < \infty$ for every $i,j \in E$. Hence, for any transient policy $v_i(R)$ is finite for every $i \in E$. If $R = (\pi^1, \pi^2, \ldots) \in C_M$ is transient, then we may write

$$v(R) = \sum_{t=1}^{\infty} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t),$$

where

$$r(\pi^t) := (\sum_a r_{ia}\pi_{ia}^t).$$

Furthermore, if $\pi^\infty \in C_S$ is transient, then we have (cf. KEMENY & SNELL [1960] p.22)

$$v(\pi^\infty) = \sum_{t=1}^{\infty} P^{t-1}(\pi)r(\pi) = (I-P(\pi))^{-1}r(\pi).$$

If a TMD-model satisfies the condition that every policy is transient, then the model is called a *transient dynamic programming problem*.

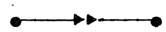
A TMD-model with $r_{ia} \geq 0$ $a \in A(i)$, $i \in E$, is said to be a *positive dynamic programming* model; if all rewards are nonpositive, then we have a *negative dynamic programming* model.

A dynamic programming problem is called *contracting* if there exists a vector $\mu >> 0$ and a scalar $\alpha \in [0,1)$ such that

$$\sum_j p_{iaj}\mu_j \leq \alpha \mu_i \quad a \in A(i),\ i \in E.$$

Any DMD-problem is contracting (re-define $p_{iaj} := \alpha p_{iaj}$ $i,j \in E$, $a \in A(i)$ and take $\mu = e$); it can easily be verified that in a contracting dynamic programming problem any policy is transient. Hence, the transient dynamic programming problem is a generalization of the contracting dynamic programming problem (in fact, these problems are equivalent as will be shown in theorem 3.2.4). The name contracting dynamic programming was introduced by van Nunen and Wessels, who have studied this model systematically (e.g. VAN NUNEN & WESSELS [1977]).

REMARK 2.2.1. In the sequel we will present examples of models and illustrate them in a picture. In these models the transition probabilities will always be degenerated, i.e. for any $a \in A(i)$ and $i \in E$ we have $p_{iaj} \neq 0$ for at most one state j. Hence, to indicate which state is the next state of the system, when in state i action a is chosen, we can use in the picture an arc from state i to state j where j is such that $p_{iaj} \neq 0$. For the different actions $1,2,\ldots,k_i$ in state i, these arcs are drawn as

(action 1),


(action 2) etc. .

In the TMD-models we add to every arc that corresponds to (i,a) and is directed from state i to state j the pair r_{ia}, p_{iaj} . For AMD-models we shall assume that $\Sigma_j\, p_{iaj} = 1$ for every $a \in A(i)$, $i \in E$. Therefore, we may add to an arc only the number r_{ia} . Figure 2.2.1 gives the picture which corresponds to the following TMD-model:
$E = \{1,2,3\}$; $A(i) = \{1,2\}$, $i \in E$;
$p_{112}=1/2$, $p_{123}=1$, $p_{211}=1/2$, $p_{222}=1/4$, $p_{313} = 1$, $p_{322} = 1/2$ (the other transition probabilities are zeros);
$r_{11}=1$, $r_{12}=0$, $r_{21}=-1$, $r_{22}=2$, $r_{31}=-2$, $r_{32}=0$.

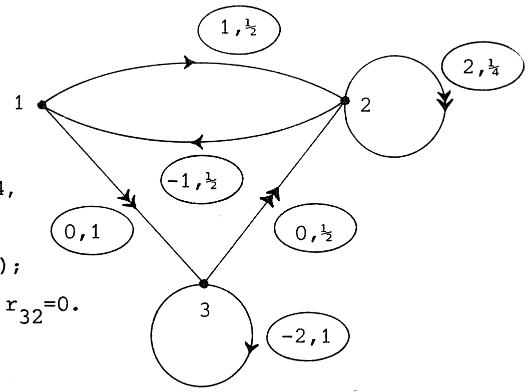


Figure 2.2.1

2.3. MARKOV CHAINS

Assume that $\Sigma_j\, p_{iaj} = 1$ for all $a \in A(i)$, $i \in E$. Then for any stationary policy π^∞ the sequence of observed states $\{x_t,\ t = 1,2,\ldots\}$ is a finite stationary *Markov chain* with transition probabilities $p_{ij} = \Sigma_a p_{iaj}\pi_{ia}$, $i,j \in E$. Hence, the theory of Markov chains plays an important role in the analysis of Markov decision models. In this section we will summarize some results for reference purposes. For the proofs we will refer to one of many books that deal with Markov chains. We assume that the reader is familiar with concepts such as: *transient state, recurrent state, ergodic set, communicating states, absorbing state and absorption probabilities*.

The Markov chain is called *completely ergodic* if all states are recurrent and there is exactly one ergodic set. If there is exactly one ergodic set plus possibly some transient states, then the Markov chain is said to be *unichained*. A subset E_0 of E is said to be *closed* under P if $p_{ij} = 0$ for all $i \in E_0$ and $j \in E \backslash E_0$.

Let $E_1, E_2, \dots, E_m$ be the ergodic sets and let F be the set of all transient states of a Markov chain with state space $E = \{1,2,\dots,N\}$. Then, by appropriate rearranging, we obtain the following form for the transition matrix P:

$$(2.3.1) \qquad P = \begin{pmatrix} P_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & P_m & 0 \\ R_1 & R_2 & \dots & R_m & Q \end{pmatrix} \begin{matrix} E_1 \\ E_2 \\ \cdot \\ \cdot \\ \cdot \\ E_m \\ F \end{matrix}$$

THEOREM 2.3.1. *The matrix* $I-Q$ *is nonsingular and* $(I-Q)^{-1} = \Sigma_{n=0}^{\infty} Q^n$.

PROOF. See KEMENY & SNELL [1960] p.46. □

DEFINITION 2.3.1. Let $B_1, B_2, \dots$ and B be real k×k matrices, and let $\tilde{B}_n = \frac{1}{n} \Sigma_{\ell=1}^{n} B_\ell$, $n \in \mathbb{N}$. If $B = \lim_{n\to\infty} \tilde{B}_n$, then we write

$$B \overset{(c)}{=} \lim_{n\to\infty} B_n \quad \text{(the notation (c) stands for } \textit{Cesaro limit}\text{).}$$

THEOREM 2.3.2.

(i) $P^* \overset{(c)}{:=} \lim_{n\to\infty} P^n$ *exists.*

(ii) $P^*P = PP^* = P^*P^* = P^*$.

(iii) $p^*_{i\cdot} = p^*_{j\cdot}$ *and* $p^*_{ij} > 0$ *for any pair* (i,j) *such that* i *and* j *belong to the same ergodic set.*

(iv) $p^*_{\cdot i} = 0$ *for any transient state i.*

PROOF. See DOOB [1953] p.175. □

DEFINITIONS 2.3.2.

(i) The matrix P^* is called the *stationary matrix* of matrix P.

(ii) Any solution of the set of equations

$$x \geq 0, \quad x^T e = 1 \quad \text{and} \quad x^T = x^T P$$

is a *stationary probability distribution* of the Markov chain.

THEOREM 2.3.3. *Let* x *be any stationary probability distribution of the Markov chain. Then*

$$x_i = \begin{cases} 0 & \text{if } i \in F \\ c_k p^*_{ii} & \text{if } i \in E_k \text{ where } c_k \text{ satisfies } \Sigma^m_{k=1} c_k = 1. \end{cases}$$

PROOF. See DOOB [1953] p.183. □

COROLLARY 2.3.1. *If* $x^T = x^T P$ *and* $E_x := \{i | x_i > 0\}$, *then* E_x *is the union of some ergodic sets and consequently,* E_x *is a closed set.*

NOTATION 2.3.1. For any transient state i we denote the absorption probability that the process will be ultimately absorbed into the ergodic set E_k by a_{ik} $k = 1,2,\ldots,m$.

THEOREM 2.3.4. *For any ergodic set* E_k, *we have*

$$p^*_{ij} = a_{ik} p^*_{jj}, \qquad i \in F,\ j \in E_k$$

and $\{a_{ik}, i \in F\}$ *is the unique solution of the linear system*

$$\tilde{a}_{ik} = \sum_{j \in E_k} p_{ij} + \sum_{j \in F} p_{ij} \tilde{a}_{jk}, \qquad i \in F.$$

PROOF. See FELLER [1967] p.403. □

If the ergodic sets and the transient states of a Markov chain are identified, then the stationary matrix P^* can be computed using the results of the theorems 2.3.4. We will describe an algorithm proposed by FOX & LANDI [1968] to find the ergodic sets and the transient states. This algorithm is based on repeated use of the following rules:

1. State i is absorbing if and only if $p_{ij} = 0$ for all $j \neq i$.
2. If state i is absorbing and $p_{ki} > 0$, then state k is transient.
3. If state i is transient and $p_{ki} > 0$, then state k is also transient.
4. If state i communicates with state j and state j communicates with state k, then state i communicates with state k.

The search for a set of communicating states is conducted by generating a chain of states such that each state can be reached from its predecessor with positive probability in one transition. If the chain encounters a state that has already been classified to be transient, then all states in the chain are transient. Otherwise, a circuit of states is obtained. Then, this circuit is replaced by one composite state. If by rule 1 the composite state is absorbing, then the states of the composite states form an ergodic set and the states in the chain that precede the circuit are transient; otherwise, extension of the chain is continued from the composite state. Hence, in a finite number of steps at least one state is classified to be recurrent or transient. This guarantees the finiteness of the following algorithm.

ALGORITHM II for identifying the ergodic sets and the transient states of a Markov chain with transition matrix P.

<u>step 1</u>: Take $S_i = \{i\}$ for every state i.

<u>step 2a</u>: Every state i such that $p_{ij} = 0$ for all $j \neq i$ is labeled as an absorbing state.

<u>step 2b</u>: For each identified absorbing state i, label state i as an ergodic set, and label every state k satisfying $p_{ki} > 0$ as transient state.

<u>step 3</u>: If all states are labeled, then go to step 6.
Otherwise, go to step 4a.

<u>step 4a</u>: Choose any unlabeled state i, set r = 1 and let $i_r = i$.

<u>step 4b</u>: Search in row i_r for a positive element, say $p_{i_r i_{r+1}}$, such that $i_r \neq i_{r+1}$.

<u>step 4c</u>: If state i_{r+1} is labeled as a transient state, then:
(i) label each state in the set $\{S_{i_1} \cup S_{i_2} \cup \cdots \cup S_{i_r}\}$ as transient,
(ii) go to step 3.
Otherwise, go to step 4d.

<u>step 4d</u>: If $i_{r+1} = i_k$ for some $k \in \{1,2,\ldots,r-1\}$, then go to step 5a.
Otherwise, r := r+1 and go to step 4b.

<u>step 5a</u>: Replace row i_k by the sum of the rows $\{i_k, i_{k+1}, \ldots, i_r\}$ and delete the rows $\{i_{k+1}, i_{k+2}, \ldots, i_r\}$ from the matrix; replace column i_k by the sum of the columns $\{i_k, i_{k+1}, \ldots, i_r\}$ and delete the columns $\{i_{k+1}, i_{k+2}, \ldots, i_r\}$; set $S_{i_k} = \bigcup_{j=k}^{r} S_{i_j}$.

<u>step 5b</u>: If the composite state i_k is absorbing, then:

(i) label i_k as an ergodic set and $i_1, i_2, \ldots, i_{k-1}$ as transient states,

(ii) label every state j which satisfies $\Sigma_{\ell=1}^{k} p_{j i_\ell} > 0$ as transient state,

(iii) go to step 3.

Otherwise, $r := k$ and go to step 4b.

step 6: The transient states are labeled as transient, and every other state i_k (whether or not composite) corresponds to an ergodic set and S_{i_k} contains the states of this ergodic set.

The results stated above imply that the stationary matrix P^* can be determined by the following algorithm.

ALGORITHM III for the computation of the stationary matrix P^*.

step 1: Identify the transient states F and the ergodic sets $E_1, E_2, \ldots, E_m$ of the Markov chain by algorithm II.

step 2: Determine for $k = 1,2,\ldots,m$

(i) the unique solution $\{x_j^k, j \in E_k\}$ of the linear system

$$\begin{cases} \sum_{j \in E_k} (\delta_{j\ell} - p_{j\ell}) \tilde{x}_j^k = 0 & \ell \in E_k \\ \sum_{j \in E_k} \tilde{x}_j^k = 1 & \end{cases}$$

(ii) the unique solution $\{a_j^k, j \in F\}$ of the linear system

$$\sum_{j \in F} (\delta_{ij} - p_{ij}) \tilde{a}_j^k = \sum_{j \in E_k} p_{ij} \qquad i \in F$$

step 3:

$$p_{ij}^* := \begin{cases} x_j^k & i \in E_k,\ j \in E_k, \quad k = 1,2,\ldots,m, \\ a_i^k x_j^k & i \in F,\ j \in E_k, \quad k = 1,2,\ldots,m, \\ 0 & \text{elsewhere.} \end{cases}$$

2.4. SUBSTOCHASTIC MATRICES

DEFINITION 2.4.1. A real n×n matrix $P = (p_{ij})$ is said to be *substochastic* if $p_{ij} \geq 0$ for all i,j and $\Sigma_j p_{ij} \leq 1$ for all i; if, moreover, $\Sigma_j p_{ij} = 1$ for all i, then P is called a *stochastic matrix*.

Throughout this section we assume that P is a substochastic matrix. In the following theorem we summarize some well-known results of substochastic matrices. For the proofs we refer to BLACKWELL [1962] and VEINOTT [1974].

THEOREM 2.4.1.
(i) $P^* \overset{(c)}{:=} \lim_{n\to\infty} P^n$ *exists and satisfies* $P^*P = PP^* = P^*P^* = P^*$.
(ii) $\lim_{\alpha\uparrow 1} (1-\alpha) \Sigma_{k=0}^{\infty} \alpha^k (P-P^*)^k = 0$.
(iii) $I - P + P^*$ *is nonsingular and moreover*

$$(I-P+P^*)^{-1} = \lim_{\alpha\uparrow 1} \sum_{k=0}^{\infty} \alpha^k (P-P^*)^k.$$

(iv) *Let* $D := (I-P+P^*)^{-1}-P^*$. *Then*

$$D = \lim_{\alpha\uparrow 1} \sum_{k=0}^{\infty} \alpha^k (P^k-P^*) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{\ell=1}^{k} (P^{\ell-1}-P^*)$$

and

$$P^*D = DP^* = (I-P)D+P^*-I = D(I-P)+P^*-I = 0.$$

DEFINITION 2.4.2. The matrices P^*, $(I-P+P^*)^{-1}$ and D are said to be the *stationary*, the *fundamental* and the *deviation matrix* of the substochastic matrix P, respectively.

LEMMA 2.4.1. *If the matrix* P *is stochastic, then* $De = 0$.

PROOF. Using theorem 2.4.1(iv), we obtain

$$De = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{\ell=1}^{k} (P^{\ell-1}-P^*)e$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{\ell=1}^{k} (P^{\ell-1}e-P^*e)$$

$$= 0. \quad \square$$

Any stochastic N×N matrix may be interpreted as the transition matrix of a Markov chain with state space $\{1,2,\dots,N\}$. In the following chapters we also encounter substochastic matrices that are not stochastic. However, such a matrix may be interpreted as a submatrix of the transition matrix $\tilde{P}$ of a Markov chain with state space $\{0,1,\dots,N\}$, where

$$(2.4.1)\qquad \tilde{P} = \left(\begin{array}{c|c} 1 & 0 \\ \hline (I-P)e & P \end{array}\right).$$

Since $\tilde{P}$ and $\tilde{P}^*$ are stochastic and, by lemma 2.4.1, $\tilde{D}e = 0$, it follows from (2.4.1) that

$$\tilde{P}^* = \left(\begin{array}{c|c} 1 & 0 \\ \hline (I-P^*)e & P^* \end{array}\right) \quad \text{and} \quad \tilde{D} = \left(\begin{array}{c|c} 0 & 0 \\ \hline -De & D \end{array}\right),$$

where P^* and D are the stationary and deviation matrix of P, respectively. The additional state 0 is an absorbing state. Suppose that there are furthermore m (possibly $m = 0$) ergodic sets $E_1, E_2, \ldots, E_m$ in the Markov chain with state space $\{0,1,\ldots,N\}$ and let F be the set of transient states. The number of states in E_k is denoted by N_k, $k = 1,2,\ldots,m$. By appropriate rearranging, we may write P in the form

$$(2.4.2)\qquad P = \left(\begin{array}{cccc|c} P_1 & 0 & \cdots & 0 & 0 \\ 0 & P_2 & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & P_m & 0 \\ \hline R_1 & R_2 & \cdots & R_m & Q \end{array}\right) \begin{array}{c} E_1 \\ E_2 \\ \cdot \\ \cdot \\ \cdot \\ E_m \\ F \end{array}$$

The matrix P_k^* has identical rows; denote this row by the N_k-vector p_k^*. Then using the result of theorem 2.3.4, it can be verified that we may write P^* and D as

$$(2.4.3)\qquad P^* = \left(\begin{array}{cccc|c} P_1^* & 0 & \cdots & 0 & 0 \\ 0 & P_2^* & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & P_m^* & 0 \\ \hline A_1 & A_2 & \cdots & A_m & 0 \end{array}\right),\ D = \left(\begin{array}{cccc|c} D_1 & 0 & \cdots & 0 & 0 \\ 0 & D_2 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & D_m & 0 \\ \hline B_1 & B_2 & \cdots & B_m & (I-Q)^{-1} \end{array}\right)$$

where

$$A_k = [(I-Q)^{-1}R_k e](p_k^*)^T \qquad k = 1,2,\ldots,m$$

$$D_k = (I-P_k+P_k^*)^{-1} - P_k^* \qquad k = 1,2,\ldots,m$$

$$B_k = (I-Q)^{-1}(R_k-A_k)(D_k+P_k^*) - A_k \qquad k = 1,2,\ldots,m.$$

If $m = 0$, then P^* is the null-matrix and $D = (I-Q)^{-1}$. For the sequel of this section, we assume that $m \geq 1$. Let i_k be an arbitrary state in the ergodic set E_k, $k = 1,2,\ldots,m$. Suppose that r is any N-vector and that B is any diagonal N×N matrix with nonnegative elements. Then we have the following result (cf. DENARDO [1971]).

LEMMA 2.4.2. *Suppose that* x *is a solution of the linear system*

$$\text{(2.4.4)} \qquad \begin{cases} (I-P)\tilde{x} = 0 \\ P^*B\tilde{x} = P^*r. \end{cases}$$

Then

$$x_i = \begin{cases} (P^*r)_{i_k} \Big/ (P^*Be)_{i_k} & i \in E_k,\ k = 1,2,\ldots,m \\ \sum_{k=1}^m a_{ik} \dfrac{(P^*r)_{i_k}}{(P^*Be)_{i_k}} & i \in F. \end{cases}$$

The following lemma gives a related result for a system of inequalities.

LEMMA 2.4.3. *Suppose that* x *is a solution of* (2.4.4) *and that* $\bar{x}$ *satisfies*

$$\begin{cases} (I-P)\bar{x} \geq 0 \\ P^*B\bar{x} \geq P^*r. \end{cases}$$

Then, $\bar{x} \geq x$.

PROOF. Let $a = (I-P)\bar{x}$. Then, $a \geq 0$ and $P^*a = 0$, implying that $a_i = 0$ $i \in E\backslash F$. Consequently, $\bar{x}_i = (P\bar{x})_i$ $i \in E\backslash F$ and also $\bar{x}_i = (P^*\bar{x})_i$, $i \in E\backslash F$. Hence, the value of $\bar{x}$ is constant on any ergodic set. Therefore, we can write

$$\bar{x}_i \geq \frac{(P^* r)_i}{(P^* Be)_i} = x_i, \quad i \in E\backslash F. \tag{2.4.5}$$

Let $\bar{x}_F$ consist of the components of $\bar{x}$ corresponding to the transient states. Then, $\bar{x} \geq P\bar{x}$, (2.4.2) and (2.4.5) imply

$$\bar{x}_F \geq \sum_{k=1}^m \bar{x}_{i_k} \cdot R_k e + Q\bar{x}_F \geq \sum_{k=1}^m x_{i_k} \cdot R_k e + Q\bar{x}_F.$$

Since (I-Q) is nonsingular and nonnegative, we obtain

$$\bar{x}_F \geq \sum_{k=1}^m x_{i_k} \cdot (I-Q)^{-1} R_k e.$$

Hence,

$$\bar{x}_j \geq \sum_{k=1}^m [(I-Q)^{-1} R_k e]_j \cdot x_{i_k}, \qquad j \in F.$$

Theorem 2.3.4 implies that

$$\bar{x}_j \geq \sum_{k=1}^m a_{jk} x_{i_k}, \qquad j \in F. \tag{2.4.6}$$

The inequalities (2.4.5) and (2.4.6) yield $\bar{x} \geq x$. □

2.5. EXISTENCE OF OPTIMAL POLICIES

The following theorem, due to DERMAN & STRAUCH [1966] and generalized by STRAUCH & VEINOTT [1966] and HORDIJK [1974] pp.115-117, indicates that we may restrict ourselves to memoryless policies.

<u>THEOREM 2.5.1</u>. *Given any initial distribution* $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$ *, any sequence of policies* $R_1, R_2, \ldots$ *and any sequence of nonnegative real numbers* $p_1, p_2, \ldots$ *with* $\Sigma_{k=1}^\infty p_k = 1$, *there exists a memoryless policy* R *such that*

$$\sum_i \beta_i \cdot \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i) = \tag{2.5.1}$$

$$\sum_i \beta_i \sum_k p_k \mathbb{P}_{R_k}(X_t = j, Y_t = a \mid X_1 = i) \qquad t \in \mathbb{N}, a \in A(j), j \in E.$$

COROLLARY 2.5.1. *Given any initial state* $i \in E$ *and any policy* $R \in C$, *there exists a policy* $R_o \in C_M$ *such that*

$$\mathbb{P}_{R_o}(X_t = j,\ Y_t = a \mid X_1 = i) = \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i)$$

$$t \in \mathbb{N},\ a \in A(j),\ j \in E.$$

We continue this section with some properties of the DMD-model. The results are folklore and for the proofs we will refer to a standard book on Markov decision processes

THEOREM 2.5.2.

(i) *The* DMD-*value-vector* v^α *is the unique solution of the functional equation*

$$(2.5.2) \qquad x_i = \max_a\{r_{ia} + \alpha \textstyle\sum_j p_{iaj}\, x_j\}, \qquad i \in E.$$

(ii) *Let* $a_i \in A(i)$ *be such that*

$$r_{ia_i} + \alpha \textstyle\sum_j p_{ia_ij}\, v_j^\alpha = \max_a\{r_{ia} + \alpha \sum_j p_{iaj}\, v_j^\alpha\}, \qquad i \in E.$$

Then the pure and stationary policy f^∞, *where* $f(i) = a_i$, $i \in E$, *is* α-*discounted optimal.*

PROOF. See ROSS [1970] pp.121-128. □

THEOREM 2.5.3. *There exists a pure and stationary Blackwell optimal policy.*

PROOF. See DERMAN [1970] pp.24-25. □

If π^∞ is a stationary policy, then $\mathbb{P}_{\pi^\infty}(X_t = j \mid X_1 = i) = (P^{t-1}(\pi))_{ij}$, $t \in \mathbb{N}$, $i,j \in E$. Hence,

$$(2.5.3) \qquad v^t(\pi^\infty) = P^{t-1}(\pi)r(\pi), \qquad t \in \mathbb{N},$$

where $r(\pi) := (\Sigma_a r_{ia}\pi_{ia})$. We also have

$$(2.5.4) \qquad \phi(\pi^\infty) = \liminf_{T\to\infty} \frac{1}{T} \textstyle\sum_{t=1}^{T} P^{t-1}(\pi)r(\pi) = P^*(\pi)r(\pi).$$

If the Markov chain induced by π^∞ is *unichained*, (i.e. there is at most one ergodic set, then $P^*(\pi)$ has identical rows, and consequently $\phi(\pi^\infty)$ has identical components.

NOTATION 2.5.1. For any stationary policy π^∞, we denote the vector $D(\pi)r(\pi)$, where $D(\pi)$ is the deviation matrix of $P(\pi)$, by $u(\pi^\infty)$:

$$(2.5.5) \qquad u(\pi^\infty) := D(\pi)r(\pi).$$

From theorem 2.4.1(iv) it follows that

$$(2.5.6) \qquad u(\pi^\infty) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{\sum_{s=1}^{t} v^s(\pi^\infty) - t\phi(\pi^\infty)\}$$

and

$$u(\pi^\infty) = \lim_{\alpha\uparrow 1} \sum_{t=1}^{\infty} \alpha^{t-1}\{P^{t-1}(\pi)r(\pi) - P^*(\pi)r(\pi)\}$$

$$= \lim_{\alpha\uparrow 1}\{v^\alpha(\pi^\infty) - \frac{\phi(\pi^\infty)}{1-\alpha}\}.$$

Hence,

$$(2.5.7) \qquad v^\alpha(\pi^\infty) = \frac{\phi(\pi^\infty)}{1-\alpha} + u(\pi^\infty) + \varepsilon(\alpha),$$

where $\lim_{\alpha\uparrow 1} \varepsilon(\alpha) = 0$.

THEOREM 2.5.4. *Any Blackwell optimal policy is average optimal as well as bias optimal.*

PROOF. From the definition of bias optimality it is obvious that Blackwell optimality implies bias optimality. In DERMAN [1970] pp.25-26 is shown that Blackwell optimality implies average optimality. □

COROLLARY 2.5.2. *There exist pure and stationary average optimal and bias optimal policies.*

REMARK 2.5.1. In chapter 5 we will show that bias optimality implies average optimality.

REMARK 2.5.2. A finite algorithm to compute a Blackwell optimal policy can be found in HORDIJK, DEKKER & KALLENBERG [1981].

CHAPTER 3

TOTAL REWARD CRITERION

3.1. INTRODUCTION AND SUMMARY

In this chapter we consider Markov decision problems with the expected total reward as optimality criterion. Already in 1953 SHAPLEY [1953] has analysed this type of problems in the context of stochastic games. The special case that we have a discounted dynamic programming problem has been studied extensively (see for instance the books written by HOWARD [1960], DERMAN [1970], ROSS [1970], MINE & OSAKI [1970] and HORDIJK [1974]). Linear programming formulations for the discounted dynamic programming problem are due to D'EPENOUX [1960] and DE GHELLINCK & EPPEN [1967].

In section 3.2 we show that a pure and stationary policy, which is optimal with regard to the total reward criterion, always exists. Furthermore, we give a slight extension of Veinott's result (VEINOTT [1969]) concerning equivalent formulations of the concept of a contracting dynamic programming problem. From these results we derive two algorithms for checking the contraction property of a given dynamic programming problem.

Section 3.3 deals with the problem of finding optimal policies in the class of transient policies. We shall show that we can obtain such optimal policies from optimal solutions of a linear programming problem. If we use the simplex method to solve this linear program, then a pure and stationary optimal policy is obtained (see algorithm VI). We also discuss a constrained dynamic programming problem, where the constraints are linear functions of the expected number of times of being in state j and then choosing action a, $a \in A(j)$, $j \in E$. Then, in general, there exists no optimal policy that also belongs to the class C_D. However, we can find by linear programming an optimal policy that is stationary (algorithm VII). Moreover, we show a one-to-one correspondence between the transient stationary policies and the feasible solutions of the proposed linear programming problem such that pure policies are mapped on extreme feasible solutions. We close this section with an application of the optimal stopping problem.

In section 3.4 we discuss the contracting dynamic programming problem. In this problem, all policies are transient; consequently, the results of section 3.3 are applicable. The results of section 3.3 can even be extended on some points (cf. theorem 3.3.4 versus theorem 3.4.8). Furthermore, we prove that, for this problem, linear programming by the simplex method is equivalent to the policy improvement method. We also show that elimination of suboptimal actions, as introduced by MACQUEEN [1967], can be implemented in the simplex method very easily using the dual variables appropriately. We close this section by the observation that discounted dynamic programming and contracting dynamic programming are equivalent models for unconstrained as well as for constrained Markovian decision problems.

Positive dynamic programming is the subject of section 3.5. We prove that, if the optimum of the linear programming problem is finite, then a pure and stationary optimal policy can be obtained directly from the linear programming solution. If the optimum is infinite, then by the linear program we can find a policy that, in general, is optimal only on a subset E_1 of the state space E. However, since $E \backslash E_1$ is closed under any policy, we may repeat the same procedure on the remaining states. In this way, we can construct a finite algorithm for positive dynamic programming (algorithm XII).

In section 3.6, where the negative dynamic programming problem is studied, we can derive a finite algorithm in a way similar to the analysis of section 3.5. In the algorithms of the sections 3.5 and 3.6 we have, besides solving linear programs, also to determine the structure of the Markov chain induced by some pure and stationary policies.

NOTATION 3.1.1. In this chapter, and also in the following chapters, we often use a vector, say x, with components x_{ia}, $a \in A(i)$, $i \in E$. However, we will also use the same notation x for the N-dimensional vector which has the components $x_i := \Sigma_a x_{ia}$, $i \in E$. Which vector is meant will always be clear from the context. Furthermore, we use the notation E_x, where E_x is defined by $E_x := \{i \in E \mid \Sigma_a x_{ia} > 0\}$.

3.2. PRELIMINARIES

In this section we discuss some properties of the TMD-value-vector v and we prove some theorems about transient policies. In order to have a well-defined concept of the expected total reward we use throughout this section the following assumption.

ASSUMPTION 3.2.1. *For any initial state* i *and any policy* R *the expected total reward* $v_i(R)$ *exists (possibly* $\pm\infty$*).*

We will show that, under the above assumption, there exists a pure and stationary optimal policy. First, we notice that the TMD-value-vector v exists (possibly $v_i = \pm\infty$ for some $i \in E$). For the proof of the existence of an optimal policy which belongs to the class C_D, we need the following lemma.

LEMMA 3.2.1. *For any initial state* i *and any policy* R, *we have*

$$\lim_{\alpha\uparrow 1} v_i^{\alpha}(R) = v_i(R).$$

PROOF. (cf. pp.65-67 in HORDIJK & TIJMS [1970]). Take any initial state $i \in E$ and any policy $R \in C$. We distinguish the following cases:

(i) $-\infty < v_i(R) < +\infty$

(ii) $v_i(R) = +\infty$

(iii) $v_i(R) = -\infty$.

case (i): Take any $\varepsilon > 0$. Then, there exists an integer T_o such that

$$|v_i(R) - \sum_{t=1}^{T} v_i^t(R)| < \varepsilon \quad \text{for every } T > T_o.$$

Since $|v_i^t(R)|$ is bounded for all t (e.g. by $\max_{i,a} |r_{ia}|$), the power series

$$v_i^{\alpha}(R) := \sum_{t=1}^{\infty} \alpha^{t-1} v_i^t(R)$$

has radius of convergence at least 1. The series $\Sigma_{t=1}^{\infty} \alpha^{t-1}$ has radius of convergence 1. Hence, for any $\alpha \in [0,1)$, we may write

$$(1-\alpha)^{-1} v_i^{\alpha}(R) = \sum_{s=1}^{\infty} \alpha^{s-1} \sum_{t=1}^{\infty} \alpha^{t-1} v_i^t(R)$$

$$= \sum_{t=1}^{\infty} (\sum_{s=1}^{t} v_i^s(R)) \alpha^{t-1}.$$

Therefore,

$$|(1-\alpha)^{-1}\{v_i^{\alpha}(R) - v_i(R)\}| \le \sum_{t=1}^{\infty} |\sum_{s=1}^{t} v_i^s(R) - v_i(R)| \alpha^{t-1} =$$

$$\sum_{t=1}^{T_o} \left|\sum_{s=1}^{t} v_i^s(R) - v_i(R)\right| \alpha^{t-1} + \sum_{t=T_o+1}^{\infty} \left|\sum_{s=1}^{t} v_i^s(R) - v_i(R)\right| \alpha^{t-1}.$$

Let $M := \max_{1 \le t \le T_o} \left|\Sigma_{s=1}^{t} v_i^s(R) - v_i(R)\right|$. Then we can write

$$(1-\alpha)^{-1} \left|v_i^\alpha(R) - v_i(R)\right| \le M \cdot \frac{1-\alpha^{T_o}}{1-\alpha} + \varepsilon \sum_{t=T_o+1}^{\infty} \alpha^{t-1} < \frac{2\varepsilon}{1-\alpha}$$

for $\alpha \in [\alpha_1, 1)$, where $\alpha_1 < 1$ satisfies $M(1-\alpha^{T_o}) < \varepsilon$ for $\alpha \ge \alpha_1$. Hence, we have shown that

$$\lim_{\alpha \uparrow 1} v_i^\alpha(R) = v_i(R).$$

<u>case (ii)</u>: Choose any $M > 0$. Then, it follows that there exists an integer T_o such that $\Sigma_{t=1}^{T} v_i^t(R) > M$ for all $T > T_o$. Similarly to case (i), we can write

$$\begin{aligned}(1-\alpha)^{-1} v_i^\alpha(R) &= \sum_{t=1}^{\infty} \left(\sum_{s=1}^{t} v_i^s(R)\right) \alpha^{t-1} \\ &= \sum_{t=1}^{T_o} \left(\sum_{s=1}^{t} v_i^s(R)\right) \alpha^{t-1} + \sum_{t=T_o+1}^{\infty} \left(\sum_{s=1}^{t} v_i^s(R)\right) \alpha^{t-1} \\ &> m \cdot \frac{1-\alpha^{T_o}}{1-\alpha} + M \cdot \frac{\alpha^{T_o}}{1-\alpha} \ge \frac{1/2}{1-\alpha} \cdot M\end{aligned}$$

for every $\alpha \in [\alpha_2, 1)$, where α_2 satisfies $\frac{3}{4} \le \alpha_2 < 1$ and $m \cdot (1-\alpha_2^{T_o}) \ge -\frac{1}{4}M$ with $m := \min_{1 \le t \le T_o} \Sigma_{s=1}^{t} v_i^s(R)$. Therefore, we have shown that $\lim_{\alpha \uparrow 1} v_i^\alpha(R) = +\infty$.

<u>case (iii)</u>: The proof is similar to the proof of case (ii). □

<u>THEOREM 3.2.1</u>. *There exists a pure and stationary optimal policy.*

<u>PROOF</u>. (cf. HORDIJK [1976]). Theorem 2.5.3 implies the existence of a real number $\alpha_o \in [0,1)$ and of a policy $f^\infty \in C_D$ such that

$$v^\alpha(f^\infty) = v^\alpha \quad \text{for all } \alpha \in [\alpha_o, 1).$$

Then, from lemma 3.2.1, it follows that

$$\begin{aligned}v_i(f^\infty) &= \lim_{\alpha \uparrow 1} v_i^\alpha(f^\infty) = \lim_{\alpha \uparrow 1} v_i^\alpha \\ &\ge \lim_{\alpha \uparrow 1} v_i^\alpha(R) = v_i(R), \quad i \in E,\ R \in C.\end{aligned}$$

Hence,

$$v_i(f^\infty) = sup_R\, v_i(R), \quad i \in E,$$

i.e. f^∞ is a pure and stationary optimal policy. □

DEFINITION 3.2.1. For any $c \in [-\infty,+\infty]$ we define $0 \cdot c := 0$; moreover, we call a vector x with components $x_i \in [-\infty,+\infty]$, $i \in E$, *p-summable* if $\Sigma_j\, p_{iaj} x_j$ is well-defined for all $a \in A(i)$, $i \in E$ (i.e. not both of the values $+\infty$ and $-\infty$ may occur in the summation).

The following example shows that, in general, the TMD-value-vector v is not p-summable.

EXAMPLE 3.2.1. $E = \{1,2,3\}$; $A(i) = \{1\}$, $i \in E$; $p_{112} = p_{113} = \frac{1}{2}$, $p_{212} = 1$, $p_{313} = 1$; $r_{11} = 0$, $r_{21} = 2$, $r_{31} = -1$. Since all action sets consist of one element, there is only one policy, say R. Assumption 3.2.1 is satisfied, namely $v_1(R) = v_2(R) = +\infty$, $v_3(R) = -\infty$. Notice that in this example $v = v(R)$. Then, $\Sigma_j\, p_{11j} v_j$ is not defined, and consequently v is not p-summable.

THEOREM 3.2.2. *If* v *is* p-*summable, then* v *satisfies the functional equation*

$$\begin{cases} x_i = max_a\{r_{ia} + \Sigma_j\, p_{iaj} x_j\}, & i \in E, \\ x \textit{ is p-summable.} & \end{cases}$$

PROOF. Theorem 3.2.1 implies that $v = v(f^\infty)$ for some pure and stationary policy f^∞. Since v is p-summable, we may write

$$\text{(3.2.1)} \qquad v_i = v_i(f^\infty) = r_{if(i)} + \sum_j p_{if(i)j} v_j(f^\infty)$$

$$\leq max_a\{r_{ia} + \sum_j p_{iaj} v_j\}, \quad i \in E.$$

Let $a_i \in A(i)$, $i \in E$, be such that

$$r_{ia_i} + \sum_j p_{ia_ij} v_j = max_a\{r_{ia} + \sum_j p_{iaj} v_j\}.$$

Take policy $R = (\pi^1, \pi^2, \ldots) \in C_M$ such that

$$\pi^1_{ia} = \begin{cases} 1 & a = a_i \\ 0 & a \neq a_i \end{cases} \quad i \in E, \text{ and } \pi^t_{ia} = \begin{cases} 1 & a = f(i) \\ 0 & a \neq f(i) \end{cases} \quad i \in E,\ t \geq 2.$$

Then we can write

$$(3.2.2) \qquad v_i \geq v_i(R) = r_{ia_i} + \sum_j p_{ia_ij} v_j(f^\infty) = $$

$$\max_a \{r_{ia} + \sum_j p_{iaj} v_j\}, \qquad i \in E.$$

The relations (3.2.1) and (3.2.2) imply

$$v_i = \max_a \{r_{ia} + \sum_j p_{iaj} v_j\}, \qquad i \in E,$$

which completes the proof. □

THEOREM 3.2.3. *If there exists a transient policy, then there also exists a transient pure and stationary policy.*

PROOF. Since the existence of a transient policy is independent of the values of the rewards, we may assume that $r_{ia} = -1$, $a \in A(i)$, $i \in E$. Let $\tilde{R}$ be any transient policy, i.e.

$$\sum_{t=1}^\infty \mathbb{P}_{\tilde{R}}(X_t = j \mid X_1 = i) < \infty \quad \text{for all } i,j \in E.$$

Hence,

$$v_i(\tilde{R}) = \sum_{t=1}^\infty \sum_j \sum_a \mathbb{P}_{\tilde{R}}(X_t = j, Y_t = a \mid X_1 = i) \cdot (-1) > -\infty, \quad i \in E.$$

Since $v_i = \sup_R v_i(R)$, $i \in E$, we have $-\infty < v_i \leq 0$, $i \in E$. Theorem 3.2.1 implies the existence of a pure and stationary policy f^∞ such that $v_i(f^\infty) = v_i$, $i \in E$. Therefore,

$$-\infty < v_i(f^\infty) = \sum_{t=1}^\infty \sum_j \sum_a \mathbb{P}_{f^\infty}(X_t = j, Y_t = a \mid X_1 = i) \cdot (-1) \leq 0, \quad i \in E.$$

Consequently,

$$\sum_{t=1}^\infty \mathbb{P}_{f^\infty}(X_t = j \mid X_1 = i) < \infty \quad \text{for every } i,j \in E,$$

i.e. f^∞ is a transient policy. □

REMARK 3.2.1. For another proof of theorem 3.2.3 we refer to remark 3.3.2.

Next, we will give some equivalent characterizations of a transient dynamic programming problem. For the presentation of this result we use the following definition and lemma.

DEFINITION 3.2.2. Suppose that we change a TMD-model in another TMD-model in the following way:

$$\tilde{E} := E \cup \{0\}$$

$$\tilde{A}(i) := \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases}$$

$$\tilde{p}_{iaj} := \begin{cases} p_{iaj} & i \neq 0,\ j \neq 0,\ a \in \tilde{A}(i) \\ 1 - \Sigma_{k=1}^{N} p_{iak} & i \neq 0,\ j = 0,\ a \in \tilde{A}(i) \\ 0 & i = 0,\ j \neq 0,\ a \in \tilde{A}(i) \\ 1 & i = 0,\ j = 0,\ a \in \tilde{A}(i) \end{cases}$$

$$\tilde{r}_{ia} := \begin{cases} r_{ia} & i \neq 0,\ a \in \tilde{A}(i) \\ 0 & i = 0,\ a \in \tilde{A}(i) \end{cases}$$

Then the transformed model is called the *extended TMD-model*.

LEMMA 3.2.2. *Let the sequence of vectors* $\{y^t, t = 0,1,\dots\}$ *be defined by*

$$y_i^0 := 1 \qquad i \in E$$

$$y_i^t := max_a \sum_j p_{iaj} y_j^{t-1},\ i \in E,\ t \in \mathbb{N},$$

and let the sequence of pure and stationary policies $\{f_t^\infty, t = 1,2,\dots\}$ *satisfy*

$$y_i^t = \sum_j p_{if_t(i)j}\, y_j^{t-1},\ i \in E,\ t \in \mathbb{N}.$$

Then,

$$y_i^t = \sum_j p_{ij}^{t+1}(R_t) = sup_R \sum_j p_{ij}^{t+1}(R),\ i \in E,\ t \in \mathbb{N}_0,$$

where

$R_t := (f_t, f_{t-1}, \ldots, f_2, f_1, f_1, \ldots)$, $t \geq 1$, *and* R_0 *is an arbitrary policy.*

PROOF. We will apply induction on t.

t = 0: For any policy R and any initial state i we have $\Sigma_j\, p^1_{ij}(R) = 1$. Hence,

$$1 = y^0_i = \sum_j p^1_{ij}(R_0) = sup_R \sum_j p^1_{ij}(R), \quad i \in E.$$

Suppose that the result is correct for t = 1,2,..., T-1. We shall show that the lemma is also true for t = T. Take any $i \in E$. By corollary 2.5.1, it is sufficient to show that $y^t_i = \Sigma_j\, p^{t+1}_{ij}(R_t) = sup_{R \in \mathcal{C}_M}\, \Sigma_j\, p^{t+1}_{ij}(R)$.
Take any arbitrary $R = (\pi^1, \pi^2, \ldots) \in \mathcal{C}_M$. Then, we obtain

$$\sum_j p^{T+1}_{ij}(R) = \sum_j \sum_k p_{ik}(\pi^1)\, p^T_{kj}(\pi^2, \pi^3, \ldots) \leq \sum_k p_{ik}(\pi^1)\, y^{T-1}_k \leq y^T_i .$$

Since R is an arbitrarily chosen policy, we obtain

$$y^T_i \geq sup_R \sum_j p^{T+1}_{ij}(R).$$

On the other hand,

$$y^T_i = max_a \sum_k p_{iak} y^{T-1}_k = \sum_k p_{ik}(f_T)\, y^{T-1}_k = \sum_k p_{ik}(f_T) \sum_j p^T_{kj}(R_{T-1})$$

$$= \sum_j p^{T+1}_{ij}(R_T) \leq sup_R \sum_j p^{T+1}_{ij}(R).$$

Hence,

$$y^T_i = \sum_j p^{T+1}_{ij}(R_T) = sup_R \sum_j p^{T+1}_{ij}(R). \qquad \square$$

THEOREM 3.2.4. *The following five statements are equivalent.*

(i) *Every pure and stationary policy is transient.*

(ii) *Every policy is transient.*

(iii) $max_i\, y^N_i < 1$, *where* y^N *is defined by* (3.2.3).

(iv) *The TMD-model is contracting.*

(v) *The linear programming problem*

$$\max\left\{\sum_i\sum_a x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij} - p_{iaj})x_{ia} \le \beta_j & j \in E \\ x_{ia} \ge 0 & i \in E,\ a \in A(i) \end{array}\right\}$$

where $\beta_j > 0$, $j \in E$, *are arbitrarily chosen,*

has a finite solution.

REMARK 3.2.2. The equivalence of the first three statements has been proven by VEINOTT [1969] for nonrandomized policies. HORDIJK [1976] has shown the equivalence of the first four statements for general policies. The equivalence between (i) and (v) is also established by DENARDO & ROTHBLUM [1979].

PROOF OF THEOREM 3.2.4.
(i) ⇒ (ii): Let i and j be two arbitrarily chosen states. Consider the dynamic programming problem with the rewards

$$r_{ka} := \begin{cases} 1 & k = j,\ a \in A(k) \\ 0 & k \ne j,\ a \in A(k). \end{cases}$$

Then, for any policy R, we have

$$\begin{aligned} v_i(R) &= \sum_{t=1}^{\infty}\sum_k\sum_a \mathbb{P}_R(X_t = k, Y_t = a \mid X_1 = i)\cdot r_{ka} \\ &= \sum_{t=1}^{\infty} \mathbb{P}_R(X_t = j \mid X_1 = i). \end{aligned}$$

Let f^∞ be a pure and stationary optimal policy (the existence of f^∞ is implied by theorem 3.2.1). Since we have assumed that f^∞ is a transient policy, we obtain

$$\sum_{t=1}^{\infty} \mathbb{P}_R(X_t = j \mid X_1 = i) = v_i(R) \le v_i = v_i(f^\infty) = \sum_{t=1}^{\infty} \mathbb{P}_{f^\infty}(X_t = j \mid X_1 = i) < \infty,$$

i.e. R is a transient policy.
(ii) ⇒ (iii): By lemma 3.2.2, it is sufficient to show that $\Sigma_j\, p_{ij}^{N+1}(R) < 1$ for all $i \in E$ and all policies $R = (f_1, f_2, \ldots)$, where $f_t^\infty \in C_D$, $t \in \mathbb{N}$. Consider the extended TMD-model. Then $\Sigma_j\, p_{iaj} = 1$ for all $a \in \tilde{A}(i)$, $i \in \tilde{E}$. Since $\tilde{A}(0) = \{1\}$ and $\tilde{p}_{010} = 1$, $\tilde{r}_{01} = 0$, any policy R that is defined for the original model corresponds uniquely to a policy $\tilde{R}$ in the extended model

and $v_i(R) = \tilde{v}_i(\tilde{R})$, $i \in E$, where $\tilde{v}_i(\tilde{R})$ is the expected total reward in the extended model. Take any $i \in E$ and choose any policy $R = (f_1, f_2, \ldots)$, where $f_t^\infty \in C_D$, $t \in \mathbb{N}$. For $k = 1,2,\ldots$ we define subsets T_k of the state space $\tilde{E}$ by

$$T_1 := \{i\}$$

$$T_k := \{j \in \tilde{E} \mid p_{ij}^k(\tilde{R}) > 0\} \quad k = 2,3,\ldots .$$

For the proof that statement (iii) follows from statement (ii) we need the following three propositions.

<u>PROPOSITION 1</u>. If, for any integer n such that $1 \le n \le N$, $0 \notin \bigcup_{\ell=1}^{n} T_\ell$ implies that $T_{n+1} \not\subset \bigcup_{\ell=1}^{n} T_\ell$, then statement (iii) holds.

<u>PROOF</u>. Since state 0 is an absorbing state, $0 \in \bigcup_{\ell=1}^{n} T_\ell$ implies that $0 \in T_{n+1}$. Suppose that $0 \notin \bigcup_{\ell=1}^{N} T_\ell$. Then $0 \notin \bigcup_{\ell=1}^{n} T_\ell$ for $n = 1,2,\ldots,N$. Then, by the assumption of the proposition, we have that $\bigcup_{\ell=1}^{n+1} T_\ell$ has at least one state more than $\bigcup_{\ell=1}^{n} T_\ell$ for all $n = 1,2,\ldots,N$. Consequently, $\bigcup_{\ell=1}^{N+1} T_\ell = \tilde{E}$ which implies that $0 \in T_{N+1}$. Hence,

$$\textstyle\sum_j p_{ij}^{N+1}(R) = 1-\tilde{p}_{i0}^{N+1}(\tilde{R}) < 1, \quad \text{i.e. statement (iii) holds.}$$

<u>PROPOSITION 2</u>. Suppose that the integer n is such that $1 \le n \le N$, $0 \notin \bigcup_{\ell=1}^{n} T_\ell$ and $T_{n+1} \subset \bigcup_{\ell=1}^{n} T_\ell$. Let the pure and stationary policy $\tilde{f}^\infty$ be defined by

$$\tilde{f}(j) := \begin{cases} f_k(j) & \text{if } j \in T_k \setminus \bigcup_{\ell=1}^{k-1} T_\ell \\ \text{arbitrarily chosen} & \text{if } j \notin \bigcup_{\ell=1}^{n} T_\ell . \end{cases}$$

Define $T_1^\star := \{i\}$ and $T_k^\star := \{j \in \tilde{E} \mid \tilde{p}_{ij}^k(\tilde{f}^\infty) > 0\}$ $k = 2,3,\ldots$.
Then,

$$T_k^\star \subset \bigcup_{\ell=1}^{n} T_\ell, \quad k \in \mathbb{N}.$$

<u>PROOF</u>. The proof is given by induction on k.

$k = 1$: $T_1^\star = T_1 \subset \bigcup_{\ell=1}^{n} T_\ell$.

Suppose that $T_k^* \subset \cup_{\ell=1}^n T_\ell$, $k = 1,2,\dots,m$. Take any $j \in T_{m+1}^*$. Then, there exists a state $s \in T_m^*$ such that $p_{s\tilde{f}(s)j} > 0$. Since $s \in \cup_{\ell=1}^n T_\ell$, we have $\tilde{f}(s) = f_k(s)$ where k satisfies $s \in T_k \backslash \cup_{\ell=1}^{k-1} T_\ell$.
From $s \in T_k$ and $\tilde{f}(s) = f_k(s)$ it follows that

$$p_{ij}^{k+1}(\tilde{R}) \ge p_{is}^k(\tilde{R}) \cdot p_{s\tilde{f}(s)j} > 0.$$

Hence,

$$j \in T_{k+1} \subset \cup_{\ell=1}^{n+1} T_\ell = \cup_{\ell=1}^n T_\ell,$$

which completes the proof that $T_{m+1}^* \subset \cup_{\ell=1}^n T_\ell$.

<u>PROPOSITION 3</u>. Suppose that we have the same assumptions as in proposition 2. Then, policy f^∞ is nontransient.

<u>PROOF</u>. Since $0 \notin \cup_{\ell=1}^n T_\ell$ and $T_k^* \subset \cup_{\ell=1}^n T_\ell$ for all $k \in \mathbb{N}$, we have $p_{i0}^k(\tilde{f}) = 0$, $k \in \mathbb{N}$. Consequently, $\Sigma_j\, p_{ij}^k(f) = 1$ for all $k \in \mathbb{N}$.
Hence,

$$\sum_{t=1}^\infty \sum_j \mathbb{P}_{f^\infty}(X_t = j \mid X_1 = i) = +\infty,$$

implying that the pure and stationary policy f^∞ is nontransient.

We can complete the proof of statement (iii) as follows. Statement (ii) implies that any policy is transient. Then, by proposition 3, the assumptions of proposition 2 are not satisfied. Therefore, by proposition 1, statement (iii) holds.

<u>(iii) ⇒ (iv)</u>: Let $a := \max_i y_i^N$ and $b := a^{1/(N+1)}$. Then, $a \le b < 1$. Take α such that $b < \alpha < 1$ and define the vector μ by

$$\mu_i := \sup_R \sum_{t=1}^\infty (1/\alpha)^{t-1} \sum_j \mathbb{P}_R(X_t = j \mid X_1 = i), \quad i \in E.$$

From lemma 3.2.2 it follows that

$$a = \max_i \sup_R \sum_j p_{ij}^{N+1}(R) = \max_i \max_{R\in C_M} \sum_j p_{ij}^{N+1}(R)$$
$$= \max_{R\in C_M} \max_i \sum_j p_{ij}^{N+1}(R) = \max_{R\in C_M} \| P^{N+1}(R) \| .$$

Hence, for any policy $R \in C_M$ and any $t \in \mathbb{N}$, we may write

$$\| P^t(R) \| \le \| P^{\lfloor t/(N+1) \rfloor \cdot (N+1)}(R) \| \le a^{\lfloor t/(N+1) \rfloor} \le a^{-1} \cdot b^t.$$

Consequently,

$$\sum_{t=1}^{\infty}(1/\alpha)^{t-1}\sum_j \mathbb{P}_R(X_t = j \mid X_1 = i) \le$$

$$\sum_{t=1}^{\infty}(1/\alpha)^{t-1}\|P^t(R)\| \le a^{-1}b\cdot \sum_{t=1}^{\infty}(b/\alpha)^{t-1} = \frac{\alpha b}{a(\alpha-b)} .$$

Therefore, μ_i is well-defined, $i \in E$.
Similarly to the proof of theorem 2.5.2 it can be shown that

$$\mu_i = max_a\{1+\frac{1}{\alpha}\sum_j p_{iaj}\mu_j\}, \quad i \in E.$$

Then, we obtain

$$\alpha\mu_i \ge \alpha + \sum_j p_{iaj}\mu_j \ge \sum_j p_{iaj}\mu_j, \quad a \in A(i),\ i \in E,$$

i.e. the TMD-model is contracting.
(iv) ⇒ (v): Suppose that the linear program has no finite solution. Since the linear program is feasible (for instance $x = 0$ is a feasible solution), the optimum value is in infinity. Then, from the theory of linear programming it follows that there exists a vector $s \ne 0$ such that

$$(3.2.4) \qquad s_{ia} \ge 0,\ a \in A(i),\ i \in E, \text{ and } \sum_i\sum_a(\delta_{ij}-p_{iaj})s_{ia} \le 0, \quad j \in E.$$

Define the stationary policy π^∞ by

$$(3.2.5) \qquad \pi_{ia} := \begin{cases} s_{ia}/s_i & a \in A(i),\ i \in E_s \\ \text{arbitrarily} & a \in A(i),\ i \in E\backslash E_s. \end{cases}$$

From (3.2.4) it follows that

$$0 \le s_j = \sum_a s_{ja} \le \sum_i\sum_a p_{iaj}s_{ia} = \sum_i(\sum_a p_{iaj}\pi_{ia})\cdot s_i = \sum_i p_{ij}(\pi)\cdot s_i, \ j \in E,$$

or in vector notation

$$(3.2.6) \qquad 0 \le s^T \le s^T P(\pi).$$

By iterating (3.2.6), we obtain

(3.2.7) $\quad 0 \le s^T \le s^T P^n(\pi) \qquad n \in \mathbb{N}.$

Since the dynamic programming problem is contracting, there exists a vector $\mu \gg 0$ and a real $\alpha \in [0,1)$ such that

$$\sum_j p_{iaj}\mu_j \le \alpha\, \mu_i \qquad a \in A(i),\ i \in E.$$

Hence,

$$0 \le P(\pi)\mu \le \alpha\, \mu$$

and consequently,

$$0 \le P^n(\pi)\mu \le \alpha^n\mu \qquad \text{for all } n \in \mathbb{N},$$

implying that $P^n(\pi) \to 0$ for $n \to \infty$.
Then, from relation (3.2.7), it follows that $s = 0$, which gives a contradiction. This completes the proof of statement (v).
(v) ⇒ (i): Suppose that statement (i) is not true. Then, there exists a pure and stationary policy f^∞ such that

$$\sum_{t=1}^{\infty} \mathbb{P}_{f^\infty}(X_t = j \mid X_1 = i) = +\infty \qquad \text{for certain } i,j \in E.$$

Then, we obtain

(3.2.8) $\quad \sum_{t=1}^{\infty} \beta^T P^{t-1}(f)e = \sum_\ell \beta_\ell \cdot \sum_{t=1}^{\infty} \sum_k p_{\ell k}^{t-1}(f) = +\infty.$

Consider the sequence $\{x^n,\ n = 1,2,\dots\}$, defined by

$$x_{ia}^n := \begin{cases} \sum_{t=1}^{n} [\beta^T P^{t-1}(f)]_i & a = f(i) \\ 0 & a \ne f(i) \end{cases} \qquad n \in \mathbb{N}.$$

Vector x^n has the following properties:

1. $x_{ia}^n \ge 0 \quad a \in A(i),\ i \in E.$

2. $\Sigma_i\Sigma_a(\delta_{ij}-p_{iaj})x^n_{ia} = \Sigma_i(\delta_{ij}-p_{ij}(f))\Sigma^n_{t=1}\Sigma_\ell\beta_\ell p^{t-1}_{\ell i}(f) =$
$\Sigma_\ell\beta_\ell\Sigma^n_{t=1}\Sigma_i p^{t-1}_{\ell i}(f)(\delta_{ij}-p_{ij}(f)) = \Sigma_\ell\beta_\ell\Sigma^n_{t=1}\{p^{t-1}_{\ell j}(f)-p^t_{\ell j}(f)\} =$
$\Sigma_\ell\beta_\ell\{\delta_{\ell j}-p^n_{\ell j}(f)\} = \beta_j-\Sigma_\ell\beta_\ell p^n_{\ell j}(f) \le \beta_j, \quad j \in E.$
3. $\Sigma_i\Sigma_a x^n_{ia} = \Sigma_i\Sigma^n_{t=1}[\beta^T P^{t-1}(f)]_i = \Sigma^n_{t=1}\beta^T P^{t-1}(f)e.$

Hence, we have a sequence $\{x^n, n = 1,2,\ldots\}$ of feasible solutions such that $\Sigma_i\Sigma_a x^n_{ia} \to +\infty$ for $n \to \infty$. This contradicts the assumption that the linear program has a finite solution. Therefore, we have shown that statement (i) is true. □

ment (i) is true. □

The characterizations (iii) and (v) of theorem 3.2.4 give two finite algorithms in order to check the contraction property for a given Markov decision problem. Below we present these algorithms.

ALGORITHM IV for the verification of the contraction property for a Markov decision problem (iterative approach).

step 1: $t := 0$; $y^0_i := 1$, $i \in E$.
step 2: $y^{t+1}_i := max_a\ \Sigma_j\ p_{iaj}y^t_j$, $i \in E$.
step 3: If $max_i\ y^{t+1}_i < 1$, then the problem is contracting (STOP), otherwise, go to step 4.
step 4: If $t = N+1$, then the problem is not contracting (STOP), otherwise, $t := t+1$ and go to step 2.

ALGORITHM V for the verification of the contraction property for a Markov decision problem (linear programming approach).

step 1: Take any vector β such that $\beta_j > 0$, $j \in E$.
step 2: Solve the linear programming problem

$$max\left\{\sum_i\sum_a x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} \le \beta_j & j \in E \\ x_{ia} \ge 0 & a \in A(i),\ i \in E \end{array}\right\}.$$

If the linear program has a finite solution, then the problem is contracting (STOP).
Otherwise, the problem is not contracting (STOP).

REMARK 3.2.3. If we use algorithm V and the algorithm shows that the problem is contracting, then we can obtain, from the dual program, a vector $\mu \gg 0$ and a scalar $\alpha \in [0,1)$ such that

$$\sum_j p_{iaj}\mu_j \leq \alpha\, \mu_i \qquad a \in A(i),\ i \in E.$$

Namely: The dual linear program is

$$\min\left\{\sum_j \beta_j \mu_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij}-p_{iaj})\mu_j \geq 1 & a \in A(i),\ i \in E \\ \mu_j \geq 0 & j \in E \end{array}\right\},$$

and has also an optimal solution, say μ. Then we have

$$\mu_i \geq 1 + \sum_j p_{iaj}\mu_j > 0, \qquad i \in E$$

and for $\alpha := 1 - (\max_i \mu_i)^{-1}$ we have $\alpha \in [0,1)$ and

$$\sum_j p_{iaj}\mu_j \leq \mu_i - 1 \leq \mu_i - \frac{\mu_i}{\max_i \mu_i} = \alpha\mu_i \qquad a \in A(i),\ i \in E.$$

3.3. OPTIMAL TRANSIENT POLICIES

In this section we discuss the problem of finding an optimal policy in the class of the transient policies, i.e. a policy R^* such that

$$(3.3.1) \qquad v_i(R^*) = \sup\{v_i(R) \mid R \text{ is a transient policy}\}, \qquad i \in E.$$

Such a policy may be of interest, for instance in the so-called optimal stopping problem (see application 3.3.1 at the end of this section). A related optimal stopping problem, whose utility function is exponential, is discussed by DENARDO & ROTHBLUM [1979]. The problem of finding an optimal policy in the class of the transient policies can also be solved for models with $\Sigma_j\, p_{iaj} > 1$ for some $i \in E$, $a \in A(i)$ (cf.HORDIJK & KALLENBERG [1981a]). Another related paper is HORDIJK & KALLENBERG [1981c].

Any policy is transient in a contracting dynamic programming problem. In that case a policy which satisfies (3.3.1) is an optimal policy in the class of all policies. In general, the problem of finding an optimal transient policy is only relevant if there exists at least one transient policy.

Therefore, we introduce the following assumption.

ASSUMPTION 3.3.1. *There exists a transient policy.*

Further on, we will show how, for a given problem, this assumption can be verified by linear programming.

The total expected reward of any transient policy is finite. However, the vector w, where

$$(3.3.2) \qquad w_i := \sup\{v_i(R) \mid R \text{ is a transient policy}\}, \qquad i \in E,$$

is not necessarily finite.

EXAMPLE 3.3.1. Consider the model of figure 3.3.1. The sequence $\{\pi^\infty(n), n = 1,2,\ldots\}$ of stationary policies defined by

$$\pi_{1a}(n) := \begin{cases} 1 - 1/n & a = 1 \\ 1/n & a = 2 \end{cases}, \qquad \pi_{21}(n) := 1, \qquad n \in \mathbb{N}$$

satisfies:

$$\sum_{t=1}^{\infty} \mathbb{P}_{\pi^\infty(n)}(X_t = j \mid X_1 = i) = \begin{cases} n & i=1, j=1 \\ 2 & i=1, j=2 \\ 0 & i=2, j=1 \\ 2 & i=2, j=2 \end{cases}$$

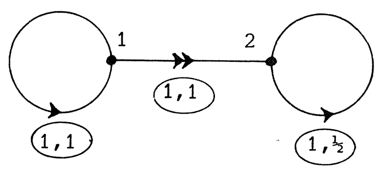


Figure 3.3.1

$$v_i(\pi^\infty(n)) = \begin{cases} n + 2 & i = 1 \\ 2 & i = 2. \end{cases}$$

Hence, every policy $\pi^\infty(n)$ is transient, but $w_1 \geq \sup_n v_1(\pi^\infty(n)) = +\infty$.

THEOREM 3.3.1. *If* w *is finite, then* w *is a solution of the functional equation*

$$x_i = \max_a \{r_{ia} + \sum_j p_{iaj} x_j\}, \qquad i \in E.$$

PROOF. Let $R = (\pi^1, \pi^2, \ldots)$ be an arbitrary transient Markov policy. Then,

$$v_i(R) = r_i(\pi^1) + \sum_j p_{ij}(\pi^1) u_j(R), \qquad i \in E,$$

where $u_j(R)$ represents the expected total reward earned from time 2, given that the state at time 2 is j. Let $\tilde{R} := (\pi^2,\pi^3,\ldots)$, then we can write

$$\infty > \sum_{t=2}^{\infty} \mathbb{P}_R(X_t = k \mid X_1 = i) = \sum_j p_{ij}(\pi^1)\sum_{t=2}^{\infty} \mathbb{P}_{\tilde{R}}(X_{t-1} = k \mid X_1 = j) \quad i,k \in E.$$

Hence, $\Sigma_{t=1}^{\infty} \mathbb{P}_{\tilde{R}}(X_t = k \mid X_1 = j) < \infty$ for every k and every j with $\Sigma_i p_{ij}(\pi^1) > 0$. Therefore, we have

$$u_j(R) = v_j(\tilde{R}) \le w_j \text{ for all j such that } p_{ij}(\pi^1) > 0 \text{ for some } i \in E.$$

Then, we obtain

$$v_i(R) \le \sum_a \pi^1_{ia}\{r_{ia} + \sum_j p_{iaj}w_j\} \le \max_a\{r_{ia} + \sum_j p_{iaj}w_j\}, \quad i \in E.$$

Theorem 2.5.1 and the fact that R is arbitrarily chosen imply that

(3.3.3) $\quad w_i \le \max_a\{r_{ia} + \sum_j p_{iaj}w_j\}, \quad i \in E.$

Take any $\varepsilon > 0$. Suppose that for every $j \in E$, $R_j := (\pi^1(j),\pi^2(j),\ldots)$ is a transient policy that satisfies $v_j(R_j) \ge w_j - \varepsilon$. Again, by theorem 2.5.1, we may assume that R_j is a Markov policy. Let $a_i \in A(i)$, $i \in E$, be such that

$$r_{ia_i} + \sum_j p_{ia_ij}w_j = \max_a\{r_{ia} + \sum_j p_{iaj}w_j\}, \quad i \in E.$$

Let $\hat{R} = (\pi^1,\pi^2,\ldots)$ be the policy with $\pi^t_{i_1a_1\ldots i_ta_t} := \pi^{t-1}_{i_ta_t}(i_2)$, $t \ge 2$, and
$\pi^1_{ia} := \begin{cases} 1 & a = a_i, \quad i \in E \\ 0 & a \ne a_i, \quad i \in E. \end{cases}$

Hence, policy $\hat{R}$ is transient and we obtain

(3.3.4) $\quad w_i \ge v_i(\hat{R}) = r_{ia_i} + \sum_j p_{ia_ij}v_j(R_j) \ge r_{ia_i} + \sum_j p_{ia_ij}(w_j - \varepsilon) \ge$

$$\max_a\{r_{ia} + \sum_j p_{iaj}w_j\} - \varepsilon, \quad i \in E.$$

Since ε is arbitrarily chosen, (3.3.3) and (3.3.4) imply that

$$w_i = \max_a\{r_{ia} + \sum_j p_{iaj}w_j\}, \quad i \in E. \quad \square$$

EXAMPLE 3.3.2. $E = \{1\}$; $A(1) = \{1,2\}$; $p_{111} = 1$, $p_{121} = \frac{1}{2}$; $r_{11} = 0$, $r_{12} = -1$. It is easy to see that $w_1 = -2$; the functional equation is $x_1 = \max\{x_1, -1 + \frac{1}{2}x_1\}$ with solution set $\{x_1 \mid x_1 \geq -2\}$. Hence, the solution of the functional equation is not unique.

DEFINITION 3.3.1. A vector $\tilde{w} \in \mathbb{R}^N$ is *TMD-superharmonic if*

$$\tilde{w}_i \geq r_{ia} + \sum_j p_{iaj}\tilde{w}_j, \quad a \in A(i),\ i \in E.$$

THEOREM 3.3.2. *Suppose that* w *is finite. Then,* w *is the smallest TMD-superharmonic vector.*

PROOF. Theorem 3.3.1 implies that w is TMD-superharmonic. Suppose that $\tilde{w}$ is also a TMD-superharmonic vector. From theorem 2.5.1 it follows that it is sufficient to prove that $\tilde{w} \geq v(R)$ for any transient Markov policy R. Let $R = (\pi^1, \pi^2, \ldots)$ be an arbitrary transient Markow policy. Since $\tilde{w}$ is TMD-superharmonic, we have

$$\tilde{w} \geq r(\pi^t) + P(\pi^t)\tilde{w}, \quad t \in \mathbb{N}. \tag{3.3.5}$$

By iterating (3.3.5), we obtain

$$\tilde{w} \geq \sum_{t=1}^n P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t) + P(\pi^1)P(\pi^2)\cdots P(\pi^n)\tilde{w}, \quad n \in \mathbb{N}.$$

Because R is a transient Markov policy

$$P(\pi^1)P(\pi^2)\cdots P(\pi^n) \to 0 \quad \text{for } n \to \infty$$

and

$$v(R) = \lim_{n\to\infty} \sum_{t=1}^n P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t).$$

Consequently,

$$\tilde{w} \geq v(R),$$

which completes the proof of the theorem. □

Theorem 3.3.2 implies that, if w is finite, then w is the unique optimal solution of the linear programming problem

(3.3.6) $\quad min\{\sum_j \beta_j \tilde{w}_j \mid \sum_j (\delta_{ij} - p_{iaj})\tilde{w}_j \geq r_{ia}, \quad a \in A(i),\ i \in E\}$

where $\beta_j > 0$, $j \in E$, are given numbers.
The dual linear programming problem is:

$$(3.3.7) \qquad max\left\{\sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \end{array}\right\}.$$

Notice that any feasible solution x of program (3.3.7) satisfies

$$x_j := \sum_a x_{ja} = \beta_j + \sum_i \sum_a p_{iaj} x_{ia} \geq \beta_j > 0, \quad j \in E.$$

We define for any feasible solution x of program (3.3.7) a stationary policy $\pi^\infty(x)$ by

$$(3.3.8) \qquad \pi_{ia}(x) := x_{ia}/x_i \qquad a \in A(i),\ i \in E.$$

Since $x_{ia} = \pi_{ia}(x) \cdot x_i$, $a \in A(i)$, $i \in E$, we can write

$$\sum_i \sum_a (\delta_{ij} - p_{iaj}) \pi_{ia}(x) \cdot x_i = \beta_j, \quad j \in E.$$

Hence, we have

$$(3.3.9) \qquad x^T = \beta^T + x^T P(\pi(x)).$$

By iterating (3.3.9), we obtain

$$x^T = \sum_{t=1}^n \beta^T P^{t-1}(\pi(x)) + x^T P^n(\pi(x)) \geq \sum_{t=1}^n \beta^T P^{t-1}(\pi(x)), \quad n \in \mathbb{N}.$$

Hence,

$$\sum_{t=1}^\infty \beta^T P^{t-1}(\pi(x)) < \infty,$$

and consequently,

$$\sum_{t=1}^\infty \mathbb{P}_{\pi^\infty(x)}(X_t = j \mid X_1 = i) = \sum_{t=1}^\infty [P^{t-1}(\pi(x))]_{ij} < \infty, \quad i,j \in E.$$

So, the policy $\pi^\infty(x)$ is transient and therefore we can write (cf. KEMENY & SNELL [1960] p.22)

$$x^T = \beta^T(I-P(\pi(x)))^{-1}. \tag{3.3.10}$$

Conversely, let π^∞ be any transient stationary policy. Then, the inverse $(I-P(\pi))^{-1}$ exists. We define the vector $x(\pi)$ by

$$x_{ia}(\pi) := [\beta^T(I-P(\pi))^{-1}]_i \cdot \pi_{ia}, \quad a \in A(i),\ i \in E. \tag{3.3.11}$$

THEOREM 3.3.3. *The mapping defined by* (3.3.11) *is a one-to-one mapping of the transient stationary policies onto the set of feasible solutions of the linear program* (3.3.7) *with* (3.3.8) *as the inverse mapping. Furthermore, the set of extreme feasible solutions of program* (3.3.7) *corresponds to the transient stationary policies which are pure.*

PROOF. First, we prove that $x(\pi)$ is a feasible solution of program (3.3.7). Let π^∞ be an arbitrarily chosen transient stationary policy. Then $x(\pi)$ satisfies

1. $x_{ia}(\pi) = [\beta^T(I-P(\pi))^{-1}]_i \cdot \pi_{ia} = [\beta^T\Sigma_{t=1}^{\infty}P^{t-1}(\pi)]_i \cdot \pi_{ia} \geq 0,\ a \in A(i), i \in E.$

2. $$\begin{aligned}\Sigma_i\Sigma_a(\delta_{ij}-p_{iaj})x_{ia}(\pi) &= \Sigma_a x_{ja}(\pi) - \Sigma_i\Sigma_a p_{iaj}x_{ia}(\pi)\\ &= [\beta^T(I-P(\pi))^{-1}]_j - [\beta^T(I-P(\pi))^{-1}P(\pi)]_j\\ &= [\beta^T(I-P(\pi))^{-1}(I-P(\pi))]_j = \beta_j, \quad j \in E.\end{aligned}$$

Hence, $x(\pi)$ is a feasible solution of (3.3.7). From (3.3.10) and (3.3.11) it follows that $x = x(\pi(x))$, implying that the mapping is onto. Since $\pi_{ia}(x(\pi)) = \pi_{ia}$, $a \in A(i)$, $i \in E$, the mapping is one-to-one and the inverse mapping is given by (3.3.8).

Let f^∞ be an arbitrarily chosen pure and stationary transient policy. Suppose that $x(f)$ is not an extreme feasible solution. Then, there exist feasible solutions x^1 and x^2 of program (3.3.7) and a real number $\lambda \in (0,1)$ such that $x^1 \neq x^2$ and $x(f) = \lambda x^1+(1-\lambda)x^2$.
Since $x_{ia}(f) = 0$, $a \neq f(i)$, $i \in E$, we also have $x^1_{ia} = x^2_{ia} = 0$, $a \neq f(i)$, $i \in E$. Hence, the N-dimensional vectors $x^1 = (x^1_{if(i)})$ and $x^2 = (x^2_{if(i)})$ are solutions of the linear system $x^T(I-P(f)) = \beta^T$.
Since f^∞ is a transient policy, the matrix $(I-P(f))$ is nonsingular and

consequently, the system has a unique solution, namely $\beta^T(I-P(f))^{-1}$. This implies that $x^1 = x^2$, giving a contradiction. Hence, we have proved that $x(f)$ is an extreme solution.
Conversely, let x be any extreme feasible solution of program (3.3.7). Since N is the number of constraints in program (3.3.7), x has at most N positive components. On the other hand, it follows from

$$\sum_a x_{ja} = \beta_j + \sum_i \sum_a p_{iaj} x_{ia} > 0, \quad j \in E,$$

that in each state j there is at least one positive component. Consequently, x has in each state j exactly one component which is positive. Hence, the corresponding policy $\pi^\infty(x)$ is a pure policy. □

For a given initial distribution $\beta = (\beta_1, \beta_2, \ldots, \beta_N)^T$, where $\beta_i > 0$ $i \in E$, we denote for any transient policy R the expected number of times of being in state j and then choosing action a by

$$\text{(3.3.12)} \qquad x_{ja}(R) := \sum_i \beta_i \cdot \sum_{t=1}^{\infty} \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i).$$

Since R is a transient policy, we have $x_{ja}(R) < \infty$, $a \in A(j)$, $j \in E$. The definitions (3.3.11) and (3.3.12) imply that

$$x_{ja}(\pi^\infty) = [\beta^T(I-P(\pi))^{-1}]_j \cdot \pi_{ja} = x_{ja}(\pi), \qquad a \in A(j),\ j \in E.$$

NOTATION 3.3.1.

$$K := \{x(R) \mid R \in C \text{ and transient}\}$$

$$K(M) := \{x(R) \mid R \in C_M \text{ and transient}\}$$

$$K(S) := \{x(R) \mid R \in C_S \text{ and transient}\}$$

$$K(D) := \{x(R) \mid R \in C_D \text{ and transient}\}$$

$$P := \left\{ x \,\middle|\, \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \end{array} \right\}.$$

THEOREM 3.3.4. $\overline{K(D)} \subset K(S) = K(M) = K = P$.

PROOF. The equality $K = K(M)$ follows from theorem 2.5.1. Since P is a convex polyhedron, theorem 3.3.3 implies that $\overline{K(D)} \subset P = K(S) \subset K(M) = K$. Therefore, it is sufficient to show that $K(M) \subset P$. Take any $x(R) \in K(M)$ and suppose that $R = (\pi^1, \pi^2, \ldots)$. Then, we obtain

$$\sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia}(R) =$$

$$= \sum_i \sum_a (\delta_{ij} - p_{iaj}) \sum_\ell \beta_\ell \cdot \lim_{n\to\infty} \sum_{t=1}^{n} \{P(\pi^1)\cdots P(\pi^{t-1})\}_{\ell i} \cdot \pi_{ia}^{t}$$

$$= \sum_\ell \beta_\ell \cdot \lim_{n\to\infty} \sum_{t=1}^{n} \sum_i \{P(\pi^1)\cdots P(\pi^{t-1})\}_{\ell i} \cdot (\delta_{ij} - p_{ij}(\pi^t))$$

$$= \sum_\ell \beta_\ell \cdot \lim_{n\to\infty} \sum_{t=1}^{n} \{P(\pi^1)\cdots P(\pi^{t-1}) \cdot (I - P(\pi^t))\}_{\ell j}$$

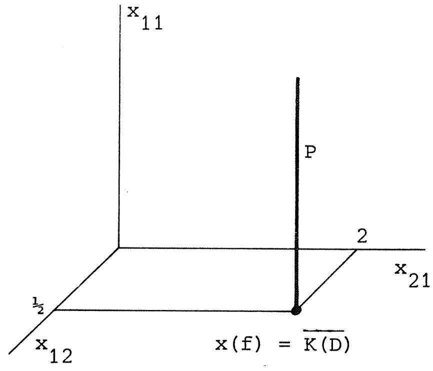
$$= \sum_\ell \beta_\ell \cdot \lim_{n\to\infty} \{I - P(\pi\ )P(\pi^2)\ldots P(\pi^n)\}_{\ell j} = \sum_\ell \beta_\ell \cdot \delta_{\ell j} = \beta_j, \quad j \in E.$$

Hence, $x(R) \in P$, which completes the proof. □

<u>REMARK 3.3.1</u>. The next example shows that $\overline{K(D)} \neq P$ is possible.

<u>EXAMPLE 3.3.3</u>. Consider the model of example 3.3.1 and take $\beta_1 = \beta_2 = \frac{1}{2}$. There is only one transient pure and stationary policy, namely f^∞, where $f(1) = 2$ and $f(2) = 1$. The solution $x(f)$ satisfies $x_{11}(f) = 0$, $x_{12}(f) = 1/2$ and $x_{21}(f) = 2$. The set P is given by

$$P = \left\{ x \;\middle|\; \begin{array}{lr} x_{12} & = \frac{1}{2} \\ -x_{12} + \frac{1}{2} x_{21} & = \frac{1}{2} \\ x_{12}, x_{21}, x_{11} & \geq 0 \end{array} \right\}.$$



<u>Figure 3.3.2</u>

Hence, $\overline{K(D)} \neq P$.

<u>REMARK 3.3.2</u>. Suppose that $K \neq \emptyset$, then also $P \neq \emptyset$. Lemma 1.2.2 implies the existence of an extreme feasible solution of program (3.3.7). Then, by theorem 3.3.3, the existence of a transient pure and stationary policy is shown. This argument provides another proof of theorem 3.2.3.

<u>REMARK 3.3.3</u>. Since assumption 3.3.1 is satisfied if and only if $P \neq \emptyset$, this assumption can be verified by linear programming: we have to check the feasibility of program (3.3.7).

REMARK 3.3.4. If the vector w is finite, then it follows from theorem 3.3.2 that the linear programming problem (3.3.7) has a finite optimum. The following theorem shows that the reverse statement is also true. Furthermore, this theorem proves the correctness of algorithm VI for the determination of an optimal transient policy.

THEOREM 3.3.5. *Let* x^* *be an extreme optimal solution of the linear programming problem* (3.3.7). *Then, the pure and stationary policy* f_*^∞, *where* $f_*(i)$ *satisfies* $x^*_{if_*(i)} > 0$, $i \in E$, *is optimal in the class of transient policies.*

PROOF. In the proof of theorem 3.3.3 we have seen that, from the fact that x^* is an extreme solution, it follows that f_*^∞ is transient and is uniquely determined by the condition $x^*_{if_*(i)} > 0$, $i \in E$.
Since $x^*_{if_*(i)} > 0$ for every $i \in E$, it follows from the complementary slackness property of linear programming that $(I-P(f_*))\ w = r(f_*)$. Hence,

$$w = (I-P(f_*))^{-1}\ r(f_*) = v(f_*^\infty),$$

i.e. f_*^∞ is an optimal transient policy. □

ALGORITHM VI for the construction of an optimal pure and stationary transient policy in a TMD-model.

step 1: Take any vector β such that $\beta_j > 0$, $j \in E$.
step 2: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$max \left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \end{array} \right\}$$

(if the problem is infeasible, then there exists no transient policy; if the problem has an infinite solution, then there exists no optimal transient policy).
step 3: Take f_*^∞ such that $x^*_{if_*(i)} > 0$, $i \in E$.

REMARK 3.3.5. Since any extreme solution x of program (3.3.7) satisfies $x_{ia_i} > 0$ for exactly one $a_i \in A(i)$ for every $i \in E$, the linear program is nondegenerated.
The following example shows that the policy f_*^∞, obtained by algorithm VI, is in general not optimal in the class of all policies.

EXAMPLE 3.3.4. Consider the model of figure 3.3.3. The corresponding linear program is:

$$\max\left\{-x_{11}-x_{21}\;\middle|\;\begin{array}{l}\tfrac{1}{2}x_{11}+x_{12}\qquad\quad -x_{22}=\tfrac{1}{2}\\ \qquad -x_{12}+\tfrac{1}{2}x_{21}+x_{22}=\tfrac{1}{2}\\ x_{11},x_{12},x_{21},x_{22}\geq 0\end{array}\right\}.$$

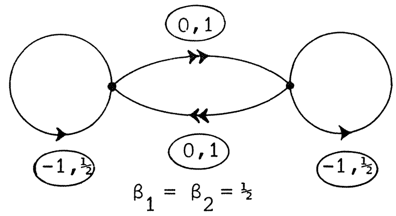


Figure 3.3.3

An extreme optimal solution is ($x^*_{11} = 0$, $x^*_{12} = \tfrac{1}{2}$, $x^*_{21} = 2$, $x^*_{22} = 0$).
The pure and stationary policy f^∞_* satisfies $f_*(1) = 2$, $f_*(2) = 1$.
If can easily be verified that $v_1(f^\infty_*) = v_2(f^\infty_*) = -2$.
However, the policy f^∞ where $f(1) = f(2) = 2$ gives $v_1(f^\infty) = v_2(f^\infty) = 0$.

THEOREM 3.3.6. *The correspondence between the transient stationary policies and the feasible solutions of the linear program preserves the optimality property, i.e.*

1. *if* π^∞ *is a stationary optimal transient policy, then* $x(\pi)$ *is an optimal solution of the dual linear programming problem* (3.3.7).
2. *If* x *is an optimal solution of the linear program* (3.3.7), *then the stationary policy* $\pi^\infty(x)$ *is an optimal transient policy.*

PROOF.

1. Since w is an optimal solution of the primal problem and $x(\pi)$ is feasible for the dual problem, it follows from theorem 1.3.4 that it is sufficient to prove that $\Sigma_i\Sigma_a\, r_{ia}x_{ia}(\pi) = \Sigma_j\beta_j w_j$. We can write

$$\sum_i\sum_a r_{ia}x_{ia}(\pi) = \sum_i\sum_a r_{ia}[\beta^T(I-P(\pi))^{-1}]_i\cdot\pi_{ia}$$

$$= \beta^T(I-P(\pi))^{-1}r(\pi) = \beta^T v(\pi^\infty) = \beta^T w,$$

which completes the proof of this part of the theorem.

2.

$$\beta^T v(\pi^\infty(x)) = \beta^T(I-P(\pi(x)))^{-1}r(\pi(x)) = \sum_i\sum_a r_{ia}x_{ia}(\pi(x))$$

$$= \sum_i\sum_a r_{ia}x_{ia} = \beta^T w.$$

Since $\beta >> 0$ and $v(\pi^\infty(x)) \leq w$, it follows that $v(\pi^\infty(x)) = w$, i.e. $\pi^\infty(x)$ is an optimal transient policy. □

REMARK 3.3.6. Theorem 3.3.6 implies that all optimal pure and stationary transient policies can be determined by the computation of all optimal extreme solutions of the dual program (3.3.7). In chapter 1 such an algorithm is presented (see algorithm I).

We continue this section with a discussion on Markov decision problems under constraints. We suppose that $\beta = (\beta_1,\beta_2,\ldots,\beta_N)^T$ is a known initial distribution such that $\beta_j > 0$ for all $j \in E$. We exclude distributions where $\beta_j = 0$ for some $j \in E$. The reason is that in that case it will in general not be possible to distinguish the transient policies from the nontransient policies (see example 3.3.7). In the unconstrained case we can find a policy R^* that is optimal simultaneously for all initial states $i \in E$. In the constrained case, a policy which is optimal for all initial states does not exist in general (see example 3.3.5). Therefore, we use the concept of optimality with regard to a given initial distribution β.

We consider constraints that are linear functions of $x(R)$, e.g.

$$\sum_i \sum_a q_{iak} x_{ia}(R) \le b_k \quad \text{for the k-th constraint.}$$

Notice that, by formula (3.3.12), the constraints depend on the initial distribution.

Markov decision problems under constraints may be of importance if we are interested in more than one reward function. Then, for instance, we want to maximize one reward function subject to the constraints that the other reward functions are bounded by some given quantities.

Linear programming seems extremely suitable for solving this kind of problems. The other standard techniques to solve unconstrained Markov decision problems (policy improvement and successive approximation) cannot handle these constrained problems. We shall show that there always exists an optimal *stationary* transient policy and we shall present an algorithm to compute one.

EXAMPLE 3.3.5. Consider the model of figure 3.3.4. Suppose that we have one reward function, which is indicated in the figure, and that we have the constraint $v_1(R) + v_2(R) \le 3$. Then we can formulate two constrained problems:

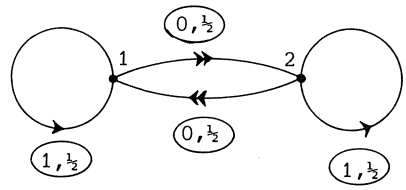


Figure 3.3.4

(1) $sup\{v_1(R) \mid v_1(R) + v_2(R) \le 3\}$ which has as optimal solution f_1^∞, where $f_1(1) = 1$ and $f_1(2) = 2$

(2) $sup\{v_2(R) \mid v_1(R) + v_2(R) \le 3\}$ which has as optimal solution f_2^∞, where $f_2(1) = 2$ and $f_2(2) = 1$.

Hence, there exists no policy which is optimal for both problems simultaneously.

The *constrained Markov decision problem* can be formulated as:

$$(3.3.13) \qquad sup\left\{ \sum_i \beta_i v_i(R) \;\middle|\; \begin{array}{ll} \sum_i \sum_a q_{iak} x_{ia}(R) \le b_k & k = 1,2,\dots,m \\ & R \text{ is transient} \end{array} \right\}.$$

In order to solve problem (3.3.13) we consider the following linear programming problem:

$$(3.3.14) \qquad max\left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j & & j \in E \\ \sum_i \sum_a q_{iak} x_{ia} & \le b_k & k = 1,2,\dots,m \\ x_{ia} & \ge 0 & a \in A(i),\ i \in E \end{array} \right\}$$

THEOREM 3.3.7.

(i) *Problem* (3.3.13) *is feasible if and only if problem* (3.3.14) *is feasible.*

(ii) *The optima of the problems* (3.3.13) *and* (3.3.14) *are equal.*

(iii) *If* x *is an optimal solution of the linear program* (3.3.14), *then* $\pi^\infty(x)$ *is an optimal solution of* (3.3.13).

(iv) *If* R *is an optimal solution of problem* (3.3.13), *then* x(R) *is an optimal solution of program* (3.3.14).

PROOF. The proof is straightforward using the following properties:

(1) K = P.

(2) Every transient policy R satisfies $\Sigma_i \beta_i v_i(R) = \Sigma_i \Sigma_a r_{ia} x_{ia}(R)$.

(3) $x = x(\pi^\infty(x))$ for every $x \in P$. □

REMARK 3.3.7. From theorem 3.3.7 it follows that, if the linear program (3.3.14) has a finite optimum, then problem (3.3.13) has an optimal solution that is stationary. The next example shows that, in general, problem (3.3.13) has no optimal solution in the class of pure and stationary policies, even in the case that $\overline{K(D)} = P$.

EXAMPLE 3.3.6. Consider the model of example 3.3.5 with the exception that $r_{11} = 0$. Take $\beta_1 = \beta_2 = \frac{1}{2}$, $m = 1$ and let the constraint be $x_{21}(R) \le \frac{1}{2}$. The polyhedron P is given by

$$P = \left\{ x \;\middle|\; \begin{array}{rl} \frac{1}{2}x_{11} + x_{12} \qquad\qquad - \frac{1}{2}x_{22} &= \frac{1}{2} \\ - \frac{1}{2}x_{12} + \frac{1}{2}x_{21} + x_{22} &= \frac{1}{2} \\ x_{11}, x_{12}, x_{21}, x_{22} &\ge 0 \end{array} \right\}.$$

We have drawn the polyhedron P in the 3-dimensional space with coordinates x_{12}, x_{21} and x_{22} (x_{11} is given by $x_{11} = 1 + x_{22} - 2x_{12}$). Let f_1^∞, f_2^∞, f_3^∞ and f_4^∞ be defined by: $f_1(1)=f_1(2)=1$; $f_2(1)=1$, $f_2(2)=2$; $f_3(1)=2$, $f_3(2)=1$; $f_4(1)=f_4(2)=2$. The vectors $x(f_k)$ are denoted in figure 3.3.5, $k = 1,2,3,4$. Since the objective function is x_{21}, it can also be seen in the picture that the linear program has no optimal solution which corresponds to a pure policy.

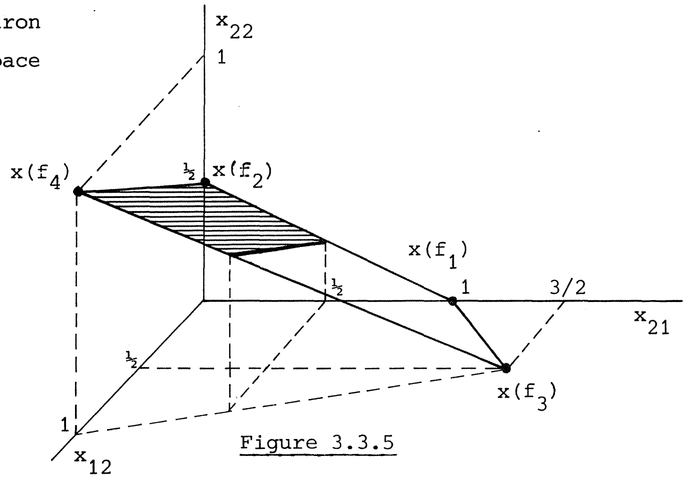


Figure 3.3.5

ALGORITHM VII for the construction of an optimal stationary transient policy in a contrained TMD-model with initial distribution $\beta \gg 0$.

step 1: Compute an optimal solution x^* of the linear programming problem

$$\max \left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j & & j \in E \\ \sum_i \sum_a q_{iak} x_{ia} \le b_k & & k = 1,2,\ldots,m \\ x_{ia} \ge 0 & & a \in A(i),\ i \in E \end{array} \right\}.$$

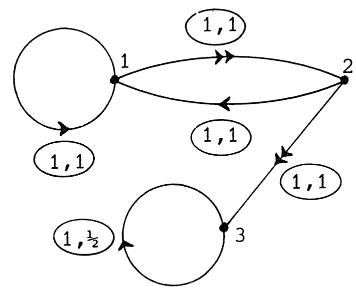
(if the program is infeasible, then the constrained TMD-problem is also infeasible; if the program has an infinite solution, then there exists no optimal transient policy).

step 2: Take ${}_*\pi^\infty$ such that ${}_*\pi_{ia} := x^*_{ia} / \Sigma_a x^*_{ia}$, $a \in A(i)$, $i \in E$.

REMARK 3.3.8. Since $x_j = \Sigma_a x_{ja} = \beta_j + \Sigma_i \Sigma_a p_{iaj} x_{ia} > 0$ for every $j \in E$, the policy $_*\pi^\infty$ is well-defined in step 2 of the algorithm. The correctness of algorithm VII is a consequence of theorem 3.3.7.

REMARK 3.3.9. If we allow that $\beta_j = 0$ for some $j \in E$, then we can loose the one-to-one correspondence between the stationary transient policies and the feasible solutions of the dual linear program (3.3.7). Furthermore, we can obtain nontransient policies, as is shown in the next example.

EXAMPLE 3.3.7. The problem is given by figure 3.3.6. Suppose that we have the constraint $-x_{12}(R) \le -\frac{1}{2}$. Then the linear program is as follows:



$\beta_1 = 0,\ \beta_2 = 0,\ \beta_3 = 1$

Figure 3.3.6

$$\max\left\{ x_{11}+x_{12}+x_{21}+x_{22}+x_{31} \;\middle|\; \begin{array}{lr} x_{12}-x_{21} & =0 \\ -x_{12}+x_{21}+x_{22} & =0 \\ \quad -x_{22}+\tfrac{1}{2}x_{31} & =1 \\ -x_{12} & \le -\tfrac{1}{2} \\ x_{11},x_{12},x_{21},x_{22},x_{31} & \ge 0 \end{array} \right\}$$

An extreme solution is: $(x^*_{11} = 0,\ x^*_{12} = x^*_{21} = \frac{1}{2},\ x^*_{22} = 0,\ x^*_{31} = 2)$.
The corresponding policy f_*^∞, where $f_*(1) = 2$, $f_*(2) = 1$, $f_*(3) = 1$, is nontransient.

APPLICATION 3.3.1. Optimal stopping problem.
In an optimal stopping problem we have two possible actions in each state. The first action corresponds with stopping and if the second action is chosen, then the process proceeds. If the stopping action is chosen in state i, then a final reward r_i is earned and the process breaks down, $i \in E$. If the second action is chosen in state i, then we receive a reward c_i and the probability of being in state j at the next time point is p_{ij}, $i,j \in E$. Our aim is to find an optimal transient policy. It is obvious that there exists a transient policy, namely the policy f^∞ where $f(i) = 1$, $i \in E$. The primal and dual linear programming problems for the optimal stopping problem are:

$$\min\left\{ \sum_j \beta_j \tilde{w}_j \;\middle|\; \begin{array}{rl} \tilde{w}_i \ge r_i & i \in E \\ \sum_j (\delta_{ij}-p_{ij})\tilde{w}_j \ge c_i & i \in E \end{array} \right\}$$

and

$$\max\left\{\sum_i r_i x_i + \sum_i c_i y_i \;\middle|\; \begin{matrix} x_j + \sum_i(\delta_{ij}-p_{ij})y_i = \beta_j & j \in E \\ x_i, y_i \geq 0 & i \in E \end{matrix}\right\}$$

respectively. The adaptation of algorithm VI to the optimal stopping problem gives the following algorithm.

ALGORITHM VIII for the construction of an optimal pure and stationary transient policy in an optimal stopping problem.

step 1: Take any vector β such that $\beta_j > 0$, $j \in E$.

step 2: Use the simplex method to compute an optimal solution (x^*, y^*) of the linear programming problem

$$\max\left\{\sum_i r_i x_i + \sum_i c_i y_i \;\middle|\; \begin{matrix} x_j + \sum_i(\delta_{ij}-p_{ij})y_i = \beta_j & j \in E \\ x_i, y_i \geq 0 & i \in E \end{matrix}\right\}$$

(if the problem has an infinite solution, then there exists no optimal transient policy).

step 3: Take f_*^∞ such that

$$f_*(i) = \begin{cases} 1 & \text{if } x_i^* > 0 \\ 2 & \text{if } y_i^* > 0 \end{cases} \qquad i \in E.$$

REMARK 3.3.10. The constraints of the linear program imply that $x_j^* + y_j^* = \beta_j + \Sigma_i\, p_{ij} y_i^* > 0$, $j \in E$. Since the simplex method gives an extreme solution and since any extreme solution has at most N (the number of constraints) positive components, we have either $x_i^* > 0$ or $y_i^* > 0$ for every $i \in E$. Hence, policy f_*^∞ is well-defined.

REMARK 3.3.11. Suppose that the linear programming problem has a finite optimum. Then, the vector w, defined by (3.3.2), is finite. Let $\Gamma := \{i \in E \mid w_i = r_i\}$. The existence of a pure and stationary optimal policy and the definition of Γ imply that an optimal stopping rule is stop on Γ and to continue on $E \backslash \Gamma$. From the complementary slackness property of linear programming, it follows that $E_{x^*} \subset \Gamma$.

REMARK 3.3.12. DERMAN ([1970], chapter 8) presents analogous formulations for the entrance-fee problem, i.e. the optimal stopping problem with $r_i = 0$ for all $i \in E$.

3.4. CONTRACTING DYNAMIC PROGRAMMING

Throughout this section we have the following *contraction assumption*.

ASSUMPTION 3.4.1. *There exists a* $\mu >> 0$, $\mu \in \mathbb{R}^N$, *and a real number* $\alpha \in [0,1)$ *such that* $\Sigma_j p_{iaj}\mu_j \le \alpha\, \mu_i$, $a \in A(i)$, $i \in E$.

In theorem 3.2.4 is shown that in a contracting dynamic programming problem *any policy is transient*. Hence, optimal transient policies are also optimal in the class C of all policies. This is true in the unconstrained case as well as in the constrained case. Therefore, we can use the results of the previous section to obtain optimal policies in both cases. Moreover, we can slightly extend some results of section 3.3. Below we summarize for the sake of completeness the results for the contracting dynamic programming problem.

THEOREM 3.4.1. *The TMD-value vector* v *is the smallest TMD-superharmonic vector.*

PROOF. Since any policy is transient, we have $v = w$. Theorem 3.2.1 implies the existence of a pure and stationary optimal (transient) policy f^∞. Then $v(f^\infty)$ is finite, and consequently v is finite. Now, apply theorem 3.3.2 to complete the proof. □

THEOREM 3.4.2. *The mapping* $x_{ia}(\pi) := [\beta^T(I-P(\pi))^{-1}]_i \cdot \pi_{ia}$, $a \in A(i)$, $i \in E$, *is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the dual linear program* (3.3.7). *The inverse mapping is given by* $\pi_{ia}(x) := x_{ia}/x$, $a \in A(i)$, $i \in E$. *Furthermore, this mapping has the property that pure policies correspond to extreme feasible solutions.*

PROOF. See theorem 3.3.3. □

THEOREM 3.4.3. *The linear programming problem* (3.3.7) *has a finite optimal solution. Moreover, if* x^* *is an optimal solution of* (3.3.7), *then any pure and stationary policy* f_*^∞ *such that* $x^*_{if_*(i)} > 0$, $i \in E$, *is an optimal policy.*

PROOF. Since v is the (finite) optimal solution of program (3.3.6), the dual program (3.3.7) also has a finite optimal solution.
Let x^* be any optimal solution of (3.3.7). Then $\Sigma_a x^*_{ia} > 0$, $i \in E$, and consequently, we can take a pure and stationary policy f_*^∞ such that $x^*_{if_*(i)} > 0$, $i \in E$. The complementary slackness property of the primal and dual linear program implies that $v = r(f_*)+P(f_*)v$. Since f_*^∞ is transient, the matrix $I-P(f_*)$ is nonsingular. Hence,

$$v = (I-P(f_*))^{-1}r(f_*) = v(f_*^\infty),$$

implying that f_*^∞ is optimal. □

As a consequence of theorem 3.4.3, a pure and stationary optimal policy can be obtained by the following algorithm.

ALGORITHM IX for the construction of a pure and stationary optimal policy in a contracting dynamic programming problem (linear programming).

step 1: Take any vector β such that $\beta_j > 0$, $j \in E$.
step 2: Compute an optimal solution x^* of the linear programming problem

$$\max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \end{array}\right\}.$$

step 3: Take f_*^∞ such that $x^*_{if_*(i)} > 0$, $i \in E$.

THEOREM 3.4.4. *The correspondence between the stationary policies and the feasible solutions of the linear program preserves the optimality property, i.e.*

1. *If* π^∞ *is a stationary optimal policy, then* $x(\pi)$ *is an optimal solution of the linear program.*
2. *If* x *is an optimal solution of the linear program, then the stationary policy* $\pi^\infty(x)$ *is an optimal policy.*

PROOF. See theorem 3.3.6. □

We continue this section with a discussion about the relation between the *policy improvement* method and the linear programming approach. The policy improvement method for discounted dynamic programming is due to HOWARD [1960]. We give the analogon for contracting dynamic programming and we establish that this method is equivalent to a particular linear programming method, called the simplex method with *block-pivoting* (cf. DANTZIG [1963] pp.201-202). Furthermore, we show that the standard simplex algorithm is equivalent to a special policy improvement algorithm.

For every $i \in E$ and every $f^\infty \in C_D$, we define a set $A(i,f)$ by

$$A(i,f) \;:=\; \{a \in A(i) \mid r_{ia} + \textstyle\sum_j p_{iaj} v_j(f^\infty) > v_i(f^\infty)\}.$$

The policy improvement method is based on the following theorem.

THEOREM 3.4.5. *Let* f^∞ *be any pure and stationary policy.*

(i) *If* $A(i,f) = \emptyset$, $i \in E$, *then* f^∞ *is an optimal policy.*

(ii) *If* $A(i,f) \neq \emptyset$ *for some* $i \in E$, *then* $v(g^\infty) > v(f^\infty)$, *where* $g^\infty \neq f^\infty$ *is any pure and stationary policy which satisfies for each* $i \in E$ *either* $g(i) = f(i)$ *or* $g(i) \in A(i,f)$.

PROOF. (The proof of this theorem is similar to the proof of theorem 3 in BLACKWELL [1962]).

(i) Since $A(i,f) = \emptyset$ for all $i \in E$, we have $r(g) + P(g)v(f^\infty) \le v(f^\infty)$ for any pure and stationary policy g^∞. Since $(I-P(g))^{-1} = \Sigma_{t=1}^{\infty} P^{t-1}(g) \ge 0$, we obtain

$$v(f^\infty) \;\ge\; (I-P(g))^{-1} r(g) \;=\; v(g^\infty)$$

for any pure and stationary policy g^∞. Hence, f^∞ is an optimal policy.

(ii) Let $g^\infty \neq f^\infty$ be such that for each $i \in E$ either $g(i) = f(i)$ or $g(i) \in A(i,f)$. Then,

$$r_i(g) + (P(g)v(f^\infty))_i \ge v_i(f^\infty), \quad i \in E,$$

with strict inequality for at least one i. Then, we obtain analogously to part (i) of the proof

$$v_i(g^\infty) = \textstyle\sum_{t=1}^{\infty} P^{t-1}(g) r(g) \ge v_i(f^\infty), \quad i \in E,$$

with strict inequality for at least one i. Hence, $v(g^\infty) > v(f^\infty)$, which completes the proof of the theorem. □

The policy improvement algorithm can be formulated as follows.

ALGORITHM X for the construction of a pure and stationary optimal policy in a contracting dynamic programming problem (policy improvement).

step 1: Take any pure and stationary policy f^∞.

step 2: Compute $v(f^\infty)$ as the unique solution of the linear system

$$x_i = r_i(f) + \sum_j p_{ij}(f)x_j, \quad i \in E.$$

step 3: Determine for every $i \in E$

$$A(i,f) := \{a \in A(i) \mid r_{ia} + \sum_j p_{iaj} v_j(f^\infty) > v_i(f^\infty)\}.$$

step 4: If $A(i,f) = \emptyset$, $i \in E$, then f^∞ is an optimal policy (STOP). Otherwise, go to step 5.

step 5: Take any policy g^∞ such that $g \neq f$ and such that for each $i \in E$ either $g(i) = f(i)$ or $g(i) \in A(i,f)$.

step 6: $f := g$ and go to step 2.

THEOREM 3.4.6. *Algorithm X determines an optimal policy in a finite number of iterations.*

PROOF. If in step 5 policy g^∞ is found as successor of f^∞, then $v(g^\infty) > v(f^\infty)$ (see theorem 3.4.5). Hence, each pure and stationary policy occurs at most once. Since $|C_D| < \infty$, the algorithm terminates after a finite number of iterations. Consequently, we finish with a policy f_*^∞ such that $A(i,f_*) = \emptyset$ for all $i \in E$. Then, theorem 3.4.5 implies that the policy f_*^∞ is optimal. □

Consider an iteration in the policy improvement algorithm. If

$$r_{ia} + \sum_j p_{iaj} v_j(f^\infty) \le v_i(f^\infty) \quad \text{for all } a \in A(i),$$

then $g(i) = f(i)$. Otherwise, we may take for $g(i)$ any action a for which

$$r_{ia} + \sum_j p_{iaj} v_j(f^\infty) > v_i(f^\infty).$$

By theorem 3.4.2, the vector $x(f^\infty)$ which is defined by formula (3.3.12) is an extreme feasible solution of the linear program (3.3.7). The linear programming tableau corresponding to this extreme feasible solution $x(f^\infty)$ has as basis matrix $(I-P(f))^T$. From theorem 1.4.1 and tableau (1.4.2), it follows that the coefficients of the transformed objective function have the values of the corresponding dual variables. Hence, the column of a nonbasic variable $x_{ia}(f^\infty)$ has in the transformed objective function the value

$$(3.4.1) \qquad d_{ia} := \tilde{w}_i - r_{ia} - \sum_j p_{iaj}\tilde{w}_j.$$

Here, $\tilde{w}_i$ is the variable which corresponds to the i-th equality of problem (3.3.7). Since $\tilde{w}_i$, $i \in E$, are unrestricted in sign, they are orthogonal to the artificial variables z_i, $i \in E$, of problem (3.3.7). Therefore, if we want to know the values $\tilde{w}_i$, $i \in E$, then we have to keep into the simplex tableau the artificial variables. Since $x_{if(i)}(f) > 0$, $i \in E$, it follows from the orthogonality of the corresponding primal and dual variables in the simplex tableau, that $d_{if(i)} = 0$, $i \in E$. Then, we obtain $\tilde{w} = r(f) + P(f)\tilde{w}$ which implies that $\tilde{w} = v(f^\infty)$. Hence, formula (3.4.1) may be written as

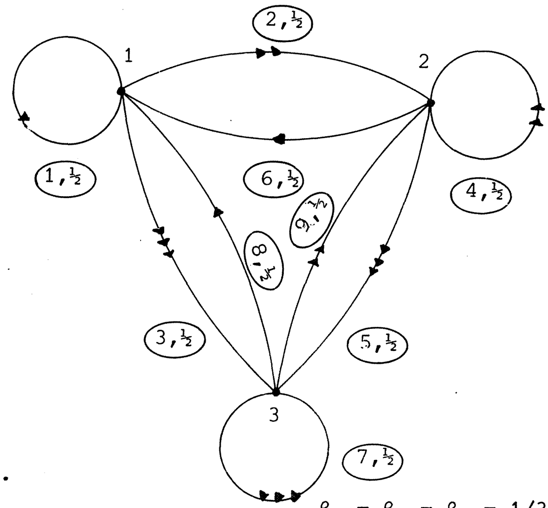
$$(3.4.2) \qquad d_{ia} := v_i(f^\infty) - r_{ia} - \sum_j p_{iaj} v_j(f^\infty).$$

Since $a \in A(i,f)$ if and only if $d_{ia} < 0$, it follows that the set of actions from which $g(i)$ may be chosen is exactly the set of possible pivot columns in the simplex method. Let $E_o := \{i \in E \mid g(i) \neq f(i)\}$. If we exchange the nonbasic variables corresponding to $g(i)$, $i \in E_o$, and the basic variables corresponding to $f(i)$, $i \in E_o$, then we obtain a linear programming algorithm in which more than one pivot step is performed simultaneously. Such an algorithm is called a block-pivoting algorithm. In the standard simplex method we choose as pivot column the column which has the most negative d_{ia}-coefficient. Since this rule may also be used in the policy improvement method, the standard simplex method is equivalent to a particular policy improvement algorithm. We summarize the above statements in the following conclusions.

CONCLUSIONS.

1. *Any policy improvement algorithm is equivalent to a block-pivoting simplex algorithm.*

2. *The standard simplex algorithm is equivalent to a particular policy improvement algorithm.*



$\beta_1 = \beta_2 = \beta_3 = 1/3$

Figure 3.4.1.

EXAMPLE 3.4.1. We compute an optimal policy for the model given in figure 3.4.1, by the policy improvement method as well as by the equivalent standard simplex method.

Policy improvement

Iteration 1:

1. $f(1) = 3,\ f(2) = 2,\ f(3) = 1.$
2. $v(f^\infty) = (28/3, 24/3, 38/3)^T.$
3. $A(1,f) = \emptyset,\ A(2,f) = \{1,3\},\ A(3,f) = \{2,3\}.$
5. d_{ia} is minimal for $i = 2,\ a = 3$: $g(1) = 3,\ g(2) = 3,\ g(3) = 1.$
6. $f(1) = 3,\ f(2) = 3,\ f(3) = 1.$

Iteration 2:

2. $v(f^\infty) = (28/3, 34/3, 38/3)^T.$
3. $A(1,f) = \emptyset,\ A(2,f) = \emptyset,\ A(3,f) = \{2,3\}.$
5. d_{ia} is minimal for $i = 3,\ a = 2.$
   $g(1) = 3,\ g(2) = 3,\ g(3) = 2.$
6. $f(1) = 3,\ f(2) = 3,\ f(3) = 2.$

Iteration 3:

2. $v(f^\infty) = (32/3, 38/3, 46/3)^T.$
3. $A(1,f) = A(2,f) = A(3,f) = \emptyset.$
4. f^∞ is optimal.

Linear programming

Iteration 1:

Policy f^∞, where $f(1) = 3,\ f(2) = 2,\ f(3) = 1$, corresponds to the simplex tableau with x_{13}, x_{22} and x_{31} as the basic variables. This tableau has the following form:

		x_{11}	x_{12}	z_1	x_{21}	z_2	x_{23}	z_3	x_{32}	x_{33}
x_{13}	2/3	2/3	4/3	4/3	-2/3	0	-1/3	2/3	2/3	1/3
x_{22}	2/3	0	-1	0	2	2	②	0	-1	0
x_{31}	2/3	1/3	2/3	2/3	-1/3	0	-2/3	4/3	4/3	2/3
x_0	10	11/3	7/3	28/3	-2/3	8	-10/3	38/3	-1/3	-2/3

Iteration 2:

The variables x_{23} and x_{22} are exchanged.

		x_{11}	x_{12}	z_1	x_{21}	x_{22}	z_2	z_3	x_{32}	x_{33}
x_{13}	7/9	2/3	7/6	4/3	-1/3	1/6	1/3	2/3	1/2	1/3
x_{23}	1/3	0	-1/2	0	1	1/2	1	0	-1/2	0
x_{31}	8/9	1/3	1/3	2/3	1/3	1/3	2/3	4/3	(1)	2/3
x_0	100/9	11/3	2/3	28/3	8/3	5/3	34/3	38/3	-2	-2/3

Iteration 3:

The variables x_{32} and x_{31} are exchanged.

		x_{11}	x_{12}	z_1	x_{21}	x_{22}	z_2	x_{31}	z_3	x_{33}
x_{13}	1/3	1/2	1	1	-1/2	0	0	-1/2	0	0
x_{23}	7/9	1/6	-1/3	1/3	7/6	2/3	4/3	1/2	2/3	1/3
x_{32}	8/9	1/3	1/3	2/3	1/3	1/3	2/3	1	4/3	2/3
x_0	116/9	13/3	4/3	32/3	10/3	7/3	38/3	2	46/3	2/3

($x^*_{11} = 0$, $x^*_{12} = 0$, $x^*_{13} = 1/3$, $x^*_{21} = 0$, $x^*_{22} = 0$, $x^*_{23} = 7/9$, $x^*_{31} = 0$, $x^*_{32} = 8/9$, $x^*_{33} = 0$) is an optimal solution. Then, f_*^∞, where $f_*(1) = 3$, $f_*(2) = 3$, and $f_*(3) = 2$, is an optimal policy.

Suppose that an upper bound b of the TMD-value-vector v is known. Then, the calculations can often been accelerated by the *elimination of suboptimal actions*. An action $a \in A(i)$ is said to be *suboptimal* if there does not exist an optimal policy $f^\infty \in C_D$ with $f(i) = a$. Since v is TMD-superharmonic and since f^∞ is optimal if and only if $v = r(f) + P(f)v$, an action $a \in A(i)$ is suboptimal if and only if

$$r_{ia} + \sum_j p_{iaj} v_j < v_i.$$

The concept of suboptimal actions was introduced by MACQUEEN [1967].

<u>THEOREM 3.4.7</u>. *Suppose that* b *is an upper bound for* v. *Let, in the simplex tableau corresponding to the extreme feasible solution* x(f), d_{ia} *be the value of the variable dual to* $x_{ia}(f)$, $a \in A(i)$, $i \in E$. *If* $a_i \in A(i)$ *satisfies*

$$(3.4.3)\qquad d_{ia_i} > \min_a d_{ia} + \sum_j p_{ia_ij}(b_j - v_j(f^\infty)),$$

then action a_i *is suboptimal.*

PROOF. Using the formulae (3.4.2) and (3.4.3) we may write

$$\begin{aligned} r_{ia_i} + \textstyle\sum_j p_{ia_ij} v_j &\le r_{ia_i} + \textstyle\sum_j p_{ia_ij} b_j = \\ &= -d_{ia_i} + v_i(f^\infty) + \textstyle\sum_j p_{ia_ij}(b_j - v_j(f^\infty)) \\ &< -d_{ia_i} + v_i(f^\infty) + d_{ia_i} - \min_a d_{ia} = \\ &\quad v_i(f^\infty) + \max_a\{r_{ia} + \textstyle\sum_j p_{iaj} v_j(f^\infty) - v_i(f^\infty)\} \\ &\le \max_a\{r_{ia} + \textstyle\sum_j p_{iaj} v_j\} = v_i. \end{aligned}$$

This completes the proof that a_i is a suboptimal action. □

Let f^∞ be any pure and stationary policy. Then, we have observed that $x(f^\infty)$ is an extreme feasible solution of the linear programming problem (3.3.7). Furthermore, we have seen that $v_j(f^\infty)$, $j \in E$, are the values of the dual variables that correspond to the artificial variables of program (3.3.7); the other dual variables, namely the variables that are orthogonal to $x_{ia}(f)$, $a \in A(i)$, $i \in E$, have the values d_{ia}, defined by (3.4.2), $a \in A(i)$, $i \in E$.

Hence, in the simplex tableau that corresponds to any pure and stationary policy f^∞, we can easily compute the vector $b(f)$, defined by

$$(3.4.4)\qquad b(f) := v(f^\infty) - \frac{\min_i \min_a d_{ia}/\mu_i}{1-\alpha} \cdot \mu,$$

where μ and α are the quantities introduced in the first paragraph of this section. If these quantities are unknown, then they can be computed by linear programming (see remark 3.2.3).

We will show that $b(f)$ is an upper bound for the TMD-value-vector v. Then, $b(f)$ can be used in the suboptimality test (3.4.3).

LEMMA 3.4.1. $b(f)$, *defined by formula* (3.4.4), *is an upper bound for the TMD-value-vector* v.

PROOF. Let $M := min_i min_a \; d_{ia}/\mu_i$. Suppose that g^∞ is a pure and stationary optimal policy (theorem 3.2.1 implies its existence).
Then,

$$M \le \frac{d_{ig(i)}}{\mu_i} = \frac{v_i(f^\infty)-r_i(g)-(P(g)v(f^\infty))_i}{\mu_i}, \quad i \in E.$$

Consequently,

$$r(g) \le (I-P(g^\infty)\, v(f^\infty) - M\cdot\mu.$$

This implies that

$$(3.4.5) \qquad v = v(g^\infty) = (I-P(g))^{-1} r(g) \le v(f^\infty) - M\cdot(I-P(g))^{-1}\mu.$$

From the contraction property it follows that

$$(3.4.6) \qquad (I-P(g))^{-1}\mu = \sum_{t=1}^{\infty} P^{t-1}(g)\mu \le \sum_{t=1}^{\infty} \alpha^{t-1}\mu = (1-\alpha)^{-1}\mu.$$

Then (3.4.5) and (3.4.6) imply that

$$v \le v(f^\infty) - \frac{M}{1-\alpha}\cdot\mu = v(f^\infty) - \frac{min_i min_a \; d_{ia}/\mu_i}{1-\alpha}\cdot\mu = b(f),$$

completing the proof of this lemma. □

REMARK 3.4.1. Any feasible solution of the linear programming problem (3.3.6) is also an upper bound for v and can be used in the suboptimality test.

REMARK 3.4.2. The use of suboptimality tests is a familiar concept in the method of successive approximations. The elimination of suboptimal actions may improve the efficiency considerably. We have seen that we can very easily implement the suboptimality test (3.4.3) in the linear programming algorithm IX. Moreover, we may expect for the linear programming approach the same acceleration as it has in the method of successive approximations.

Next, we will discuss the constrained Markov decision problem. Let $\beta = (\beta_1,\beta_2,\ldots,\beta_N)^T$ be any given initial distribution. In contrast with section 3.3, we allow in this section that $\beta_j = 0$ for some $j \in E$. In the same way as in section 3.3, we define the vector $x(R)$ for $R \in \mathcal{C}$ and the sets K, K(M), K(S), K(D) and P. Notice that in this section the restriction to transient policies can be dropped.

The *constrained Markov decision problem* is then formulated by

$$(3.4.7)\qquad sup_R\{\sum_i\beta_i v_i(R) \mid \sum_i\sum_a q_{iak}x_{ia}(R) \le b_k \quad k = 1,2,\dots,m\}.$$

THEOREM 3.4.8. $\overline{K(D)} = K(S) = K(M) = K = P$.

PROOF. The proof is similar to the proof of theorem 3.3.4. However, it is not a direct consequence because we have in theorem 3.3.4 $\overline{K(D)} \subset K(S)$; furthermore, we here allow that $\beta_j = 0$ for some $j \in E$.
We first prove that $K(S) = P$. For any $x \in P$, we define a stationary policy $\pi^\infty(x)$ by

$$(3.4.8)\qquad \pi_{ia}(x) := \begin{cases} x_{ia}/x_i & a \in A(i),\ i \in E_x \\ \text{arbitrarily} & a \in A(i),\ i \notin E_x. \end{cases}$$

Since $x_{ia} = \pi_{ia}(x)\cdot x_i$ for all $a \in A(i)$ and $i \in E$, we obtain $\beta^T = x^T(I-P(\pi(x))$. Consequently,

$$x_{ia} = [\beta^T(I-P(\pi(x)))^{-1}]_i\cdot\pi_{ia} = x_{ia}(\pi^\infty(x)) \qquad a \in A(i),\ i \in E.$$

Hence, $x \in K(S)$.
Conversely, let $x(\pi^\infty) \in K(S)$. Then, analogously to the proof of theorem 3.3.3, it follows that $x(\pi^\infty) = x(\pi) \in P$, where $x_{ia}(\pi) := [\beta^T(I-P(\pi))^{-1}]_i\cdot\pi_{ia}$.
In the same way as in theorem 3.3.4 it can be shown that

$$\overline{K(D)} \subset K(S) = K(M) = K = P.$$

Suppose that $\overline{K(D)} \neq K(S)$. Then there exists a stationary policy π^∞ such that $x(\pi) \notin \overline{K(D)}$. Since $\overline{K(D)}$ is a closed convex set, it follows from theorem 1.2.1 that there exist real numbers r_{ja}, $a \in A(j)$, $j \in E$, such that

$$\sum_j\sum_a r_{ja}x_{ja}(\pi) > \sum_j\sum_a r_{ja}x_{ja} \quad \text{for all } x \in \overline{K(D)}.$$

Hence,

$$(3.4.9)\qquad \sum_i\beta_i v_i(\pi^\infty) = \sum_j\sum_a r_{ja}x_{ja}(\pi) > \sum_j\sum_a r_{ja}x_{ja}(f) = \sum_i\beta_i v_i(f^\infty)$$

for all $f \in C_D$. Since there exists a pure and stationary policy which is optimal with respect to the rewards r_{ja}, relation (3.4.9) gives a contradiction. Hence, we have shown that $\overline{K(D)} = K(S)$, which completes the proof of the theorem. □

REMARK 3.4.3. We can also prove, similar to theorem 3.3.3, that x is an extreme point of P if and only if $x \in K(D)$. However, we can loose the one-to-one correspondence if $\beta_j = 0$ for some $j \in E$. E.g., choose in the model of example 3.3.4 $\beta_1 = 0$, $\beta_2 = 1$, and let f_1^∞, f_2^∞ be such that $f_1(1) = f_1(2) = 1$, $f_2(1) = 2$, $f_2(2) = 1$. Then, it can easily be verified that $x(f_1) = x(f_2) = x$, where $x_{11} = x_{12} = x_{21} = 0$, $x_{22} = 2$.

In order to solve (3.4.7) we consider the following linear programming problem

$$(3.4.10) \qquad \max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = \beta_j & & j \in E \\ \sum_i\sum_a q_{iak}x_{ia} & \le b_k & k = 1,2,\dots,m \\ & x_{ia} \ge 0 & a \in A(i),\ i \in E \end{array}\right\}.$$

Analogously to theorem 3.3.7, we can prove the following theorem.

THEOREM 3.4.9.

(i) *Problem* (3.4.7) *is feasible if and only if problem* (3.4.10) *is feasible.*

(ii) *The optima of the problems* (3.4.7) *and* (3.4.10) *are equal.*

(iii) *If* x *is an optimal solution of the linear program* (3.4.10), *then the stationary policy* $\pi^\infty(x)$, *defined by* (3.4.8) *is an optimal solution of problem* (3.4.7).

(iv) *If* R *is an optimal solution of problem* (3.4.7), *then* x(R) *is an optimal solution of the linear programming problem* (3.4.10).

Theorem 3.4.9 provides an algorithm for contracting dynamic programming with additional constraints.

ALGORITHM XI for the construction of a stationary optimal policy in a contracting dynamic programming problem with additional constraints and with initial distribution $\beta \ge 0$.

step 1: Determine an optimal solution x^* of the linear programming problem

$$max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = \beta_j & & j \in E \\ \sum_i\sum_a q_{iak}x_{ia} & \le b_k & k = 1,2,\ldots,m \\ & x_{ia} \ge 0 & a \in A(i),\ i \in E \end{array}\right\}$$

(if the problem is infeasible, then the constrained TMD-problem is also infeasible).

step 2: Take ${}_*\pi^\infty$ such that

$${}_*\pi_{ia} := \begin{cases} x^*_{ia}/x^*_i & a \in A(i),\ i \in E_{x^*} \\ \text{arbitrarily} & a \in A(i),\ i \notin E_{x^*} \end{cases}$$

REMARK 3.4.4. The DMD-problem, which may be viewed as a TMD-problem with $\Sigma_j p_{iaj} = \alpha < 1$ for all states i and all actions $a \in A(i)$, is a special case of the contracting dynamic programming problem. In fact, these models are equivalent in the following sense: the expected total reward of any policy R in the two models differs by a multiplicative factor which only depends on the initial state.

To prove this equivalence, we consider for a contracting dynamic programming model (E,A,p,r) a transformed model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, which is defined as follows:

$$\tilde{E} := \{0,1,\ldots,N\}$$

$$\tilde{A}(i) := \begin{cases} A(i) & i \in E \\ \{1\} & i = 0 \end{cases}$$

$$\tilde{p}_{iaj} := \begin{cases} \mu_i^{-1}\ p_{iaj}\mu_j & i \in E,\ a \in \tilde{A}(i),\ j \in E \\ \alpha - \mu_i^{-1}\ \sum_k p_{iak}\mu_k & i \in E,\ a \in \tilde{A}(i),\ j = 0 \\ \alpha & i = 0,\ a = 1,\quad j = 0 \\ 0 & i = 0,\ a = 1,\quad j \in E \end{cases}$$

$$\tilde{r}_{ia} := \begin{cases} r_{ia}/\mu_i & a \in \tilde{A}(i),\ i \in E \\ 0 & a = 1,\quad i = 0. \end{cases}$$

The model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$ is a DMD-problem, namely

$$\sum_{j\in\tilde{E}} \tilde{p}_{iaj} = \tilde{p}_{iao} + \sum_{j\in E} \tilde{p}_{iaj} = \alpha \qquad a \in \tilde{A}(i),\ i \in \tilde{E}.$$

In order to analyse the rewards, we may, by theorem 2.5.1, restrict ourselves to Markov policies. Since state 0 is absorbing and the reward in state 0 is zero, $\tilde{v}(R) = \hat{v}(R)$ for any policy R, where $\hat{v}(R)$ is the expected total reward in the model $(\hat{E},\hat{A},\hat{p},\hat{r})$, defined by

$$\begin{aligned}
\hat{E} &:= E \\
\hat{A}(i) &:= \tilde{A}(i) && i \in \hat{E} \\
\hat{p}_{iaj} &:= \tilde{p}_{iaj} && i \in \hat{E},\ a \in \hat{A}(i),\ j \in \hat{E} \\
\hat{r}_{ia} &:= \tilde{r}_{ia} && i \in \hat{E},\ a \in \hat{A}(i).
\end{aligned}$$

Let $R = (\pi^1,\pi^2,\ldots)$ be any Markov policy in model $(\hat{E},\hat{A},\hat{p},\hat{r})$. We observe that

$$[\hat{P}(\pi^1)\hat{P}(\pi^2)\cdots\hat{P}(\pi^t)]_{ij} = \frac{1}{\mu_i}[P(\pi^1)P(\pi^2)\cdots P(\pi^t)]_{ij}\cdot\mu_j$$

for all $i,j \in E$ and $t \in \mathbb{N}$. Therefore, we can write

$$\begin{aligned}
\hat{v}_i(R) &= \lim_{n\to\infty} \sum_{t=1}^{n} [\hat{P}(\pi^1)\hat{P}(\pi^2)\cdots\hat{P}(\pi^{t-1})\hat{r}(\pi^t)]_i \\
&= \lim_{n\to\infty} \sum_{t=1}^{n} \sum_j \mu_i^{-1}\cdot\{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij}\cdot\mu_j\cdot \frac{r_j(\pi^t)}{\mu_j} \\
&= \mu_i^{-1}\cdot\lim_{n\to\infty} \sum_{t=1}^{n} [P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)]_i \\
&= \mu_i^{-1}\, v_i(R), \qquad i \in E.
\end{aligned}$$

Hence, it follows that a policy is optimal in the undiscounted TMD-model if and only if the policy is optimal in the corresponding DMD-model. The transformations, that were used above, are due to VEINOTT [1969] (see also VAN HEE, HORDIJK & VAN DER WAL [1977]).

Next, we consider what happens in a constrained dynamic programming problem. Suppose that we want to solve problem (3.4.7) for a contracting dynamic programming problem. Then we solve the following constrained problem for the corresponding discounted model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$:

$$(3.4.11) \qquad sup_R\{\tilde{\beta}^T\tilde{v}(R) \mid \sum_i\sum_a\tilde{q}_{iak}\tilde{x}_{ia}(R) \le \tilde{b}_k \qquad k = 1,2,\dots,m\},$$

where

$$\tilde{\beta}_i := \begin{cases} \beta_i\mu_i & i \in E \\ 0 & i = 0 \end{cases}$$

$$\tilde{q}_{iak} := \begin{cases} q_{iak}/\mu_i & a \in \tilde{A}(i),\ i \in E,\quad k = 1,2,\dots,m \\ 0 & a = 1, \quad i = 0, \quad k = 1,2,\dots,m \end{cases}$$

$$\tilde{b}_k := b_k \qquad k = 1,2,\dots,m.$$

The equivalence between the problems (3.4.7) and (3.4.11) is a consequence of the following properties:

$$\begin{aligned}
\text{(i)}\ \tilde{x}_{ja}(R) &= \sum_i\tilde{\beta}_i\cdot\sum_{t=1}^{\infty}\{\tilde{P}(\pi^1)\tilde{P}(\pi^2)\cdots\tilde{P}(\pi^{t-1})\}_{ij}\cdot\pi_{ja}^t \\
&= \sum_i\beta_i\mu_i\cdot\sum_{t=1}^{\infty}\mu_i^{-1}\{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij}\mu_j\cdot\pi_{ja}^t \\
&= \mu_j\cdot\sum_i\beta_i\cdot\sum_{t=1}^{\infty}\{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij}\cdot\pi_{ja}^t \\
&= \mu_j\cdot x_{ja}(R) \qquad a \in A(j),\ j \in E.
\end{aligned}$$

$$\begin{aligned}
\text{(ii)}\ \tilde{\beta}^T\tilde{v}(R) &= \sum_j\sum_a\tilde{r}_{ja}\tilde{x}_{ja}(R) = \sum_j\sum_a r_{ja}\mu_j^{-1}\mu_j x_{ja}(R) \\
&= \sum_j\sum_a r_{ja}x_{ja}(R) = \beta^T v(R).
\end{aligned}$$

$$\text{(iii)}\ \sum_i\sum_a\tilde{q}_{iak}\tilde{x}_{ia}(R) = \sum_i\sum_a q_{iak}\mu_i^{-1}\mu_i x_{ia}(R) = \sum_i\sum_a q_{iak}x_{ia}(R).$$

CONCLUSION: *Discounting dynamic programming and contracting dynamic programming are equivalent models for unconstrained as well as for constrained Markov decision models.*

3.5. POSITIVE DYNAMIC PROGRAMMING

Positive dynamic programming problems are dynamic programming problems that satisfy the following assumption.

ASSUMPTION 3.5.1. $r_{ia} \geq 0 \quad a \in A(i),\ i \in E.$

THEOREM 3.5.1. *v is the smallest nonnegative TMD-superharmonic vector.*

PROOF. (cf. HORDIJK [1974] p.25). From theorem 3.2.2 and assumption 3.5.1, it follows that v is a nonnegative TMD-superharmonic vector. Suppose that $\tilde{w}$ is also a nonnegative TMD-superharmonic vector. Since $v_i = \max\{v_i(f^\infty) \mid f^\infty \in C_D\}$, $i \in E$, it is sufficient to show that $v(f^\infty) \leq \tilde{w}$ for every $f^\infty \in C_D$. Let f^∞ be an arbitrarily chosen pure and stationary policy. Then the superharmonicity of $\tilde{w}$ implies that $\tilde{w} \geq r(f) + P(f)\tilde{w}$. By iterating this inequality, we obtain

$$\tilde{w} \geq \sum_{t=1}^{n} P^{t-1}(f)r(f) + P^n(f)\tilde{w} \geq \sum_{t=1}^{n} P^{t-1}(f)r(f) \qquad n \in \mathbb{N}.$$

Hence, for $n \to \infty$ we find

$$\tilde{w} \geq \sum_{t=1}^{\infty} P^{t-1}(f)r(f) = v(f^\infty),$$

which completes the proof of the theorem. □

In order to find an optimal policy, theorem 3.5.1 suggests the use of the following linear program:

$$(3.5.1) \qquad \min\left\{ \sum_j \beta_j \tilde{w}_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij} - p_{iaj})\tilde{w}_j \geq r_{ia} & a \in A(i),\ i \in E \\ \tilde{w}_j \geq 0 & j \in E \end{array} \right\},$$

where $\beta_j > 0$, $j \in E$, are given numbers. The dual linear programming problem is:

$$(3.5.2) \qquad \max\left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} \leq \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \end{array} \right\}.$$

This dual program is feasible (e.g. $x = 0$ is a feasible solution). Therefore, there are two possibilities: the optimum of (3.5.2) is finite or infinite. We will treat these possibilities as two separate cases and we will see that the construction of optimal policies (in class C of all policies) is different. If the dual program has a finite solution, then an optimal policy can be obtained directly from the optimal solution of the linear program. In the infinite case, we need some analysis of the underlying Markov chain structure. Fortunately, also in this case we can present a finite algorithm to determine an optimal pure and stationary policy.

THEOREM 3.5.2. *Suppose that* x^* *is an extreme optimal solution of the dual program. Then, the pure and stationary policy* f_*^∞, defined by

$$f_*(i) := \begin{cases} a_i \textit{ such that } x^*_{ia_i} > 0 & i \in E_{x^*} \\ \textit{arbitrarily} & i \notin E_{x^*}, \end{cases}$$

is an optimal policy.

PROOF. By introducing slack variables, we can write the constraints of the problems (3.5.1) and (3.5.2) as follows

$$(3.5.3) \qquad \begin{cases} \sum_j (\delta_{ij}-p_{iaj})\tilde{w}_j - u_{ia} = r_{ia} & a \in A(i),\ i \in E \\ \tilde{w}_j \geq 0 & j \in E \\ u_{ia} \geq 0 & a \in A(i),\ i \in E \end{cases}$$

and

$$(3.5.4) \qquad \begin{cases} \sum_i \sum_a (\delta_{ij}-p_{iaj})x_{ia} + y_j = \beta_j & j \in E \\ x_{ia} \geq 0 & a \in A(i),\ i \in E \\ y_j \geq 0 & j \in E \end{cases}$$

respectively.

Theorem 3.5.1 implies that v is the optimal solution of problem (3.5.1). Let

$$u^*_{ia} := \sum_j (\delta_{ij}-p_{iaj})v_j - r_{ia} \qquad a \in A(i),\ i \in E$$

$$y^*_j := \beta_j - \sum_i \sum_a (\delta_{ij}-p_{iaj})x^*_{ia} \qquad j \in E.$$

Then, it follows from the theory of linear programming that

$$\sum_j \beta_j v_j = \sum_i \sum_a r_{ia} x^*_{ia}$$

and

$$\sum_j y_j^* v_j = \sum_i \sum_a u_{ia}^* x_{ia}^* = 0.$$

Since x^* is an extreme point and the dual program has N constraints, the vector $(x^*,y^*)^T$ has at most N positive components. Then,

$$\sum_a x_{ja}^* + y_j^* = \beta_j + \sum_i \sum_a p_{iaj} x_{ia}^* \geq \beta_j > 0, \quad j \in E,$$

implies that for every $j \in E_{x^*}$ there is exactly one action $a_j \in A(j)$ such that $x_{ja_j}^* > 0$. Hence, the policy f_*^∞ is uniquely determined on E_{x^*}. Also it turns out that any extreme solution is nondegenerated. Furthermore we can write

$$(x^*)^T = (\beta - y^*)^T + (x^*)^T P(f_*).$$

By iterating this equality, we obtain

$$(x^*)^T = (\beta - y^*)^T \sum_{t=1}^n P^{t-1}(f_*) + (x^*)^T P^n(f_*) \qquad n \in \mathbb{N}.$$

Consequently,

$$(x^*)^T r(f_*) = (\beta - y^*)^T \sum_{t=1}^n P^{t-1}(f_*) r(f_*) + (x^*)^T P^n(f_*) r(f_*), \; n \in \mathbb{N}.$$

Since $v(f_*^\infty) = \Sigma_{t=1}^\infty P^{t-1}(f_*) r(f_*) \leq v$ and v is finite, it follows that

$$\lim_{n\to\infty} P^n(f_*) r(f_*) = 0.$$

Therefore, we get

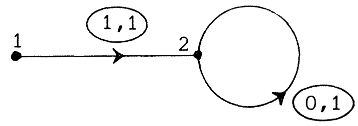
$$\beta^T v = \sum_i \sum_a r_{ia} x_{ia}^* = (x^*)^T r(f_*) = (\beta - y^*)^T v(f_*^\infty) \leq \beta^T v(f_*^\infty),$$

implying that f_*^∞ is an optimal policy. □

REMARK 3.5.1. If we use the simplex method to solve the linear programming problem (3.5.2) and it turns out that this problem has a finite optimum, then an optimal extreme solution is obtained.

REMARK 3.5.2. If the Markov decision problem is contracting, then the linear programs have finite solutions. The following example shows that the converse statement is not true, in general; in this example, an optimal nontransient policy is found.

EXAMPLE 3.5.1. The problem of figure 3.5.1 has only one policy and this policy is nontransient. The dual program is



$\beta_1 = \beta_2 = \frac{1}{2}$

Figure 3.5.1

$$\max\left\{x_{11} \;\middle|\; \begin{matrix} x_{11} \le \tfrac{1}{2};\ x_{11} \ge 0 \\ \\ -x_{11} \le \tfrac{1}{2};\ x_{21} \ge 0 \end{matrix}\right\}.$$

This problem has a finite optimum, namely $x_{11} = \frac{1}{2}$.

Suppose that the dual program (3.5.2) has an *infinite optimum*. Then, if we solve this problem by the simplex method starting with the extreme feasible solution $x = 0$, we obtain after a finite number of iterations a simplex tableau with a nonpositive column. In this column, the coefficient of the transformed objective function is strictly negative. Therefore, we have in this tableau an extreme feasible solution x and a direction vector s such that

(i) $x(\lambda) := x + \lambda s$ is feasible for all $\lambda \ge 0$.

(ii) $\Sigma_i \Sigma_a r_{ia} x_{ia}(\lambda) \to +\infty$ for $\lambda \to \infty$.

Since x is an extreme solution, it follows from the proof of theorem 3.5.2 that for every $j \in E_x$, x_{ja_j} is positive for exactly one action $a_j \in A(j)$. Therefore, if the linear program (3.5.2) has an infinite solution, we can find by the simplex method an extreme feasible solution x, a direction vector s and actions $a_j \in A(j)$, $j \in E_x$, such that:

$$\sum_i \sum_a (\delta_{ij} - p_{iaj}) s_{ia} \le 0 \qquad j \in E \tag{3.5.5}$$

$$s_{ia} \ge 0 \qquad a \in A(i),\ i \in E \tag{3.5.6}$$

$$\sum_i \sum_a r_{ia} s_{ia} > 0 \tag{3.5.7}$$

$$\sum_a x_{ja} = x_{ja_j} > 0 \qquad j \in E_x. \tag{3.5.8}$$

Corresponding to the direction vector s, we define a stationary policy π^∞ by

$$\pi_{ia} := \begin{cases} s_{ia}/s_i & a \in A(i),\ i \in E_s \\ \text{arbitrarily} & a \in A(i),\ i \notin E_s. \end{cases} \tag{3.5.9}$$

<u>THEOREM 3.5.3</u>. *The policy* π^∞, *defined by* (3.5.9), *can be chosen from* C_D.

<u>PROOF</u>. Let $a^*_{\cdot\ell}$ be the nonpositive column in the simplex tableau from which the infinite solution is obtained. Suppose that this column corresponds to the nonbasic variable x_{ka_o}. Then the direction vector s is given by

$$s_{ja} := \begin{cases} -a^*_{i_j\ell} & \text{if } j \in E_x,\ a = a_j \text{ and } x_{ja_j} \text{ is the basic variable} \\ & \text{corresponding to row } i_j \text{ of the simplex tableau} \\ 1 & \text{if } j = k,\ a = a_o \\ 0 & \text{elsewhere.} \end{cases}$$

Hence, to prove that π^∞ can be chosen from C_D, it is sufficient to show that $\Sigma_a s_{ka} = s_{ka_o}$. Assume the contrary. Then, $k \in E_x$ and $s_{ka_k} > 0$. For every $i \in E\backslash E_s$, we choose an arbitrary action $a_i \in A(i)$ and we take $\pi_{ia_i} := 1$ and $\pi_{ia} := 0$, $a \neq a_i$. Then it can be verified that

$$(3.5.10) \qquad P(\pi) = \delta \cdot P(f_1) + (1-\delta) \cdot P(f_2),$$

where $\delta = \varepsilon(1-\varepsilon)^{-1}$ with $\varepsilon = s_{ka_k}$ and $f_1^\infty, f_2^\infty \in C_D$ such that

$$f_1(i) := a_i,\ i \in E, \text{ and } f_2(i) := \begin{cases} f_1(i) & i \neq k \\ a_o & i = k. \end{cases}$$

From (3.5.5)-(3.5.7) and (3.5.9) it follows that

$$0 < \textstyle\sum_j s_j \le \sum_j \sum_i \sum_a p_{iaj} \pi_{ia} s_i = \sum_i s_i (\sum_j p_{ij}(\pi)) \le \sum_i s_i.$$

Hence

$$(3.5.11) \qquad \textstyle\sum_j p_{ij}(\pi) = 1 \qquad i \in E_s,$$

$$(3.5.12) \qquad s^T e = s^T P(\pi) e.$$

Since $s^T \le s^T P(\pi)$, (3.5.12) implies that $s^T = s^T P(\pi)$ and consequently, $s^T = s^T P^*(\pi)$. Therefore, $E_s \subset R(\pi)$, where $R(\pi)$ is the set of recurrent states in the Markov chain induced by $P(\pi)$, and E_s is closed under $P(\pi)$. By (3.5.10) and (3.5.11) we also have

$$\textstyle\sum_j p_{ij}(f_k) = 1, \quad i \in E_s, \text{ and } E_s \text{ is closed under } P(f_k)\ k = 1,2.$$

Therefore, we find

$$(3.5.13)\qquad \sum_j p_{ij}^{(n)}(f_1) = \sum_{j\in E_s} p_{ij}^{(n)}(f_1) = 1 \qquad i \in E_s,\ n \in \mathbb{N}.$$

Since x is an extreme feasible solution and since $E_s \subset E_x$, we have on the other hand

$$(3.5.14)\qquad x_{ja_j} = \beta_j + \sum_i \sum_a p_{iaj} x_{ia} \geq \beta_j + \sum_{i\in E_s} p_{ij}(f_1) x_{ia_i} \qquad j \in E_s.$$

Because E_s is closed under $P(f_1)$, we obtain by iterating (3.5.14)

$$x_{ja_j} \geq \sum_{i\in E_s} \beta_i \cdot \sum_{t=1}^n p_{ij}^{(t-1)}(f_1) + \sum_{i\in E_s} p_{ij}^{(n)}(f_1) x_{ia_i} \quad j \in E_s,\ n \in \mathbb{N}.$$

Consequently, $\Sigma_{t=1}^\infty\ p_{ij}^{(t-1)}(f_1) < \infty$, $i \in E_s$, $j \in E_s$, implying that $p_{ij}^{(n)}(f_1) \to 0$ for $n \to \infty$, $i \in E_s$, $j \in E_s$. Then $\Sigma_{j\in E_s} p_{ij}^{(n)}(f_1) \to 0$ for $n \to \infty$, $i \in E_s$, which contradicts to (3.5.13). This yields the theorem. □

Let f_s^∞ be the policy, defined by (3.5.9) and for which, as has been shown in theorem 3.5.3, we may assume that it belongs to $\mathcal{C}_D$.

<u>THEOREM 3.5.4</u>. $v_j(f_s^\infty) = +\infty$ *for at least one state* j.

<u>PROOF</u>. From the proof of theorem 3.5.3 it follows that $E_s \subset R(f_s)$ and that E_s is closed under $P(f_s)$.
Furthermore, (3.5.7) implies that $s^T r(f_s) > 0$. Hence, there exists a state $\ell \in E_s$ such that $r_\ell(f_s) > 0$. For any state j in the same ergodic set as state ℓ, we have

$$v_j(f_s^\infty) = \sum_{t=1}^\infty [P^{t-1}(f_s) r(f_s)]_j = \lim_{n\to\infty} n \cdot \frac{1}{n} \sum_{t=1}^n [P^{t-1}(f_s) r(f_s)]_j$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n [P^{t-1}(f_s) r(f_s)]_j = [P^*(f_s) r(f_s)]_j \geq p_{j\ell}^*(f_s) r_\ell(f_s) > 0.$$

Consequently, $v_j(f_s^\infty) = +\infty$. □

We can find a pure and stationary policy in the following way. First, we identify all ergodic sets in E_s which have a state ℓ such that $r_\ell(f_s) > 0$. For any state i in these ergodic sets we define $f_*(i) = f_s(i)$. Outside these ergodic sets, we choose actions which lead to these ergodic sets, if possible. Then f_*^∞ has for certain initial states, say E_o, a total reward $+\infty$, and $E\backslash E_o$ is closed under any policy. We repeat the same approach on $E\backslash E_o$. This method is outlined in the following algorithm.

ALGORITHM XII for the construction of a pure and stationary optimal policy in positive dynamic programming.

step 1: Use the simplex method to solve the linear program

$$(3.5.15) \qquad max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{matrix} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} \le \beta_j & j \in E \\ x_{ia} \ge 0 & a \in A(i),\ i \in E \end{matrix}\right\}.$$

If a finite optimal solution x^* is obtained, then go to step 2.
If an infinite optimum is discovered, then go to step 3.

step 2: Choose $f_*^\infty \in C_D$ such that $x^*_{if_*(i)} > 0$, $i \in E_{x^*}$.
Then, f_*^∞ is an optimal policy (STOP).

step 3: Let $a^*_{\cdot\ell}$ be the nonpositive column in the simplex tableau from which the infinite solution is obtained. Suppose that this column corresponds to the nonbasic variable x_{ka_o}, and suppose that the simplex tableau corresponds to the extreme feasible solution x. Define s by

$$s_{ja} := \begin{cases} -a^*_{i_j\ell} & \text{if } j \in E_x \text{ and } x_{ja_j} \text{ is the basic variable of row } i_j \text{ of the simplex tableau} \\ 1 & \text{if } j = k \text{ and } a = a_o \\ 0 & \text{elsewhere.} \end{cases}$$

step 4: Take $f_*^\infty \in C_D$ such that $s_{if_*(i)} > 0 \quad i \in E_s$.

step 5: Determine on E_s the ergodic sets in the Markov chain induced by $P(f_*)$ (see algorithm II).

step 6: Determine the union E_o of the ergodic sets under $P(f_*)$, which contain a state j such that $r_j(f_*) > 0$.

step 7: If $E_o = E$, then f_*^∞ is an optimal policy (STOP).
Otherwise, go to step 8.

step 8: Search for a triple $i \in E\backslash E_o$, $a_i \in A(i)$, $j \in E_o$ such that $p_{ia_ij} > 0$.
If such triple is found: $f_*(i) := a_i$, $E_o := E_o \cup \{i\}$, go to step 7.
Otherwise, go to step 9.

step 9: For $E := E \backslash E_o$ repeat the algorithm, starting in step 1.

THEOREM 3.5.5. *Algorithm XII determines a pure and stationary optimal policy f_*^∞ in a finite number of iterations.*

PROOF. If the linear programming problem, that is solved in step 1, has a finite optimal solution x^*, then theorem 3.5.2 implies that the policy f_*^∞ is optimal. Suppose that program (3.5.15) has an infinite solution. Then, by the theorems 3.5.3 and 3.5.4, the policy f_*^∞ which is defined in step 4 satisfies $v_j(f_*^\infty) = +\infty$ for every $j \in E_o$, where E_o is the nonempty set defined in step 6. Hence, if $E_o = E$, then the algorithm terminates in step 7 with an optimal policy f_*^∞.
Now, suppose that $E_o \neq E$. Then, in step 8 the policy f_*^∞ may be re-defined in a state $i \in E \backslash E_o$ such that $p_{ij}(f_*) > 0$ for at least one state $j \in E_o$. Consequently, $v_i(f_*^\infty) \geq r_i(f_*) + p_{ij}(f_*) \cdot v_j(f_*^\infty) = +\infty$. Next, E_o is replaced by $E_o \cup \{i\}$ and the steps 7 and 8 are repeated. We remark that the property that $v_j(f_*^\infty) = +\infty$ for all $j \in E_o$ is maintained. If step 9 is reached, then $p_{iaj} = 0$ for all triples (i,a,j) such that $i \in E \backslash E_o$, $a \in A(i)$, $j \in E_o$. Hence, the set $E \backslash E_o$ is closed under any policy, i.e. $p_{ij}^t(R) = 0$ for all $i \in E \backslash E_o$, $j \in E_o$, $t \in \mathbb{N}$, $R \in \mathcal{C}$. Therefore, we may repeat the algorithm on the state space $E \backslash E_o$. Since $E_o \neq \emptyset$ at each iteration, algorithm XII determines a pure and stationary optimal policy in at most N iterations. □

EXAMPLE 3.5.2. We shall show the working of algorithm XII in order to find a pure and stationary optimal policy for the model of figure 3.5.2.

Iteration 1:

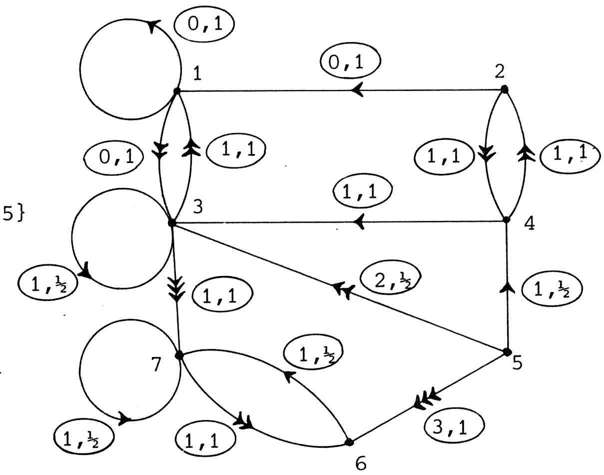
1. Starting the simplex method with $x = 0$, we find an infinite solution in the following simplex tableau (the column of x_{11} is deleted since all components are equal to zero):

		x_{12}	x_{21}	y_2	x_{31}	x_{32}	x_{33}	x_{41}	↓ x_{42}	x_{51}	x_{52}	x_{53}	x_{61}	x_{71}	x_{72}
y_1	1/7	1	-1			-1									
x_{22}	1/7		1	1					-1						
y_3	1/7	-1			½	1	1	-1			-½				
y_4	2/7		1	1				1		-½					
y_5	1/7									1	1	1			
y_6	1/7											-1	1		-1
y_7	1/7												-½	½	1
x_0	1/7		1	1	-1	-1	-1	-1	-2	-1	-2	-3	-1	-1	-1

3. $k = 4$, $a_o = 2$;
   $s_{22} = 1$, $s_{42} = 1$: $E_s = \{2,4\}$.
4. $f_*(2) = 2$, $f_*(4) = 2$.
5. $E_1 = \{2,4\}$
6. $E_o = \{2,4\}$
8. $p_{514} > 0$: $f_*(5) = 1$, $E_o = \{2,4,5\}$
9. $E = \{1,3,6,7\}$

Iteration 2:

1. Starting the simplex method with $x = 0$, we find an infinite solution in the following simplex tableau:



$\beta_i = 1/7 \quad i = 1,2,\ldots,7$

Figure 3.5.2

		$\downarrow x_{12}$	x_{31}	y_3	x_{33}	x_{61}	x_{71}	x_{72}
y_1	2/7		½	1	1			
x_{32}	1/7	-1	½	1	1			
y_6	1/7					1		-1
y_7	1/7				-1	-½	½	1
x_0	1/7	-1	-½	1		-1	-1	-1

3. $k = 1$, $a_o = 2$
   $s_{32} = 1$, $s_{12} = 1$: $E_s = \{1,3\}$
4. $f_*(1) = 2$, $f_*(3) = 2$
5. $E_1 = \{1,3\}$
6. $E_o = \{1,3\}$
7. $E = \{6,7\}$.

Iteration 3:

1. A finite optimal solution is obtained:
   $x^*_{61} = 4/7$, $x^*_{72} = 3/7$
2. $f_*(6) = 1$, $f_*(7) = 2$: f^∞_* is optimal, where
   $f_*(1) = 2$, $f_*(2) = 2$, $f_*(3) = 2$, $f_*(4) = 2$,
   $f_*(5) = 1$, $f_*(6) = 1$, $f_*(7) = 2$.

		y_6	x_{71}	y_7
x_{61}	4/7	2	1	2
x_{72}	3/7	1	1	2
x_0	1	3	1	4

REMARK 3.5.3. The results of the sections 3.5 and 3.6 are based on KALLENBERG [1983].

3.6. NEGATIVE DYNAMIC PROGRAMMING

Negative dynamic programming problems are dynamic programming problems which satisfy the following assumption.

ASSUMPTION 3.6.1. $r_{ia} \le 0 \quad a \in A(i),\ i \in E$.

If there exists a transient policy R, then we have $v(R) \le v \le 0$. Hence, the TMD-value-vector v is finite. In contrast with section 3.3, in this section we also allow nontransient policies. Intuitively, it is obvious that nontransient optimal policies must contain an ergodic set such that the corresponding rewards are zero for each state in this ergodic set. Such ergodic sets can be obtained from an average optimal policy. The computation of an average optimal policy by linear programming will be discussed in section 4.2 (see algorithm XIV). However, in chapter 4 we have assumed that the transition probabilities satisfy $\Sigma_j p_{iaj} = 1$ for all $i \in E$, $a \in A(i)$. Therefore, we have to use the extended TMD-model given by definition 3.2.2. The state space of the extended model is again denoted by E.

<u>THEOREM 3.6.1</u>. *Let* f_1^∞ *be any pure and stationary average optimal policy.*

(i) $v_i = -\infty$ *for every* i *such that* $\phi_i(f_1^\infty) < 0$.

(ii) $v_i = v_i(f_1^\infty) = 0$ *for every* i *such that* $\phi_i(f_1^\infty) = 0$ *and* i *is a recurrent state in the Markov chain induced by* $P(f_1)$.

<u>PROOF</u>.

(i) From (2.5.7) it follows that for any pure and stationary policy f^∞, we have

$$v^\alpha(f^\infty) = (1-\alpha)^{-1}\cdot\phi(f^\infty) + u(f^\infty) + \varepsilon(\alpha),$$

where $\varepsilon(\alpha) \to 0$ for $\alpha \uparrow 1$. Since $\phi(f^\infty) \le \phi(f_1^\infty)$ and $v(f^\infty) = \lim_{\alpha\uparrow 1} v^\alpha(f^\infty)$ (see lemma 3.2.1), we obtain

$$(3.6.1)\qquad v(f^\infty) = \lim_{\alpha\uparrow 1}\{(1-\alpha)^{-1}\cdot\phi(f^\infty) + u(f^\infty) + \varepsilon(\alpha)\}$$

$$\le \lim_{\alpha\uparrow 1}\{(1-\alpha)^{-1}\cdot\phi(f_1^\infty) + u(f^\infty) + \varepsilon(\alpha)\}.$$

Let $i \in E$ such that $\phi_i(f_1^\infty) < 0$. Then (3.6.1) implies that $v_i(f^\infty) = -\infty$. Since f^∞ is arbitrarily chosen and since there exists a pure and stationary optimal policy (theorem 3.2.1), it follows that $v_i = -\infty$.

(ii) Suppose that $\phi_i(f_1^\infty) = 0$ and $i \in E_k$, where E_k is an ergodic set in the Markov chain induced by $P(f_1)$. Then (cf. (2.4.3))

$$p_{ij}^*(f_1) > 0 \quad j \in E_k,\ p_{ij}^*(f_1) = 0 \quad j \notin E_k \quad \text{and}$$

$$p_{ij}^t(f_1) = 0 \quad j \notin E_k,\ t \in \mathbb{N}_0.$$

Since

$$0 = \phi_i(f_1^\infty) = \sum_j p^*_{ij}(f_1)r_j(f_1) = \sum_{j\in E_k} p^*_{ij}(f_1)r_j(f_1),$$

we get

$$r_j(f_1) = 0 \qquad j \in E_k.$$

Hence,

$$v_i(f_1^\infty) = \sum_{t=1}^\infty \sum_j p_{ij}^{t-1}(f_1)r_j(f_1) = \sum_{t=1}^\infty \sum_{j\in E_k} p_{ij}^{t-1}(f_1)r_j(f_1) = 0.$$

Consequently, $v_i = v_i(f_1^\infty) = 0$, completing the proof. □

From theorem 3.6.1 it follows that if $\phi_i(f_1^\infty) < 0$, where $f_1^\infty \in C_D$ is an average optimal policy then $v_i(R) = -\infty$ for all $R \in C$. Hence, for the determination of an optimal policy, we may remove from the state space the states in which $\phi_i(f_1^\infty) < 0$. Therefore, we may restrict ourselves to the states i in which $\phi_i(f_1^\infty) = 0$. We can find (e.g. by algorithm II) the set $R(f_1)$ of states that are recurrent in the Markov chain induced by $P(f_1)$. Theorem 3.6.1 implies that in the states of $R(f_1)$ f_1^∞ takes already optimal actions. If all states belong to $R(f_1)$, then we have found an optimal policy. Otherwise, we try to find in $E\backslash R(f_1)$ an ergodic set with respect to another average optimal policy, say f_2^∞. Therefore, we change the model in the following way

$$E := E\backslash R(f_1) \cup \{0\}$$

$$A(i) := \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases}$$

$$p_{iaj} := \begin{cases} p_{iaj} & i \neq 0,\ j \neq 0,\ a \in A(i) \\ \sum_{k\in R(f_1)} p_{iak} & i \neq 0,\ j = 0,\ a \in A(i) \\ 1 & i = 0,\ j = 0,\ a \in A(i) \\ 0 & i = 0,\ j \neq 0,\ a \in A(i) \end{cases}$$

$$r_{ia} := \begin{cases} r_{ia} & i \neq 0,\ a \in A(i) \\ -1 & i = 0,\ a \in A(i). \end{cases}$$

In this new model we compute an average optimal policy, say f_2^∞. Then, there are two possibilities:

1. $\phi_i(f_2^\infty) = 0$ for at least one state i:
   We remove the states j for which $\phi_j(f_2^\infty) < 0$. Let E_1 be the set of removed states. Then, the state 0 belongs to E_1.
   If the remaining state space coincides with $R(f_2)$, then $v_i(f_2^\infty) = 0$ for all remaining states, and consequently, f_2^∞ gives optimal actions for these states.
   Otherwise, we repeat the analysis described above to obtain recurrent states in $E \backslash R(f_2)$.
2. $\phi_i(f_2^\infty) < 0$ for all states i:
   Redefine $r_{01} := 0$, $p_{01j} := 0$ for all j. For the remaining states together with the set E_1 of already removed states, we compute an optimal transient policy by algorithm VI.

Every time that we encounter possibility 1, the state space decreases with at least one state. Hence, after a finite number of iterations either we have possibility 2 or we have an average optimal policy f_2^∞ such that all states i with $\phi_i(f_2^\infty) = 0$ are recurrent in the Markov chain induced by $P(f_2)$. Hence, the following algorithm gives in a finite number of iterations a pure and stationary policy f_*^∞. In theorem 3.6.2 we will show that f_*^∞ is an optimal policy.

ALGORITHM XIII for the construction of a pure and stationary policy in negative dynamic programming.

<u>step 1</u>: If $\Sigma_j\, p_{iaj} < 1$ for at least one pair (i,a), where $a \in A(i)$, $i \in E$, then construct the extended model in the following way:

$$E := E \cup \{0\}$$

$$A(i) := \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases}$$

$$p_{iaj} := \begin{cases} p_{iaj} & i \neq 0,\ j \neq 0,\ a \in A(i) \\ 1-\sum_{k\neq 0} p_{iak} & i \neq 0,\ j = 0,\ a \in A(i) \\ 0 & i = 0,\ j \neq 0,\ a \in A(i) \\ 1 & i = 0,\ j = 0,\ a \in A(i) \end{cases}$$

$$r_{ia} := \begin{cases} r_{ia} & i \neq 0,\ a \in A(i) \\ 0 & i = 0,\ a \in A(i). \end{cases}$$

step 2: Compute an average optimal policy f_1^∞ by algorithm XIV.

step 3a: Let $E_0 := \{i \mid \phi_i(f_1^\infty) < 0\}$; $E_1 := \emptyset$; $A_1(i) := A(i)$ for all $i \in E\backslash E_0$.

step 3b: Define $f_*(i) := f_1(i)$, $i \in E_0$.

step 3c: If $E_0 = E$, then go to step 9.
Otherwise, go to step 3d.

step 3d: For every $a \in A(i)$, where $i \in E\backslash E_0$, such that $\Sigma_{j \in E_0} p_{iaj} > 0$ do $A(i) := A(i)\backslash\{a\}$.

step 3e: $E := E\backslash E_0$.

step 4a: Determine by algorithm II the set $R(f_1)$ of the recurrent states of E in the Markov chain induced by $P(f_1)$.

step 4b: Define $f_*(i) := f_1(i)$, $i \in R(f_1)$.

step 4c: If $R(f_1) = E$, then go to step 7a.
Otherwise, go to step 4d.

step 4d: $E := E\backslash R(f_1) \cup \{0\}$

$$A(i) := \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases}$$

$$p_{iaj} := \begin{cases} p_{iaj} & i \neq 0,\ j \neq 0,\ a \in A(i) \\ \sum_{k \in R(f_1)} p_{iak} & i \neq 0,\ j = 0,\ a \in A(i) \\ 1 & i = 0,\ j = 0,\ a \in A(i) \\ 0 & i = 0,\ j \neq 0,\ a \in A(i) \end{cases}$$

$$r_{ia} := \begin{cases} r_{ia} & i \neq 0,\ a \in A(i) \\ -1 & i = 0,\ a \in A(i). \end{cases}$$

step 5: Compute an average optimal policy f_1^∞ by algorithm XIV.

step 6a: $E_2 := \{i \mid \phi_i(f_1^\infty) < 0\}$.

step 6b: If $E = E_2$, then $E_1 := E_1 \cup (E\backslash\{0\})$ and go to step 7a.
Otherwise, $E_1 := E_1 \cup (E_2\backslash\{0\})$ and go to step 6c.

step 6c: For every $a \in A(i)$, where $i \in E\backslash E_2$, such that $\Sigma_{j \in E_2} p_{iaj} > 0$ do $A(i) := A(i)\backslash\{a\}$.

step 6d: $E := E\backslash E_2$ and go to step 4a.

step 7a: If $E_1 = \emptyset$, then go to step 9.

Otherwise, go to step 7b.

step 7b: $E := E_1 \cup \{0\}$

$$A(i) := \begin{cases} A_1(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases}$$

$$p_{iaj} := \begin{cases} p_{iaj} & i \neq 0,\ j \neq 0,\ a \in A(i) \\ 1-\sum_{k\in E_1} p_{iak} & i \neq 0,\ j = 0,\ a \in A(i) \\ 0 & i = 0,\ j \in E,\ a \in A(i) \end{cases}$$

$$r_{ia} := \begin{cases} r_{ia} & i \neq 0,\ a \in A(i) \\ 0 & i = 0,\ a \in A(i) \end{cases}$$

step 8a: Compute an optimal transient policy f_o^∞ by algorithm VI.

step 8b: Define $f_*(i) := f_o(i) \quad i \in E$.

step 9: f_*^∞ is an optimal policy.

EXAMPLE 3.6.1. We illustrate algorithm XIII for the negative dynamic programming problem of figure 3.6.1 (without the dotted part).

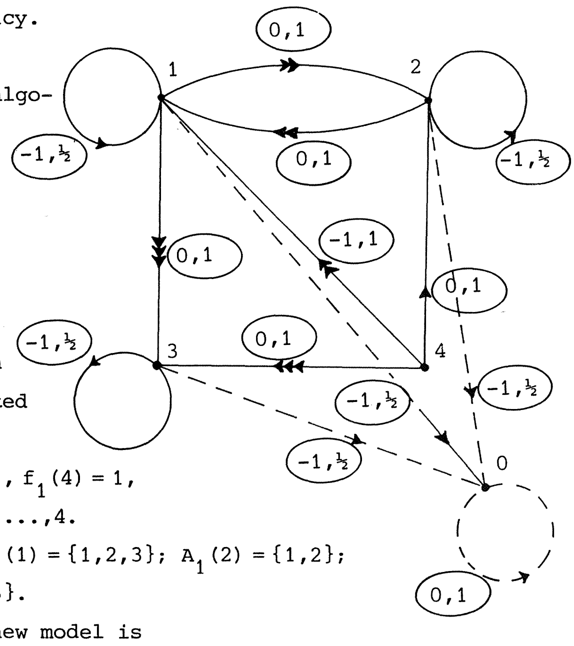


Figure 3.6.1

Iteration 1:

1. The extended model is drawn in figure 3.6.1 by the dotted lines.
2. $f_1(1) = 1$, $f_1(2) = 1$, $f_1(3) = 1$, $f_1(4) = 1$, $f_1(0) = 1$; $\phi_i(f_1^\infty) = 0$, $i = 0,1,\ldots,4$.
3. $E_0 = \emptyset$; $E_1 = \emptyset$; $A_1(0) = \{1\}$; $A_1(1) = \{1,2,3\}$; $A_1(2) = \{1,2\}$; $A_1(3) = \{1\}$; $A_1(4) = \{1,2,3\}$.
4. $R(f_1) = \{0\}$; $f_*(0) = 1$; the new model is the same as the old model except that $r_{01} = -1$.
5. $f_1(1) = 2$, $f_1(2) = 2$, $f_1(3) = 1$, $f_1(4) = 1$, $f_1(0) = 1$. $\phi_1(f_1^\infty) = \phi_2(f_1^\infty) = \phi_4(f_1^\infty) = 0$, $\phi_3(f_1^\infty) = \phi_0(f_1^\infty) = -1$.
6. $E_2 = \{3,0\}$; $E_1 = \{3\}$; $A(1) = \{2\}$, $A(2) = \{2\}$, $A(4) = \{1,2\}$; $E = \{1,2,4\}$.

Iteration 2:

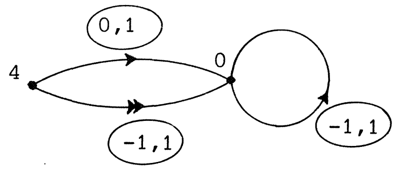


Figure 3.6.2

4. $R(f_1) = \{1,2\}$; $f_*(1) = f_*(2) = 2$; The model is reduced to the model of figure 3.6.2; $E = \{4,0\}$.
5. $f_1(4) = 1$, $f_1(0) = 1$; $\phi_4(f_1^\infty) = \phi_0(f_1^\infty) = -1$.
6. $E_2 = \{0,4\}$; $E_1 = \{3,4\}$
7. We obtain the model of figure 3.6.3
8. $f_o(3) = 1$, $f_o(4) = 1$, $f_o(0) = 1$.
   $f_*(3) = 1$, $f_*(4) = 1$
9. f_*^∞, where $f_*(1) = 2$, $f_*(2) = 2$, $f_*(3) = 1$, $f_*(4) = 1$, is an optimal policy.

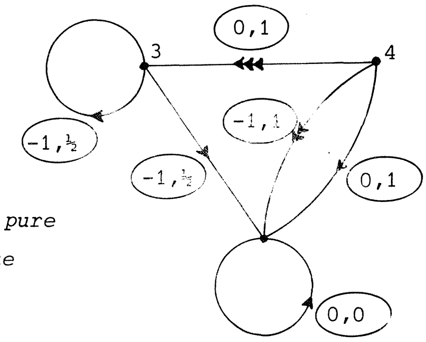


Figure 3.6.3

THEOREM 3.6.2. *Algorithm XIII determines a pure and stationary optimal policy* f_*^∞ *in a finite number of iterations.*

PROOF. First, we consider the finiteness. The only loop in the algorithm may possibly occur in the steps 4 until 6. However, each time that we go back to step 4, the number of states in E decreases, namely:

The model defined in step 4d has state 0 as absorbing state and $\phi_0(f_1^\infty) = -1$. Then $0 \in E_2$, where E_2 is defined in step 6a. Hence, if we go back to step 4a the number of states decreases by $|R(f_1)|$.

Consequently, algorithm XIII determines a pure and stationary policy f_*^∞ in a finite number of iterations. This policy f_*^∞ has the following properties:

(i) $v_i(f_*^\infty) = v_i = -\infty$ for all $i \in E_0$.

(ii) $v_i(f_*^\infty) = v_i = 0$ for all $i \in E\setminus(E_0 \cup E_1)$

(iii) f_*^∞ is an optimal transient policy in the model defined in step 7b.

We will show that f_o^∞, computed in step 8a, is an optimal policy for the model defined in step 7b. Suppose that f_o^∞ is not optimal. Then, there exists a nontransient optimal policy, say f_2^∞. Since f_2^∞ is nontransient, we have $R(f_2) \cap E_1 \neq \emptyset$. From the construction of E_1 (see step 6b) it follows that $\phi_i(f_2^\infty) < 0$, $i \in R(f_2)$. Then relation (3.6.1) implies that $v_i(f_2^\infty) = -\infty$, $i \in R(f_2)$, which is in contradiction with the assumption that f_2^∞ is optimal. Next, we will prove that f_*^∞ is an optimal policy. By the properties (i) and (ii), it is sufficient to show that $v_i(f_*^\infty) \geq v_i(f^\infty)$ for all $i \in E_1$ and all $f^\infty \in C_D$.

For any $f^\infty \in C_D$, let $\tilde{v}(f^\infty)$ be the expected total reward obtained in the model of step 7b when policy f^∞ is used. Since $r_j(f_*) = 0$ for every $j \in E\backslash (E_0 \cup E_1)$, we can write for $i \in E_1$ and $f^\infty \in C_D$:

$$\begin{aligned} v_i(f_*^\infty) &= \tilde{v}_i(f_*^\infty) \\ &\geq \tilde{v}_i(f^\infty) = \\ &\quad \text{expected total reward until a state of } E\backslash E_1 \text{ is reached} \\ &\geq \text{expected total reward over the infinite horizon} = \\ &\quad v_i(f^\infty). \end{aligned}$$

This completes the proof of the theorem. □

<u>REMARK 3.6.1</u>. From theorem 3.6.1 and relation (3.6.1) it follows that an optimal policy can also be obtained in the following way:

1. Construct the extended model with $\Sigma_j p_{iaj} = 1$ for all $i \in E$, $a \in A(i)$.
2. Compute an average optimal policy f_1^∞ by algorithm XIV.
3. Define $f_*(i) := f_1(i)$ for $i \in E_0 \cup E_2$, where

$$E_0 := \{j \mid \phi_j(f_1^\infty) < 0\} \quad \text{and} \quad E_2 := (E\backslash E_0) \cap R(f_1).$$

4. Construct the model with state space $E := E\backslash(E_0 \cup E_2) \cup \{0\}$ as in step 4d of algorithm XIII but with $r_{01} := 0$ instead of $r_{01} := -1$.
5. Compute a bias optimal policy f_2^∞ by algorithm XXII or XXIII presented in chapter 5, i.e. f_2^∞ satisfies

$$u(f_2^\infty) = \max\{u(f^\infty) \mid \phi(f^\infty) = 0\}.$$

6. Define $f_*(i) := f_2(i) \quad i \neq 0$.

The policy f_*^∞ is an optimal policy since for all states i and policies f^∞ such that $\phi_i(f^\infty) = 0$, we have (cf. (3.6.1)) $v_i(f^\infty) = u_i(f^\infty)$.

CHAPTER 4

AVERAGE REWARD CRITERION

4.1. INTRODUCTION AND SUMMARY

The linear programming approach for the average reward criterion was introduced by DE GHELLINCK [1960] and MANNE [1960]. They have proposed a linear program from which a pure and stationary optimal policy can be obtained if for any stationary policy π^{∞} the Markov chain induced by $P(\pi)$ is completely ergodic.

The first analysis of a linear program in the general multichain case has been presented in DENARDO & FOX [1968] and DENARDO [1970b]. DERMAN [1970] has streamlined and slightly improved their results. He has shown that, in order to find an optimal policy, there have to be solved two linear programming problems and one combinatorial problem, in the worst case. DIRICKX & RAO [1979] have shown that the second linear program may be replaced by a search procedure. In the first part of this chapter we will show that a pure and stationary optimal policy can be obtained by solving only the first one of the two linear programming problems introduced in DENARDO & FOX [1968].

In section 4.2 we review some relevant theorems which lead to the linear programming formulation. The main result is that a pure and stationary optimal policy can be obtained directly from any extreme optimal solution of the linear program. Since the simplex method gives such an extreme optimal solution, we have an elegant algorithm for the construction of a pure and stationary average optimal policy in the multichain case.

In section 4.3 we study the correspondence between feasible solutions of the linear program and stationary policies of the Markov decision problem. In contrast with the contracting dynamic programming problem, it is not possible to construct a one-to-one correspondence between these feasible solutions and these (randomized) stationary policies. As it turns out, we have to use equivalence classes of feasible solutions. We will construct a one-to-one correspondence between the stationary policies and the representatives of the equivalence classes. Furthermore, this mapping preserves the optimality property, i.e. optimal solutions are mapped on optimal policies and optimal policies correspond to representatives which are optimal solutions of the linear program.

Then in section 4.4 we compare the linear programming approach with the policy improvement algorithm. We can conclude that the policy improvement algorithm is equivalent to an algorithm for the optimal solution of the linear program in which successive solutions are extreme but not necessarily adjacent points of the set of feasible solutions. Such an algorithm is called a block-pivoting algorithm.

In the sections 4.5 and 4.6 we give simplified algorithms for some special models. In section 4.5 we discuss the case that the Markov chain induced by P(f), where f^∞ is any pure and stationary optimal policy, is unichained. Then a pure and stationary optimal policy can be obtained by the solution of a linear program, that needs half of the number of constraints and variables in comparison with the program used in section 4.2, plus a simple search procedure.

Section 4.6 deals with the completely ergodic as well as with the unichain case. In both cases a pure and stationary average optimal policy can be obtained directly (without the search procedure) from the smaller linear program used in section 4.5.

We close this chapter with a discussion about the constrained Markov decision model. In this model there are some additional constraints for the limit points of the expected state - action frequencies. Such models are of importance e.g. if there are more than one reward or cost functions. In contrast with the policy improvement method and the method of successive approximations, the linear programming method can also solve this kind of models. In general, these models have no stationary optimal policies. First, we shall prove some properties of the set of limit points of the state-action frequencies. We present an algorithm for the construction of an average optimal policy for a constrained Markov decision problem. However, this algorithm requires an enormous quantity of calculations. Fortunately, in many cases an optimal stationary policy can be computed. We give sufficient conditions for the existence of optimal stationary policies. These conditions include the unichain case. We also present an algorithm by which a stationary, but not necessary optimal, policy is computed. We give some numerical results about its performance. These results are very encouraging; a stationary optimal policy was always found, if one exists, for the 400 test problems that we have analysed. In the unichain case a stationary optimal policy always exists and we present a simple algorithm to construct one.

The results of the sections 4.2, 4.3 and 4.7 are based upon HORDIJK & KALLENBERG [1978a], [1978b], [1979a], [1979b] and [1981b].

4.2. LINEAR PROGRAMMING FORMULATION

We assume in this chapter that $\Sigma_j p_{iaj} = 1$ for every pair (i,a) $a \in A(i)$, $i \in E$. If this assumption is not satisfied, then we can change the model into the extended model as described in definition 3.2.2. From definition 3.2.2 and the analysis in section 2.4 it follows that $\phi(\pi^\infty) = \tilde{\phi}(\pi^\infty)$ for every $\pi^\infty \in C_S$, where $\tilde{\phi}(\pi^\infty)$ denotes the expected average reward in the extended model. Since there exists a stationary average optimal policy (cf. corollary 2.5.2), the assumption that $\Sigma_j p_{iaj} = 1$ for every pair (i,a) $a \in A(i)$, $i \in E$ is no real restriction for the determination of an average optimal policy.

Before we give the linear program from which an average optimal policy can be obtained, we first present some theorems and we introduce the concept of superharmonicity for the AMD-model.

<u>THEOREM 4.2.1</u>. *Let f_o^∞ be a Blackwell optimal pure and stationary policy. Then $\phi^o := \phi(f_o^\infty)$ and $u^o := D(f_o)r(f_o)$ satisfy the pair of optimality equations*

$$(4.2.1) \qquad \tilde{\phi}_i = \max_{a \in A(i)} \sum_j p_{iaj} \tilde{\phi}_j, \qquad i \in E.$$

$$(4.2.2) \qquad \tilde{\phi}_i + \tilde{u}_i = \max_{a \in \bar{A}(i)} \{r_{ia} + \sum_j p_{iaj} \tilde{u}_j\}, \qquad i \in E,$$

where $\bar{A}(i) := \{a \in A(i) \mid \tilde{\phi}_i = \Sigma_j p_{iaj} \tilde{\phi}_j\}$, $i \in E$.

<u>PROOF</u>. Since f_o^∞ is a Blackwell optimal policy, there exists a nonnegative real number $\alpha_o < 1$ such that $v^\alpha(f_o^\infty) = v^\alpha$ for all $\alpha \in [\alpha_o, 1)$. From theorem 3.4.1 it follows that

$$v_i^\alpha(f_o^\infty) \geq r_{ia} + \alpha \sum_j p_{iaj} v_j^\alpha(f_o^\infty) \qquad a \in A(i),\ i \in E,\ \alpha \in [\alpha_o, 1).$$

Equation (2.5.7) implies that

$$(4.2.3) \qquad v_i^\alpha(f_o^\infty) = (1-\alpha)^{-1} \phi_i^o + u_i^o + \varepsilon_i(\alpha), \qquad i \in E,$$

where $\lim_{\alpha\uparrow 1} \varepsilon_i(\alpha) = 0$, $i \in E$.
Hence, we obtain

$$(1-\alpha)^{-1}\phi_i^\circ + u_i^\circ + \varepsilon_i(\alpha) \geq$$

$$r_{ia} + \alpha \sum_j p_{iaj}\{(1-\alpha)^{-1}\phi_j^\circ + u_j^\circ + \varepsilon_j(\alpha)\} =$$

$$r_{ia} + \{1-(1-\alpha)\} \sum_j p_{iaj}\{(1-\alpha)^{-1}\phi_j^\circ + u_j^\circ + \varepsilon_j(\alpha)\} =$$

$$(1-\alpha)^{-1}\sum_j p_{iaj}\phi_j^\circ + r_{ia} + \sum_j p_{iaj}u_j^\circ - \sum_j p_{iaj}\phi_j^\circ + \delta_i(\alpha),$$

where

$$\delta_i(\alpha) := \sum_j p_{iaj}\{\varepsilon_j(\alpha) - (1-\alpha)u_j^\circ - (1-\alpha)\varepsilon_j(\alpha)\}, \quad a \in A(i),\ i \in E,\ \alpha \in [\alpha_0, 1).$$

Therefore, $\lim_{\alpha\uparrow 1}\delta_i(\alpha) = 0$, $i \in E$, and we get

$$\phi_i^\circ \geq \sum_j p_{iaj}\phi_j^\circ \qquad a \in A(i),\ i \in E,$$

$$u_i^\circ \geq r_{ia} + \sum_j p_{iaj}u_j^\circ - \sum_j p_{iaj}\phi_j^\circ = r_{ia} + \sum_j p_{iaj}u_j^\circ - \phi_i^\circ \qquad a \in \bar{A}(i),\ i \in E.$$

For the actions $a_i := f_\circ(i)$, $i \in E$, we have equality since by theorem 2.4.1 we obtain

$$\phi^\circ = \phi(f_\circ^\infty) = P^*(f_\circ)r(f_\circ) = P(f_\circ)P^*(f_\circ)r(f_\circ) = P(f_\circ)\phi^\circ$$

and

$$\phi^\circ + u^\circ - P(f_\circ)u^\circ = P^*(f_\circ)r(f_\circ) + (I-P(f_\circ))D(f_\circ)r(f_\circ)$$

$$= P^*(f_\circ)r(f_\circ) + (I-P^*(f_\circ))r(f_\circ) = r(f_\circ).$$

Consequently, we have proved that

$$\phi_i^\circ = \max_{a\in A(i)} \sum_j p_{iaj}\phi_j^\circ, \qquad i \in E,$$

$$\phi_i^\circ + u_i^\circ = \max_{a\in \bar{A}(i)}\{r_{ia} + \sum_j p_{iaj}u_j^\circ\}, \qquad i \in E. \qquad \square$$

<u>DEFINITION 4.2.1</u>. A vector $\tilde{\phi} \in \mathbb{R}^N$ is *AMD-superharmonic* if there exists a vector $\tilde{u} \in \mathbb{R}^N$ such that

$$(4.2.4) \qquad \tilde{\phi}_i \geq \sum_j p_{iaj}\tilde{\phi}_j \qquad a \in A(i),\ i \in E,$$

and

$$(4.2.5) \qquad \tilde{\phi}_i + \tilde{u}_i \geq r_{ia} + \sum_j p_{iaj}\tilde{u}_j \qquad a \in A(i),\ i \in E.$$

REMARK 4.2.1. The inequalities (4.2.4) and (4.2.5) have to hold for all actions. Since in (4.2.2) the inequalities have to be satisfied only for the actions which yield equality in the first set of equations, the AMD-superharmonicity is a stronger condition.

THEOREM 4.2.2. *The AMD-value-vector ϕ is the smallest AMD-superharmonic vector.*

PROOF. Let f_o^∞ be any Blackwell optimal pure and stationary policy. From the property that f_o^∞ is average optimal (cf. theorem 2.5.4) and from theorem 4.2.1 it follows that

$$\phi_i \geq \sum_j p_{iaj}\phi_j \qquad a \in A(i),\ i \in E,$$

$$\phi_i + u_i^o \geq r_{ia} + \sum_j p_{iaj}u_j^o \qquad a \in \bar{A}(i),\ i \in E,$$

where

$$u_i^o := u_i(f_o^\infty) \quad \text{and} \quad \bar{A}(i) := \{a \in A(i) \mid \phi_i = \sum_j p_{iaj}\phi_j\}, \qquad i \in E.$$

Define:

$$A^*(i) := \{a \in A(i) \mid \phi_i + u_i^o < r_{ia} + \sum_j p_{iaj}u_j^o\} \qquad i \in E.$$

$$s_{ia} := \phi_i - \sum_j p_{iaj}\phi_j \qquad a \in A(i),\ i \in E.$$

$$t_{ia} := \phi_i + u_i^o - r_{ia} - \sum_j p_{iaj}u_j^o \qquad a \in A(i),\ i \in E.$$

Then, $A^*(i) \cap \bar{A}(i) = \emptyset$, $i \in E$, and
$s_{ia} \geq 0$ $a \in A(i), i \in E$; $s_{ia} > 0$ $a \notin \bar{A}(i), i \in E$.
$t_{ia} \geq 0$ $a \notin A^*(i), i \in E$; $t_{ia} < 0$ $a \in A^*(i), i \in E$.

$A^*(i)$		$\bar{A}(i)$
$s_{ia} > 0$	$s_{ia} > 0$	$s_{ia} = 0$
$t_{ia} < 0$	$t_{ia} \geq 0$	$t_{ia} \geq 0$

Let

$$M := \min\left\{\frac{t_{ia}}{s_{ia}} \;\middle|\; a \in A^*(i),\ i \in E\right\}, \quad \text{and } u := u^o - M\phi$$

(if $A^*(i) = \emptyset$ for all $i \in E$, then we define $M := 0$)
For $a \in \bar{A}(i)$, we have

$$\phi_i = \sum_j p_{iaj}\phi_j$$

and

$$\phi_i + u_i = \phi_i + u_i^\circ - M \sum_j p_{iaj}\phi_j \geq r_{ia} + \sum_j p_{iaj}u_j^\circ - M \sum_j p_{iaj}\phi_j =$$
$$r_{ia} + \sum_j p_{iaj}u_j.$$

For $a \in A^*(i)$, we obtain

$$\phi_i > \sum_j p_{iaj}\phi_j$$

and

$$\phi_i + u_i = \phi_i + u_i^\circ - M(s_{ia} + \sum_j p_{iaj}\phi_j)$$
$$= r_{ia} + \sum_j p_{iaj}(u_j^\circ - M\phi_j) + (t_{ia} - Ms_{ia})$$
$$\geq r_{ia} + \sum_j p_{iaj}u_j.$$

If $a \notin A^*(i) \cup \bar{A}(i)$ then we get

$$\phi_i > \sum_j p_{iaj}\phi_j$$

and

$$\phi_i + u_i = \phi_i + u_i^\circ - M\phi_i = t_{ia} + r_{ia} + \sum_j p_{iaj}u_j^\circ - M\phi_i$$
$$\geq t_{ia} + r_{ia} + \sum_j p_{iaj}(u_j^\circ - M\phi_j) \geq r_{ia} + \sum_j p_{iaj}u_j.$$

Hence, we have shown that the AMD-value-vector ϕ is AMD-superharmonic. Suppose that $\tilde{\phi} \in \mathbb{R}^N$ is also AMD-superharmonic. Then (4.2.4) implies that

(4.2.6) $$\tilde{\phi} \geq P(f_o)\tilde{\phi}.$$

By iterating (4.2.6) we obtain $\tilde{\phi} \geq P^t(f_o)\tilde{\phi}$ for all $t \in \mathbb{N}$, and consequently

(4.2.7) $$\tilde{\phi} \geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^t(f_o)\tilde{\phi} = P^*(f_o)\tilde{\phi}.$$

From (4.2.5) it follows that

(4.2.8) $$P^*(f_o)(\tilde{\phi}+\tilde{u}) \geq P^*(f_o)(r(f_o) + P(f_o)\tilde{u}) = \phi(f_o^\infty)+P^*(f_o)\tilde{u} =$$
$$\phi + P^*(f_o)\tilde{u}.$$

Then, using (4.2.7) and (4.2.8), we can complete the proof as follows.

$$\tilde{\phi} \geq P^*(f_o)\tilde{\phi} \geq \phi,$$

i.e. ϕ is the smallest AMD-superharmonic vector. □

Next, we shall show that a pure and stationary average optimal policy is also an optimal policy if the following stronger criterion should be used:

(4.2.9) $$\hat{\phi}_i(R) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_j \sum_a \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja}, \quad i \in E.$$

Notice that for any $\pi^\infty \in C_S$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_j \sum_a \mathbb{P}_{\pi^\infty}(X_t = j,\ Y_t = a \mid X_1 = i) \cdot r_{ja} =$$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [P^{t-1}(\pi) r(\pi)]_i = [P^*(\pi) r(\pi)]_i, \quad i \in E,$$

and consequently,

$$\phi(\pi^\infty) = \hat{\phi}(\pi^\infty) = P^*(\pi) r(\pi).$$

THEOREM 4.2.3. *Let f^∞ be any pure and stationary average optimal policy. Then, $\hat{\phi}(f^\infty) \geq \hat{\phi}(R)$ for all $R \in C$.*

PROOF. From theorem 2.5.1 it follows that it is sufficient to prove that

$$\hat{\phi}(f^\infty) \geq \hat{\phi}(R) \quad \text{for all } R \in C_M.$$

Let $R = (\pi^1, \pi^2, \ldots)$ be an arbitrarily chosen Markov policy. Since ϕ is AMD-superharmonic, there exists a $u \in \mathbb{R}^N$ such that

$$\phi_i \pi^t_{ia} \geq \sum_j p_{iaj} \pi^t_{ia} \phi_j \qquad a \in A(i),\ i \in E,\ t \in \mathbb{N},$$

and

$$\phi_i \pi^t_{ia} + u_i \pi^t_{ia} \geq r_{ia} \pi^t_{ia} + \sum_j p_{iaj} \pi^t_{ia} u_j \qquad a \in A(i),\ i \in E,\ t \in \mathbb{N}.$$

Consequently,

$$P(\pi^t)\phi \leq \phi \quad \text{and} \quad r(\pi^t) \leq \phi + u - P(\pi^t) u \qquad t \in \mathbb{N}.$$

Hence, we obtain

$$\sum_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t) \le \sum_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1}).$$

$$\{\phi+u-P(\pi^t)u\} \le T\cdot\phi + u - P(\pi^1)P(\pi^2)\cdots P(\pi^T)u \qquad T \in \mathbb{N}.$$

Since $\frac{1}{T}\{u-P(\pi^1)P(\pi^2)\cdots P(\pi^T)u\} \to 0$ for $T \to \infty$, we can write

$$\hat{\phi}_i(R) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} [P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)]_i$$

$$\le \limsup_{T\to\infty} \frac{1}{T} \{T\cdot\phi + u - P(\pi^1)P(\pi^2)\cdots P(\pi^T)u\}_i =$$

$$\phi_i = \phi_i(f^\infty), \quad i \in E.$$

This completes the proof. □

COROLLARY 4.2.1. *Any pure and stationary average optimal policy is also optimal for the stronger criterion with utility function* (4.2.9).

REMARK 4.2.2. In DERMAN [1970] p.26 the above result is also mentioned. However, as was pointed out by HORDIJK & TIJMS [1970] p.93, Derman's proof is incorrect.

We will formulate a pair of dual linear programs and we will show that a pure and stationary average optimal policy can be obtained from the optimal solution of the dual program. Since ϕ is the smallest AMD-superharmonic vector, it is plausible to consider the following linear programming problem

$$(4.2.10) \qquad \min\left\{\sum_j \beta_j \tilde{\phi}_j \;\middle|\; \begin{array}{ll} \tilde{\phi}_i \ge \sum_j p_{iaj}\tilde{\phi}_j & a \in A(i),\ i \in E \\ \tilde{\phi}_i + \tilde{u}_i \ge r_{ia} + \sum_j p_{iaj}\tilde{u}_j & a \in A(i),\ i \in E \end{array}\right\}$$

where $\beta_j > 0$, $j \in E$, are given numbers with $\Sigma_j\beta_j = 1$. The dual linear programming problem is

$$(4.2.11) \qquad \max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} & & = 0\ ,\ j \in E \\ \sum_a x_{ja} & + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} & = \beta_j,\ j \in E \\ & x_{ia}, y_{ia} \ge 0 & a \in A(i),\ i \in E \end{array}\right\}$$

REMARK 4.2.3. From theorem 4.2.2 it follows that there exists a vector

$u \in \mathbb{R}^N$ such that (ϕ,u) is an optimal solution of the primal program (4.2.10). Then theorem 1.3.4 implies that the dual program (4.2.11) has also an optimal solution, say (x^*,y^*), which satisfies $\Sigma_j\beta_j\phi_j = \Sigma_i\Sigma_a r_{ia}x^*_{ia}$.

THEOREM 4.2.4. *If* (x^*,y^*) *is an optimal solution of the linear program* (4.2.11) *such that* (x^*,y^*) *is an extreme point of the set of feasible solutions, then the policy* f_*^∞, *where*

$$f_*(i) := a_i \text{ such that} \begin{cases} x^*_{ia_i} > 0 & i \in E_{x^*} \\ y^*_{ia_i} > 0 & i \in E\backslash E_{x^*}, \end{cases}$$

is an average optimal policy.

REMARK 4.2.4. The above theorem says that an optimal policy is obtained by taking an arbitrary action for which the x^*-variable is positive, if possible; otherwise, by taking an arbitrary action for which the y^*-variable is positive. Indeed, it is possible to obtain an optimal solution where in some states there is more than one positive variable (see example 4.2.1). In that case we can construct different policies. Any of these policies is average optimal.

PROOF OF THEOREM 4.2.4. From the constraints of program (4.2.11) it follows that

$$\sum_a x^*_{ja} + \sum_a y^*_{ja} = \beta_j + \sum_i\sum_a p_{iaj}y^*_{ia} \geq \beta_j > 0, \quad j \in E.$$

Hence, the policy f_*^∞ is well-defined. Let (ϕ,u) be an optimal solution of the primal problem (4.2.10).

The remaining part of the proof has the following structure. First, we give three separate propositions. After presenting the proofs of these propositions, we complete the proof of the theorem by some final conclusions.

PROPOSITION 4.2.1.

$$\sum_j(\delta_{ij}-p_{if_*(i)j})\phi_j = 0 \qquad i \in E,$$

$$\phi_i + \sum_j(\delta_{ij}-p_{if_*(i)j})u_j = r_i(f_*) \qquad i \in E_{x^*}.$$

PROOF. Since $x^*_{if_*(i)} > 0$, $i \in E_{x^*}$, and $y^*_{if_*(i)} > 0$, $i \in E\backslash E_{x^*}$, it follows from the complementary slackness property of linear programming (see corollary 1.3.1) that

$$\sum_j(\delta_{ij}-p_{if_*(i)j})\phi_j = 0 \qquad i \in E\backslash E_{x^*},$$

and

$$\phi_i + \sum_j(\delta_{ij}-p_{if_*(i)j})u_j = r_i(f_*) \qquad i \in E_{x^*}.$$

The primal program (4.2.10) implies

$$\sum_j(\delta_{ij}-p_{iaj})\phi_j \geq 0 \qquad a \in A(i),\ i \in E.$$

Suppose that

$$\sum_j(\delta_{kj}-p_{kf_*(k)j})\phi_j > 0 \qquad \text{for some } k \in E_{x^*}.$$

Since $x^*_{kf_*(k)} > 0$, we obtain

$$\sum_j(\delta_{kj}-p_{kf_*(k)j})\phi_j \cdot x^*_{kf_*(k)} > 0.$$

Furthermore, we have

$$\sum_j(\delta_{ij}-p_{iaj})\phi_j \cdot x^*_{ia} \geq 0 \qquad a \in A(i),\ i \in E.$$

Hence, we get

$$\sum_i\sum_a\sum_j(\delta_{ij}-p_{iaj})\phi_j \cdot x^*_{ia} > 0.$$

On the other hand, it follows from the constraints of program (4.2.11) that

$$\sum_i\sum_a\sum_j(\delta_{ij}-p_{iaj})\phi_j \cdot x^*_{ia} = \sum_j\{\sum_i\sum_a(\delta_{ij}-p_{iaj})x^*_{ia}\}\cdot\phi_j = 0.$$

This contradiction implies that $\Sigma_j(\delta_{ij}-p_{if_*(i)j})\phi_j = 0$, $i \in E_{x^*}$, which completes the proof.

<u>PROPOSITION 4.2.2</u>. E_{x^*} *is closed under* $P(f_*)$, *i.e.* $p_{if_*(i)j} = 0$ $i \in E_{x^*}$, $j \notin E_{x^*}$.

<u>PROOF</u>. Suppose that $p_{kf_*(k)\ell} > 0$ for some $k \in E_{x^*}$ and $\ell \in E\backslash E_{x^*}$. From the constraints of program (4.2.11) it follows that

$$0 = \sum_a x^*_{\ell a} = \sum_i\sum_a p_{ia\ell}x^*_{ia} \geq p_{kf_*(k)\ell}x^*_{kf_*(k)} > 0,$$

implying a contradiction.

PROPOSITION 4.2.3. *The states of* $E\backslash E_{x^*}$ *are transient in the Markov chain induced by* $P(f_*)$.

PROOF. Suppose that there is a state $j \in E\backslash E_{x^*}$ which is nontransient. Since E_{x^*} is closed under $P(f_*)$ (see proposition 4.2.2), there has to exist a non-empty set $J \subset E\backslash E_{x^*}$ which is ergodic. Because (x^*,y^*) is an extreme point and $y^*_{jf_*(j)} > 0$, $j \in J$, theorem 1.1.2 implies that the corresponding columns $\{q^j,\ j \in J\}$, where

$$q^j_k := \begin{cases} 0 & k = 1,2,\dots,N \\ \delta_{j(k-N)} - p_{jf_*(j)(k-N)} & k = N+1,N+2,\dots,2N, \end{cases}$$

are linearly independent. Let $J = \{j_1,j_2,\dots,j_m\}$. Since J is an ergodic set, we have

$$0 = p_{jf_*(j)(k-N)} = \delta_{j(k-N)} \qquad j \in J,\ k-N \notin J.$$

Hence, $q^j_k = 0$ for all $k \notin \{N+j_1,N+j_2,\dots,N+j_m\}$. Therefore, the vectors $\{b^1,b^2,\dots,b^m\}$, where

$$b^i_k := q^{j_i}_{N+j_k} \qquad i,k = 1,2,\dots,m,$$

are also linearly independent. However,

$$\begin{aligned} \textstyle\sum_{k=1}^m b^i_k &= \textstyle\sum_{k=1}^m (\delta_{j_i j_k} - p_{j_i f_*(j_i) j_k}) \\ &= 1 - \textstyle\sum_{k=1}^m p_{j_i f_*(j_i) j_k} \\ &= 1 - \textstyle\sum_k p_{j_i f_*(j_i) k} = 0, \end{aligned}$$

which is contradictory to the independency of $\{b^1,b^2,\dots,b^m\}$. This completes the proof of the proposition 4.2.3.

Now, we can finish the proof of theorem 4.2.4 by the following arguments. From proposition 4.2.1 it follows that $P(f_*)\phi = \phi$ and consequently, $P^*(f_*)\phi = \phi$. Since the states of $E\backslash E_{x^*}$ are transient under $P(f_*)$ (see proposition 4.2.3), we have $p^*_{ik}(f^*) = 0$, $i \in E$, $k \in E\backslash E_{x^*}$.
Hence,

$$\phi_i(f_*^\infty) = (P^*(f_*)r(f_*))_i$$
$$= \sum_k p_{ik}^*(f_*)r_k(f_*)$$
$$= \sum_{k\in E_{x^*}} p_{ik}^*(f_*)\{\phi_k+\sum_j(\delta_{kj}-p_{kf_*(k)j})u_j\}$$
$$= \sum_k p_{ik}^*(f_*)\phi_k + \sum_j\{\sum_k p_{ik}^*(f_*)\cdot(\delta_{kj}-p_{kf_*(k)j})\}u_j, \quad i \in E.$$

Because $P^*(f_*)\phi = \phi$ and $P^*(f_*)(I-P(f_*)) = 0$ (cf. theorem 2.4.1), we obtain

$$\phi_i(f_*^\infty) = \phi_i, \quad i \in E,$$

i.e. f_*^∞ is an average optimal policy. □

The solution of a linear program by the simplex method always gives an optimal solution which is an extreme point of the set of feasible solutions. Hence, the above theorem implies that a pure and stationary average optimal policy is obtained by the following algorithm.

ALGORITHM XIV for the construction of a pure and stationary average optimal policy (multichain case).

step 1: Take any choice for the numbers β_j such that $\beta_j > 0$, $j \in E$, and $\Sigma_j\beta_j = 1$.

step 2: Use the simplex method to compute an optimal solution (x^*,y^*) of the linear programming problem

$$\max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{lr} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} & = 0\ ,\ j\in E \\ \sum_a x_{ja} + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} & = \beta_j,\ j\in E \\ x_{ia},y_{ia} \ge 0 & a\in A(i),\ i\in E \end{array}\right\}$$

step 3: For each $i \in E$ choose an arbitrary action a_i from the set $A^*(i)$, where

$$A^*(i) := \begin{cases} \{a \mid x_{ia}^* > 0\} & i \in E_{x^*} \\ \{a \mid y_{ia}^* > 0\} & i \in E\backslash E_{x^*} \end{cases}$$

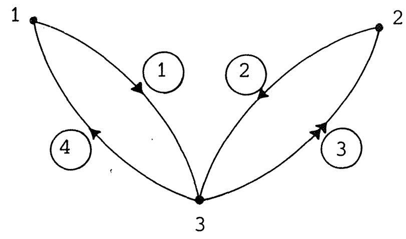
step 4: f^∞, where $f(i) := a_i$, $i \in E$, is a pure and stationary average optimal policy.

<u>EXAMPLE 4.2.1</u>. The data of the model can be found in figure 4.2.1 and should be interpreted as exposed in remark 2.2.1. The linear program is:

maximize $x_{11} + 2x_{21} + 4x_{31} + 3x_{22}$

subject to

$$
\begin{array}{l}
x_{11} - x_{31} = 0 \\
x_{21} - x_{32} = 0 \\
-x_{11} - x_{21} + x_{31} + x_{32} = 0 \\
x_{11} + y_{11} - y_{31} = \tfrac{1}{4} \\
x_{21} + y_{21} - y_{32} = \tfrac{1}{4} \\
x_{31} + x_{32} - y_{11} - y_{21} + y_{31} + y_{32} = \tfrac{1}{2} \\
x_{11}, x_{21}, x_{31}, x_{32}, y_{11}, y_{21}, y_{31}, y_{32} \geq 0.
\end{array}
$$



$\beta_1 = \beta_2 = \tfrac{1}{4},\ \beta_3 = \tfrac{1}{2}$

<u>Figure 4.2.1</u>

The solution (x^*, y^*), where $x^*_{11} = x^*_{21} = x^*_{31} = x^*_{32} = \tfrac{1}{4}$, $y^*_{11} = y^*_{21} = y^*_{31} = y^*_{32} = 0$, is an extreme point of the set of feasible solutions and is also an optimal solution. In state 3 there are two actions for which the corresponding variables x^*_{31} and x^*_{32} are positive. Hence, we can construct two pure and stationary average optimal policies, namely f_1^∞ and f_2^∞, where $f_1(1) = f_2(1) = f_1(2) = f_2(2) = f_1(3) = 1$ and $f_2(3) = 2$.

<u>REMARK 4.2.5</u>. For every optimal solution (x^*, y^*) which is an extreme point of the set of feasible solutions, we define

$$
A_i^*\{(x^*, y^*)\} := \begin{cases} \{a \mid x^*_{ia} > 0\} & i \in E_{x^*} \\ \{a \mid y^*_{ia} > 0\} & i \in E \backslash E_{x^*} \end{cases}
$$

$$
F^*\{(x^*, y^*)\} := \{f^\infty \in C_D \mid f(i) \in A_i^*\{(x^*, y^*)\}, \quad i \in E\}
$$

$$
F^* := \cup F^*\{(x^*, y^*)\}.
$$

From theorem 4.2.4 it follows that any $f^\infty \in F^*$ is average optimal. Conversely, for any pure and stationary optimal policy f^∞, there is an extreme optimal solution $(x(f), y(f))$ such that $f^\infty \in F^*\{(x(f), y(f))\}$ (this fact is shown in the theorems 4.3.3 and 4.3.4). Hence, all pure and stationary optimal policies can be determined by the computation of all extreme optimal solutions of program (4.2.11). In chapter 1 we have derived an algorithm to perform this computation (algorithm I).

4.3. RELATIONS BETWEEN STATIONARY POLICIES AND FEASIBLE SOLUTIONS

For any feasible solution (x,y) of the linear programming problem (4.2.11) we define a stationary policy $\pi^{\infty}(x,y)$ by

$$
(4.3.1)\qquad \pi_{ia}(x,y) := \begin{cases} x_{ia}\Big/\sum_a x_{ia} & a \in A(i),\ i \in E_x \\ y_{ia}\Big/\sum_a y_{ia} & a \in A(i),\ i \in E\backslash E_x. \end{cases}
$$

Unfortunately, in contrast with the contracting dynamic programming model, in the AMD-model it is possible that two different feasible solutions are mapped on the same stationary policy. We give an example.

EXAMPLE 4.3.1. Figure 4.3.1 presents the AMD-model. The formulation of the linear program becomes:

maximize $x_{11}+x_{21}+x_{22}+x_{31}+x_{32}+x_{41}$

subject to

$$
\begin{array}{rl}
x_{11} - x_{32} & = 0 \\
-x_{11} + x_{21} + x_{22} & = 0 \\
-x_{21} + x_{32} & = 0 \\
-x_{22} & = 0 \\
x_{11} + y_{11} - y_{32} & = \tfrac14 \\
x_{21} + x_{22} - y_{11} + y_{21} + y_{22} & = \tfrac14 \\
x_{31} + x_{32} - y_{21} + y_{32} & = \tfrac14 \\
x_{41} - y_{22} & = \tfrac14 \\
x_{11},x_{21},x_{22},x_{31},x_{32},x_{41},y_{11},y_{21},y_{22},y_{32} & \ge 0
\end{array}
$$

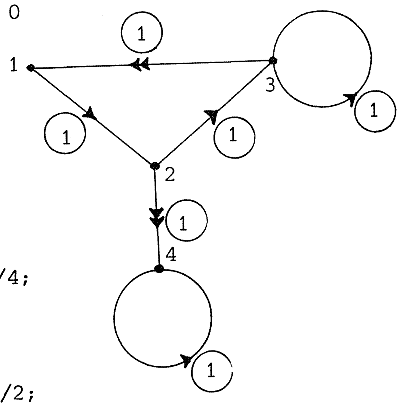
The following two feasible solutions (x^1,y^1) and (x^2,y^2) are mapped on the same pure and stationary policy f^{∞}, where $f(1)=f(2)=1$, $f(3)=2$ and $f(4)=1$:

$x^1_{11}=1/4, x^1_{21}=1/4, x^1_{22}=0, x^1_{31}=0, x^1_{32}=1/4, x^1_{41}=1/4;$

$y^1_{11}=0,\ y^1_{21}=0,\ y^1_{22}=0,\ y^1_{32}=0.$

$x^2_{11}=1/6, x^2_{21}=1/6, x^2_{22}=0, x^2_{31}=0, x^2_{32}=1/6, x^2_{41}=1/2;$

$y^2_{11}=1/6,\ y^2_{21}=0,\ y^2_{22}=1/4,\ y^2_{32}=1/12.$



$\beta_1=\beta_2=\beta_3=\beta_4=\tfrac14$

Figure 4.3.1

Since there is no one-to-one correspondence between the stationary policies and the feasible solutions of the linear programming problem (4.2.11), we introduce an equivalence relation. We call two feasible solutions (x^1,y^1) and (x^2,y^2) *equivalent* if $\pi_{ia}(x^1,y^1) = \pi_{ia}(x^2,y^2)$ $a \in A(i)$, $i \in E$. This equivalence relation divides the set of feasible solutions in equivalence classes.

Conversely, let π^∞ be a stationary policy. Consider the Markov chain induced by $P(\pi)$. Suppose that there are m ergodic sets, say $E_1,E_2,\ldots,E_m$ and let F be the set of the transient states. We define the vectors $x(\pi)$ and $y(\pi)$ by

$$(4.3.2)\qquad \begin{cases} x_{ia}(\pi) := [\beta^T P^*(\pi)]_i \cdot \pi_{ia} & a \in A(i),\ i \in E \\ y_{ia}(\pi) := [\beta^T D(\pi) + \gamma^T P^*(\pi)]_i \cdot \pi_{ia} & a \in A(i),\ i \in E \end{cases}$$

where

$$(4.3.3)\qquad \gamma_i := \begin{cases} 0 & i \in F \\ \max_{\ell \in E_j} \{-\sum_k \beta_k d_{k\ell}(\pi) / \sum_{k \in E_j} p^*_{k\ell}(\pi)\} & i \in E_j,\ 1 \le j \le m. \end{cases}$$

Notice that γ is constant on every ergodic set.

<u>THEOREM 4.3.1</u>. $(x(\pi),y(\pi))$, *defined by* (4.3.2), *is a feasible solution of the linear programming problem* (4.2.11).

<u>PROOF</u>. In the proof we will use some properties of the matrices $P^*(\pi)$ and $D(\pi)$ as mentioned in theorem 2.4.1.

$$\sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia}(\pi) = \sum_a x_{ja}(\pi) - \sum_i \sum_a p_{iaj} x_{ia}(\pi)$$

$$= [\beta^T P^*(\pi)]_j - [\beta^T P^*(\pi) P(\pi)]_j = 0, \quad j \in E.$$

$$\sum_a x_{ja} + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia}(\pi)$$

$$= [\beta^T P^*(\pi)]_j + [\beta^T D(\pi) + \gamma^T P^*(\pi)]_j - [\beta^T D(\pi) P(\pi) + \gamma^T P^*(\pi) P(\pi)]_j$$

$$= [\beta^T \{P^*(\pi) + D(\pi)(I - P(\pi))\}]_j + [\gamma^T P^*(\pi)(I - P(\pi))]_j$$

$$= [\beta^T \{P^*(\pi) + I - P^*(\pi)\}]_j = \beta_j, \quad j \in E.$$

$$x_{ia}(\pi) \geq 0 \qquad a \in A(i),\ i \in E.$$

$$y_{ia}(\pi) = \{\sum_k \beta_k d_{ki}(\pi) + \sum_k \gamma_k p^*_{ki}(\pi)\}\cdot\pi_{ia} \qquad a \in A(i),\ i \in E.$$

If $i \in F$, then $p^*_{\cdot i}(\pi) = 0$ and $d_{\cdot i}(\pi) = \Sigma_{t=1}^{\infty} p^{t-1}_{\cdot i}(\pi) \geq 0$. Consequently,

$$y_{ia}(\pi) = \sum_k \beta_k \cdot \sum_{t=1}^{\infty} p^{t-1}_{ki}(\pi)\cdot\pi_{ia} \geq 0 \qquad a \in A(i),\ i \in F.$$

If $i \notin F$, say $i \in E_j$, then $\gamma_k p^*_{ki}(\pi) = 0$ for every $k \notin E_j$. Hence, we get

$$\begin{aligned} y_{ia}(\pi) &= \{\sum_k \beta_k d_{ki}(\pi) + \sum_{k\in E_j} \gamma_k p^*_{ki}(\pi)\}\cdot\pi_{ia} \\ &= \{\sum_k \beta_k d_{ki}(\pi) + \gamma_i \cdot \sum_{k\in E_j} p^*_{ki}(\pi)\}\cdot\pi_{ia} \\ &\geq \{\sum_k \beta_k d_{ki}(\pi) - (\sum_k \beta_k d_{ki}(\pi))\cdot(\sum_{k\in E_j} p^*_{ki}(\pi))^{-1}\cdot(\sum_{k\in E_j} p^*_{ki}(\pi))\}\cdot\pi_{ia} \\ &= 0 \qquad a \in A(i),\ i \notin F. \end{aligned}$$

This completes the proof of the theorem. □

For a stationary policy π^∞, let $(X(\pi),Y(\pi))$ be the class of corresponding equivalent feasible solutions. We choose the element $(x(\pi),y(\pi))$ as the *representative* of this equivalence class.

THEOREM 4.3.2. *The mapping defined by* (4.3.2) *is a one-to-one mapping of the stationary policies onto the set of representatives with* (4.3.1) *as the inverse mapping.*

PROOF. It is obvious that the stationary policies are mapped onto the set of representatives. Suppose that $\pi^1 \neq \pi^2$ and $(x(\pi^1),y(\pi^1)) = (x(\pi^2),y(\pi^2))$. Then, we obtain

$$\pi^1_{ia} = x_{ia}(\pi^1)/\sum_a x_{ia}(\pi^1) = x_{ia}(\pi^2)/\sum_a x_{ia}(\pi^2) = \pi^2_{ia} \qquad a \in A(i),\ i \in E_{x(\pi^1)} = E_{x(\pi^2)},$$

and

$$\pi^1_{ia} = y_{ia}(\pi^1)/\sum_a y_{ia}(\pi^1) = y_{ia}(\pi^2)/\sum_a y_{ia}(\pi^2) = \pi^2_{ia} \qquad a \in A(i),\ i \in E\backslash E_{x(\pi^1)} = E\backslash E_{x(\pi^2)}.$$

Hence, $\pi^1 = \pi^2$ implying a contradiction. □

REMARK 4.3.1. Suppose that (x,y) is a feasible solution of program (4.2.11). Then, if we define $x_j := \Sigma_a x_{ja}$, $j \in E$, we have

$$x_j = \sum_a x_{ja} = \sum_i \sum_a p_{iaj} x_{ia} = \sum_i \sum_a p_{iaj} \pi_{ia} x_i = \sum_i x_i p_{ij}(\pi), \qquad j \in E,$$

and

$$\begin{aligned}\sum_j x_j &= \sum_j \{\beta_j - \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia}\} \\ &= \sum_j \beta_j - \sum_i \sum_a \{\sum_j \delta_{ij} - \sum_j p_{iaj}\} y_{ia} = \sum_j \beta_j = 1.\end{aligned}$$

Hence, x is a stationary probability distribution of the Markov chain induced by $P(\pi(x,y))$.

Conversely, if x is a stationary probability distribution of the Markov chain induced by $P(\pi)$ for some stationary policy π^∞, then in general x cannot be completed by a y such that (x,y) is a feasible solution of the linear programming problem (4.2.11). For instance, in the AMD-model of example 4.3.1 $x := (1/3, 1/3, 1/3, 0)^T$ is a stationary probability distribution of the Markov chain induced by $P(f)$, where f satisfies $f(1) = f(2) = 1$, $f(3) = 2$ and $f(4) = 1$. There is no corresponding feasible solution since for any feasible solution $x_{41} \geq \frac{1}{4}$.

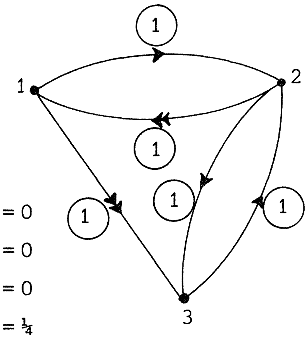
From example 4.3.1 it also follows that $X(\pi)$ can have more than one element. If the Markov chain induced by $P(\pi)$ is unichained, then it follows from theorem 2.3.3 that the stationary probability distribution is unique. Hence, $X(\pi)$ consists of one element: $X(\pi) = \{x(\pi)\}$. Similarly to theorem 4.3.1, it can be shown that any (x,y), where $x = x(\pi)$ and $y \in Y^\circ(\pi) := \{y \mid y_{ia} = y_{ia}(\pi) + [c^T P^*(\pi)]_i \cdot \pi_{ia}$ for some $c \geq 0\}$ is a feasible solution of program (4.2.11). Hence $Y^\circ(\pi) \subset Y(\pi)$. The next example shows that it may occur that $Y^\circ(\pi) \neq Y(\pi)$.

EXAMPLE 4.3.2. Consider the model of figure 4.3.2. The linear programming problem is:

maximize $x_{11} + x_{12} + x_{21} + x_{22} + x_{31}$

subject to

$$\begin{array}{rcl}
x_{11} + x_{12} - x_{22} & = & 0 \\
-x_{11} + x_{21} + x_{22} - x_{31} & = & 0 \\
-x_{12} - x_{21} + x_{31} & = & 0 \\
x_{11} + x_{12} + y_{11} + y_{12} - y_{22} & = & \frac{1}{4} \\
x_{21} + x_{22} - y_{11} + y_{21} + y_{22} - y_{31} & = & \frac{1}{2} \\
x_{31} - y_{12} - y_{21} + y_{31} & = & \frac{1}{4}
\end{array}$$



$\beta_1 = \frac{1}{4}$, $\beta_2 = \frac{1}{2}$, $\beta_3 = \frac{1}{4}$

Figure 4.3.2

$x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, y_{11}, y_{12}, y_{21}, y_{22}, y_{31} \geq 0$

Take the pure and stationary policy f^∞ such that $f(i) = 1$, $i \in E$. Then,

$x_{11}(f) = 0$, $x_{12}(f) = 0$, $x_{21}(f) = \frac{1}{2}$, $x_{22}(f) = 0$, $x_{31}(f) = \frac{1}{2}$;

$y_{11}(f) = \frac{1}{4}$, $y_{12}(f) = 0$, $y_{21}(f) = \frac{1}{4}$, $y_{22}(f) = 0$, $y_{31}(f) = 0$.

The feasible solution (x,y), where $x = x(f)$ and $y_{11} = \frac{1}{2}$, $y_{12} = 0$, $y_{21} = \frac{1}{2}$, $y_{22} = \frac{1}{4}$, $y_{31} = \frac{1}{4}$ is an element of $Y(f)$. Suppose that $y \in Y^\circ(f)$. Since state 1 is transient under $P(f)$, each $\tilde{y} \in Y^\circ(f)$ satisfies $\tilde{y}_{11} = y_{11}(f) = \frac{1}{4}$. Hence, $Y^\circ(f) \neq Y(f)$.

THEOREM 4.3.3. *The correspondence between the stationary policies and the feasible solutions of the linear program* (4.2.11) *preserves the optimality property, i.e.*

1. *If* π^∞ *is a stationary average optimal policy, then* $(x(\pi),y(\pi))$ *is an optimal solution of the linear program* (4.2.11).
2. *If* (x,y) *is an optimal solution of the linear program* (4.2.11), *then the stationary policy* $\pi^\infty(x,y)$ *is an average optimal policy.*

PROOF.

1. Let (ϕ,u) be an optimal solution of the linear programming problem (4.2.10). Since $(x(\pi),y(\pi))$ is a feasible solution of program (4.2.11), it follows from the theory of linear programming (cf. theorem 1.3.4) that it is sufficient to prove that $\Sigma_i\Sigma_a r_{ia}x_{ia}(\pi) = \Sigma_j\beta_j\phi_j$. We have

$$\sum_i\sum_a r_{ia}x_{ia}(\pi) = \sum_i\sum_a r_{ia}[\beta^T P^*(\pi)]_i \cdot \pi_{ia}$$

$$= \beta^T P^*(\pi)r(\pi) = \beta^T\phi(\pi^\infty) = \beta^T\phi,$$

   which completes the proof of the first part of the theorem.
2. The proof has the same structure as the proof of theorem 4.2.4. We first present three propositions which are similar to the propositions 4.2.1, 4.2.2 and 4.2.3. Then we complete the proof. Throughout the proof (ϕ,u) is an optimal solution of program (4.2.10).

PROPOSITION 4.3.1.

$$\sum_j(\delta_{ij}-p_{iaj})\phi_j = 0 \qquad a \in A^\circ(i),\ i \in E,$$

$$\phi_i + \sum_j(\delta_{ij}-p_{iaj})u_j = r_{ia} \qquad a \in A^\circ(i),\ i \in E_x,$$

where

$$A^{\circ}(i) := \{a \in A(i) \mid \pi_{ia}(x,y) > 0\}, \qquad i \in E.$$

PROOF. Since $x_{ia} > 0$, $a \in A^{\circ}(i)$, $i \in E_x$ and $y_{ia} > 0$, $a \in A^{\circ}(i)$, $i \in E\backslash E_x$, it follows from the complementary slackness property of linear programming (see corollary 1.3.1) that

$$\sum_j (\delta_{ij} - p_{iaj})\phi_j = 0 \qquad a \in A^{\circ}(i),\ i \in E\backslash E_x$$

and

$$\phi_i + \sum_j (\delta_{ij} - p_{iaj})u_j = r_{ia} \qquad a \in A^{\circ}(i),\ i \in E_x.$$

Suppose that $\Sigma_j (\delta_{kj} - p_{ka_kj})\phi_j \neq 0$ for some $a_k \in A^{\circ}(k)$ and $k \in E_x$. Since $\pi_{ka_k}(x,y) > 0$, we also have $x_{ka_k} > 0$, and consequently

$$\sum_j (\delta_{kj} - p_{ka_kj})\phi_j \cdot x_{ka_k} > 0.$$

Moreover, we have

$$\sum_j (\delta_{ij} - p_{iaj})\phi_j \cdot x_{ia} \geq 0 \qquad a \in A(i),\ i \in E.$$

Hence, we obtain

$$\sum_i \sum_a \sum_j (\delta_{ij} - p_{iaj})\phi_j \cdot x_{ia} > 0.$$

On the other hand, the constraints of (4.2.11) imply that

$$\sum_i \sum_a \sum_j (\delta_{ij} - p_{iaj})\phi_j \cdot x_{ia} = \sum_j \{\sum_i \sum_a (\delta_{ij} - p_{iaj})x_{ia}\}\phi_j = 0.$$

This contradiction completes the proof.

PROPOSITION 4.3.2. E_x *is closed under* $P(\pi(x,y))$.

PROOF. Suppose that $p_{k\ell}(\pi(x,y)) > 0$ for some $k \in E_x$ and $\ell \in E\backslash E_x$. Since $p_{k\ell}(\pi(x,y)) = \Sigma_a p_{ka\ell}\pi_{ka}(x,y)$, there exists an action a_k such that $p_{ka_k\ell} > 0$ and $\pi_{ka_k}(x,y) > 0$. From the constraints of program (4.2.11) it follows that

$$0 = \sum_a x_{\ell a} = \sum_i \sum_a p_{ia\ell} x_{ia} \geq p_{ka_k\ell} x_{ka_k} > 0,$$

implying a contradiction.

PROPOSITION 4.3.3. *For any feasible solution* (x,y) *of the linear program* (4.2.11), E_x *is the set of recurrent states in the Markov chain induced by* $P(\pi(x,y))$.

PROOF. Let $x_i := \Sigma_a x_{ia}$ and $y_i := \Sigma_a y_{ia}$, $i \in E$. We have seen in remark 4.3.1 that x is a stationary probability distribution in the Markov chain induced by $P(\pi(x,y))$. Theorem 2.3.3 implies that $F \subset E \backslash E_x$, where F is the set of transient states in this Markov chain. Suppose that $F \neq E\backslash E_x$. Since E_x is closed under $P(\pi(x,y))$, there is an ergodic set $E_1 \subset E\backslash E_x$. Hence, we can write

$$0 = \sum_{j \notin E_1} \sum_{i \in E_1} p_{ij}(\pi(x,y)).$$

Then, we also have

$$0 = \sum_{j \notin E_1} \sum_{i \in E_1} \sum_a p_{iaj} y_{ia}$$

and

$$\begin{aligned} 0 < \sum_{j \in E_1} \beta_j &= \sum_{j \in E_1} y_j - \sum_{j \in E_1} \sum_i \sum_a p_{iaj} y_{ia} = \\ &\sum_{j \in E_1} y_j - \sum_j \sum_{i \in E_1} \sum_a p_{iaj} y_{ia} + \sum_{j \notin E_1} \sum_{i \in E_1} \sum_a p_{iaj} y_{ia} - \sum_{j \in E_1} \sum_{i \notin E_1} \sum_a p_{iaj} y_{ia} = \\ &\sum_{j \in E_1} y_j - \sum_{i \in E_1} \sum_a y_{ia} - \sum_{j \in E_1} \sum_{i \notin E_1} \sum_a p_{iaj} y_{ia} = \\ &- \sum_{j \in E_1} \sum_{i \notin E_1} \sum_a p_{iaj} y_{ia} \le 0 \end{aligned}$$

implying a contradiction. This yields the proof.

We complete the proof as follows. From proposition 4.3.1 it follows that $P^*(\pi(x,y))\phi = \phi$. Since $E\backslash E_x$ is the set of transient states, we have $p^*_{\cdot i}(\pi(x,y)) = 0$, $i \in E\backslash E_x$. Then, using proposition 4.3.1 we can write

$$\begin{aligned} \phi(\pi^\infty(x,y)) &= P^*(\pi(x,y))r(\pi(x,y)) \\ &= P^*(\pi(x,y))\{\phi + (I - P(\pi(x,y)))u\} \\ &= P^*(\pi(x,y))\phi \\ &= \phi. \end{aligned}$$

Hence, $\pi^\infty(x,y)$ is an average optimal policy. □

REMARK 4.3.2. Proposition 4.3.3 differs from proposition 4.2.3 by the fact that in theorem 4.2.4 the states of E_x may contain transient states. Consider for instance example 4.2.1. The policy f_1^∞ is average optimal, $E_x = E$, but state 2 is transient in the Markov chain induced by $P(f_1)$.

REMARK 4.3.3. If π^∞ is an optimal stationary policy and if (x,y) is a feasible solution of program (4.2.11) such that $\pi^\infty(x,y) = \pi^\infty$, then in general (x,y) is not an optimal solution of (4.2.11). Below we give an example.

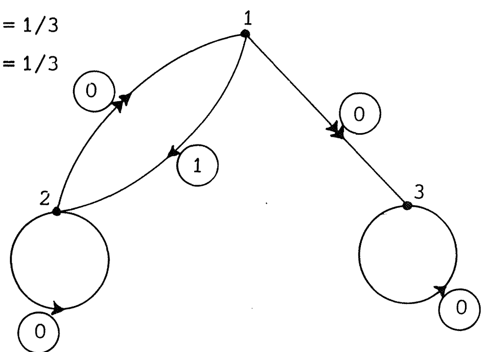
EXAMPLE 4.3.3. Consider the model of figure 4.3.3. The corresponding linear programming problem is:

maximize x_{11}

subject to

$$\begin{array}{l}
x_{11} + x_{12} - x_{22} = 0 \\
-x_{11} + x_{22} = 0 \\
-x_{12} = 0 \\
x_{11} + x_{12} + y_{11} + y_{12} - y_{22} = 1/3 \\
x_{21} + x_{22} - y_{11} + y_{22} = 1/3 \\
x_{31} - y_{12} = 1/3 \\
x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, y_{11}, y_{12}, y_{22} \geq 0
\end{array}$$

The pure and stationary policy f^∞ such that $f(1) = 1$, $f(2) = 2$, $f(3) = 1$ is average optimal. The vector (x,y), where $x_{11} = 1/6$, $x_{12} = x_{21} = 0$, $x_{22} = 1/6$, $x_{31} = 2/3$, $y_{11} = 0$, $y_{12} = 1/3$, $y_{22} = 1/6$, is a feasible solution and $\pi^\infty(x,y) = f^\infty$. However, (x,y) is not an optimal solution of the linear programming problem.



$\beta_1 = \beta_2 = \beta_3 = 1/3$

Figure 4.3.3

THEOREM 4.3.4. *Let* f^∞ *be any pure and stationary policy. Then the corresponding vector* $(x(f),y(f))$, *defined by* (4.3.2), *is an extreme feasible solution of the linear programming problem* (4.2.11).

PROOF. Suppose that $(x(f),y(f))$ is not an extreme point of the set of feasible solutions of program (4.2.11). Then there exist different feasible solutions (x^1,y^1) and (x^2,y^2) such that for some $\lambda \in (0,1)$

$$\begin{cases} x(f) = \lambda x^1 + (1-\lambda)x^2 \\ y(f) = \lambda y^1 + (1-\lambda)y^2 \end{cases}$$

Since

$$x_{ia}(f) = y_{ia}(f) = 0 \qquad a \neq f(i),\ i \in E,$$

we have

$$x^1_{ia} = x^2_{ia} = y^1_{ia} = y^2_{ia} = 0 \qquad a \neq f(i),\ i \in E.$$

Let

$$P := P(f), \qquad \tilde{x} := (x_{if(i)}(f)), \qquad \tilde{y} := (y_{if(i)}(f)),$$

$$\tilde{x}^1 := (x^1_{if(i)}), \qquad \tilde{x}^2 := (x^2_{if(i)}), \qquad \tilde{y}^1 := (y^1_{if(i)})$$

and

$$\tilde{y}^2 := (y^2_{if(i)}).$$

Then $(\tilde{x},\tilde{y})$, $(\tilde{x}^1,\tilde{y}^1)$ and $(\tilde{x}^2,\tilde{y}^2)$ are solutions of the linear system

$$\text{(4.3.4)} \qquad \begin{cases} x^T(I-P) & = 0 \\ x^T \qquad + y^T(I-P) & = \beta^T. \end{cases}$$

Hence, for any solution (x,y) of (4.3.4) we obtain $x^T = x^TP$, and consequently $x^T = x^TP^* = \beta^TP^* - y^T(I-P)P^* = \beta^TP^*$, implying that

$$\tilde{x} = \tilde{x}^1 = \tilde{x}^2 = \beta^TP^*.$$

We also get

$$y^T(I-P+P^*) = \beta^T - x^T + y^TP^* = \beta^T(I-P^*) + y^TP^*.$$

From theorem 2.4.1 it follows that

$$\text{(4.3.5)} \qquad y^T = \beta^T(I-P^*)(D+P^*) + y^TP^*(D+P^*) = \beta^TD + y^TP^*.$$

Consider the Markov chain induced by the transition matrix P. Suppose that there are m ergodic sets, say $E_1, E_2, \dots, E_m$, and let F be the set of transient states. Then, (4.3.5) implies that any solution (x,y) of (4.3.4) satisfies $y_i = (\beta^TD)_i$, $i \in F$. Consequently,

$$\tilde{y}_i = \tilde{y}^1_i = \tilde{y}^2_i, \qquad i \in F.$$

By the definition of γ given in (4.3.3), there is in each ergodic set E_k a state, say i_k, such that $\tilde{y}_{i_k} = 0$. Then also $\tilde{y}^1_{i_k} = \tilde{y}^2_{i_k} = 0$. Since $(\tilde{x}^1,\tilde{y}^1)$ and $(\tilde{x}^2,\tilde{y}^2)$ are solutions of the linear system (4.3.4) which satisfy $\tilde{x}^1 = \tilde{x}^2$, $\tilde{y}^1_i - \tilde{y}^2_i = 0$, $i \in F$, $\tilde{y}^1_{i_k} - \tilde{y}^2_{i_k} = 0$, $k = 1,2,\dots,m$, we obtain from (4.3.4)

$$(4.3.6) \qquad \begin{cases} \tilde{y}^1_i - \tilde{y}^2_i = \sum_{\ell\in E_k} (\tilde{y}^1_\ell - \tilde{y}^2_\ell) p_{\ell i} & i \in E_k \\ \tilde{y}^1_{i_k} - \tilde{y}^2_{i_k} = 0, & \end{cases} \qquad k = 1,2,\dots,m$$

Let $z_i := \tilde{y}^1_i - \tilde{y}^2_i$, $i \in E_k$ and $q_{ij} := p_{ij}$, $i.j \in E_k$. Then, we have

$$z^T = z^T Q = z^T Q^*.$$

Since E_k is an ergodic set, theorem 2.3.2 implies that $q^*_{i\cdot} = q^*_{j\cdot} >> 0$ for all $i,j \in E_k$. Hence, we get

$$\begin{cases} z_i = q^*_{ii} \cdot \sum_{j\in E_k} z_j & i \in E_k \\ z_{i_k} = 0. & \end{cases}$$

Then,

$$\sum_{j\in E_k} z_j = 0 \quad \text{and consequently,} \quad z_i = 0 \quad i \in E_k.$$

Therefore, we have shown that $\tilde{y}^1 = \tilde{y}^2$, which completes the proof that $(x(f),y(f))$ is an extreme point. □

<u>REMARK 4.3.4</u>. In example 4.2.1 we have found an extreme point (x^*,y^*) of the set of feasible solutions such that the corresponding policy is not pure. Hence, the reverse statement of theorem 4.3.4 is in general not true.

<u>REMARK 4.3.5</u>. Take any stationary policy π^∞ and let $R(\pi)$ be the set of recurrent states in the Markov chain induced by $P(\pi)$. Then proposition 4.3.3 implies that for every feasible solution (x,y) of (4.2.11) such that $(x,y) \in (X(\pi),Y(\pi))$ $E_x = R(\pi)$. Consequently, elements in the same equivalence class have the same positive x-components.

4.4. POLICY IMPROVEMENT AND LINEAR PROGRAMMING

In this section we shall discuss some relations between the policy improvement method and the linear programming approach. The idea of policy improvement was introduced by HOWARD [1960]. BLACKWELL [1962] has given an elegant mathematical foundation of the policy improvement method, treating the average reward case as a limiting case of the α-discounted reward case. By Blackwell's algorithm a pure and stationary average optimal policy is obtained. VEINOTT [1966] and DENARDO [1970a] have generalized this algorithm to an algorithm by which a pure and stationary bias optimal policy can be determined. MILLER & VEINOTT [1969] have extended these results. They present a Laurent expansion in $(1-\alpha)$ for $v^{\alpha}(f^{\infty})$ by which algorithms can be constructed in order to find optimal policies with regard to more selective criteria. In particular, a finite algorithm was proposed to obtain a Blackwell optimal policy. Other references on this subject are DENARDO & MILLER [1968], VEINOTT [1969], DENARDO [1973], VEINOTT [1974] and HORDIJK [1976].

THEOREM 4.4.1. *For any pure and stationary policy* f^{∞}, *the linear system*

$$(4.4.1)\qquad \begin{cases} (I-P(f))\tilde{\phi} & & = 0 \\ \tilde{\phi} + (I-P(f))\tilde{u} & & = r(f) \\ & \tilde{u} + (I-P(f))\tilde{z} & = 0 \end{cases}$$

has a feasible solution $(\tilde{\phi},\tilde{u},\tilde{z})$. *Moreover any feasible solution* $(\tilde{\phi},\tilde{u},\tilde{z})$ *of* (4.4.1) *satisfies* $\tilde{\phi} = \phi(f^{\infty})$ *and* $\tilde{u} = u(f^{\infty})$.

PROOF. (cf. HORDIJK [1976]). In the proof we use repeatedly the results of theorem 2.4.1. Let $\tilde{\phi} := \phi(f^{\infty})$, $\tilde{u} := u(f^{\infty})$ and $\tilde{z} := -D(f)u(f^{\infty})$. Then, we obtain

$$\begin{aligned}
(I-P(f))\tilde{\phi} &= (I-P(f))P^{*}(f)r(f) \\
&= 0. \\
\tilde{\phi} + (I-P(f))\tilde{u} &= \{P^{*}(f) + (I-P(f))D(f)\}r(f) \\
&= \{P^{*}(f) + I-P^{*}(f)\}r(f) \\
&= r(f). \\
\tilde{u} + (I-P(f))\tilde{z} &= D(f)\{I-(I-P(f))D(f)\}r(f) \\
&= D(f)P^{*}(f)r(f) \\
&= 0.
\end{aligned}$$

Suppose that $(\tilde{\phi},\tilde{u},\tilde{z})$ is a feasible solution of (4.4.1). Then we have

$$\tilde{\phi} = P(f)\tilde{\phi} = P^*(f)\tilde{\phi}$$

$$= P^*(f)\{r(f) - (I-P(f))\tilde{u}\} = P^*(f)r(f)$$

$$= \phi(f^\infty).$$

$$\tilde{u} = (I-P(f) + P^*(f))^{-1}(I-P(f) + P^*(f))\tilde{u}$$

$$= (D(f)+P^*(f))(I-P(f)+P^*(f))\tilde{u}$$

$$= (D(f)+P^*(f))(r(f)-\tilde{\phi})$$

$$= D(f)r(f) = u(f^\infty). \quad \square$$

We define for any pure and stationary policy f^∞ the sets $A(i,f)$, $i \in E$, by

$$(4.4.2) \qquad A(i,f) = \left\{ a \in A(i) \left| \begin{array}{l} \sum_j p_{iaj}\phi_j(f^\infty) > \phi_i(f^\infty) \quad \text{or} \quad \sum_j p_{iaj}\phi_j(f^\infty) = \\ \phi_i(f^\infty) \ \& \ r_{ia} + \sum_j p_{iaj}u_j(f^\infty) > \phi_i(f^\infty)+u_i(f^\infty) \end{array} \right. \right\}.$$

THEOREM 4.4.2. *Let f^∞ be a pure and stationary policy.*

1. *If $A(i,f) = \emptyset$ for all $i \in E$, then f^∞ is an average optimal policy.*
2. *If $A(i,f) \neq \emptyset$ for some $i \in E$, and g^∞ is a pure and stationary policy such that $g(i) \in A(i,f)$ for at least one $i \in E$ and $g(i) = f(i)$ whenever $g(i) \notin A(i,f)$, then $\phi(g^\infty) \geq \phi(f^\infty)$ and $v^\alpha(g^\infty) > v^\alpha(f^\infty)$ for all α sufficiently near to 1.*

PROOF. (cf. BLACKWELL [1962]).
1. Let g^∞ be an arbitrarily chosen pure and stationary policy. Since $A(i,f) = \emptyset$ for all $i \in E$, we have

$$(4.4.3) \qquad P(g)\phi(f^\infty) \leq \phi(f^\infty) \quad \text{and} \quad r_i(g) + (P(g)u(f^\infty))_i \leq \phi_i(f^\infty)+u_i(f^\infty)$$

for each i which satisfies $(P(g)\phi(f^\infty))_i = \phi_i(f^\infty)$.
Let $R := (g,f,f,\ldots)$. Then $v^\alpha(R) = r(g) + \alpha P(g)v^\alpha(f^\infty)$ and it follows from (2.5.7) that we can write

$$v^\alpha(R) = r(g) + \{1-(1-\alpha)\}P(g)\{(1-\alpha)^{-1}\cdot\phi(f^\infty) + u(f^\infty) + \varepsilon^1(\alpha)\}$$

$$= (1-\alpha)^{-1}\cdot P(g)\phi(f^\infty) + r(g) + P(g)u(f^\infty) - P(g)\phi(f^\infty) + \varepsilon^2(\alpha),$$

where $\lim_{\alpha\uparrow 1} \varepsilon^k(\alpha) = 0$ for $k = 1,2$. Hence, (4.4.4) $$v^{\alpha}(f^{\infty})-v^{\alpha}(R) = (1-\alpha)^{-1}\cdot\{\phi(f^{\infty})-P(g)\phi(f^{\infty})\} + u(f^{\infty}) + P(g)\phi(f^{\infty}) - r(g) - P(g)u(f^{\infty}) + \varepsilon^{3}(\alpha),$$

where $\lim_{\alpha\uparrow 1} \varepsilon^{3}(\alpha) = 0$.
Therefore, it follows from (4.4.3) and (4.4.4) that for α sufficiently near to 1

(4.4.5) $$v^{\alpha}(f^{\infty}) - v^{\alpha}(R) \geq \varepsilon^{3}(\alpha) \quad \text{and} \quad \lim_{\alpha\uparrow 1} \varepsilon^{3}(\alpha) = 0.$$

Let $\varepsilon(\alpha) := \mathit{min}_i \varepsilon_i^{3}(\alpha)$. Then,

(4.4.6) $$v^{\alpha}(f^{\infty}) \geq v^{\alpha}(R) + \varepsilon(\alpha)\cdot e = r(g) + \varepsilon(\alpha)\cdot e + \alpha P(g) v^{\alpha}(f^{\infty}).$$

By iterating (4.4.6), we obtain

(4.4.7) $$v^{\alpha}(f^{\infty}) \geq \sum_{t=1}^{\infty} \alpha^{t-1} P^{t-1}(g)(r(g) + \varepsilon(\alpha)\cdot e) = v^{\alpha}(g^{\infty}) + \frac{\varepsilon(\alpha)}{1-\alpha}\cdot e$$

From (2.5.7) and (4.4.7) it follows that

$$\frac{\phi(f^{\infty})-\phi(g^{\infty})-\varepsilon(\alpha)\cdot e}{1-\alpha} + u(f^{\infty})-u(g^{\infty})+\varepsilon^{4}(\alpha) \geq 0 \qquad \alpha \in [0,1),$$

where $\lim_{\alpha\uparrow 1} \varepsilon^{4}(\alpha) = 0$ and $\lim_{\alpha\uparrow 1} \varepsilon(\alpha) = 0$.
Consequently,

(4.4.8) $$\phi(f^{\infty}) \geq \phi(g^{\infty}).$$

Since g^{∞} has been chosen arbitrarily and since there exists an average optimal policy in the class of pure and stationary policies, (4.4.8) implies that f^{∞} is an average optimal policy.
2. Let g^{∞} be such that $g(i) \in A(i,f)$ for at least one $i \in E$ and $g(i) = f(i)$ if $g(i) \notin A(i,f)$. Define the policy R by $R := (g,f,f,\ldots)$. Notice that (4.4.4) is also valid in this case. Then, it follows that

(a) if $g(i) = f(i)$, then $v_i^{\alpha}(R) = v_i^{\alpha}(f^{\infty})$.

(b) if $g(i) \neq f(i)$, then $v_i^{\alpha}(R) > v_i^{\alpha}(f^{\infty})$ for α sufficiently near to 1.

Hence,

(4.4.9) $$v^{\alpha}(f^{\infty}) < v^{\alpha}(R) = r(g) + \alpha P(g) v^{\alpha}(f^{\infty}) \qquad \alpha \in [\alpha_o,1), \text{ where } \alpha_o \in [0,1).$$

By iterating (4.4.9) we obtain

$$(4.4.10) \qquad v^\alpha(f^\infty) < \sum_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(g)r(g) = v^\alpha(g^\infty) \qquad \alpha \in [\alpha_o,1).$$

Since

$$0 < v^\alpha(g^\infty) - v^\alpha(f^\infty) = \frac{\phi(g^\infty)-\phi(f^\infty)}{1-\alpha} + u(g^\infty) - u(f^\infty) + \varepsilon^5(\alpha), \quad \alpha \in [\alpha_o,1),$$

where $\lim_{\alpha\uparrow 1} \varepsilon^5(\alpha) = 0$, we get

$$(4.4.11) \qquad \phi(g^\infty) \geq \phi(f^\infty).$$

Combining (4.4.10) and (4.4.11) completes the proof. □

Next, we formulate and prove the correctness of the following policy improvement algorithm.

ALGORITHM XV for the construction of a pure and stationary average optimal policy by the policy improvement method (multichain case). case).

<u>step 1</u>: Take an arbitrary $f^\infty \in C_D$.
<u>step 2</u>: Compute $\phi(f^\infty)$ and $u(f^\infty)$ by solving the linear system

$$\begin{cases} (I-P(f))\tilde{\phi} & = 0 \\ \tilde{\phi} + (I-P(f))\tilde{u} & = r(f) \\ \tilde{u} + (I-P(f))\tilde{z} & = 0 \end{cases}$$

<u>step 3</u>: Determine for every $i \in E$

$$A(i,f) := \left\{ a \in A(i) \,\middle|\, \begin{array}{l} \sum_j p_{iaj}\phi_j(f^\infty) > \phi_i(f^\infty) \text{ or } \sum_j p_{iaj}\phi_j(f^\infty) = \\ \phi_i(f^\infty) \;\&\; r_{ia} + \sum_j p_{iaj}u_j(f^\infty) > \phi_i(f^\infty) + u_i(f^\infty) \end{array} \right\}$$

<u>step 4</u>: If $A(i,f) = \emptyset$ for all $i \in E$, then f^∞ is an average optimal policy (STOP).
Otherwise, go to step 5.
<u>step 5</u>: Take g^∞ such that

$$\begin{cases} g(i) \in A(i,f) & \text{if } A(i,f) \neq \emptyset \\ g(i) = f(i) & \text{if } A(i,f) = \emptyset \end{cases}, \quad i \in E.$$

step 6: $f^\infty := g^\infty$ and go to step 2.

THEOREM 4.4.3. *The policy improvement algorithm XV provides an average optimal policy within a finite number of iterations.*

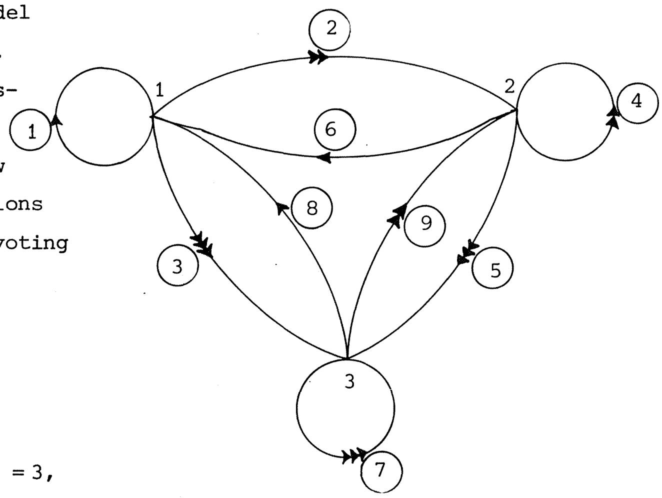
PROOF. If in the algorithm the policy g^∞ is taken as successor of f^∞, then it follows from theorem 4.4.2 that $v^\alpha(g^\infty) > v^\alpha(f^\infty)$ for α near enough to 1. Therefore, each pure and stationary policy can occur only once. Since there are a finite number of pure and stationary policies, the policy improvement algorithm terminates after a finite number of iterations with a policy $f^\infty \in C_D$ which satisfies $A(i,f) = \emptyset$ for all $i \in E$. This policy f^∞ is by theorem 4.4.2 an average optimal policy. □

Let f_k^∞ be the pure and stationary policy obtained in the k-th step of algorithm XV. In theorem 4.3.4 we have shown that $(x(f_k),y(f_k))$, defined by (4.3.2), is an extreme point of the set of feasible solutions of the linear program (4.2.11). The value of the objective function satisfies

$$\sum_i \sum_a r_{ia} x_{ia}(f_k) = \sum_i r_i(f_k)(\beta^T P^*(f_k))_i = \beta^T P^*(f_k) r(f_k) = \beta^T \phi(f_k^\infty).$$

The successive policies f_k^∞, $k = 1,2,\dots$, correspond to extreme points of the set of feasible solutions of program (4.2.11). From theorem 4.4.2 we know that the values of the objective function are nondecreasing and it follows also from theorem 4.4.2 that cycling cannot occur. The successive extreme points $(x(f_k),y(f_k))$, $k = 1,2,\dots$, are not necessarily adjacent. Hence, the policy iteration algorithm is not equivalent to the standard simplex algorithm but rather to another linear programming algorithm in which pivot operations on many variables are performed simultaneously. Such an algorithm is called a *block-pivoting* algorithm and may be viewed as a special case of the general class of methods of feasible directions as introduced by ZOUTENDIJK [1960].

CONCLUSION: *The policy improvement algorithm is equivalent to a block-pivoting simplex algorithm.*



$\beta_1 = \beta_2 = \beta_3 = 1/3$

Figure 4.4.1

EXAMPLE 4.4.1. For the model given in figure 4.4.1 (cf. HOWARD [1960] p.65) we display the policy improvement algorithm and we show how the successive iterations can be viewed as block-pivoting in the simplex algorithm.

Policy improvement

Iteration 1:

1. Take f_1^∞ such that $f_1(1) = 3$, $f_1(2) = 2$, $f_1(3) = 1$.
2. $\phi(f_1^\infty) = (11/2,4,11/2)^T$; $u(f_1^\infty) = (-5/4,0,5/4)^T$.
3. $A(1,f_1) = \emptyset$; $A(2,f_1) = \{1,3\}$; $A(3,f_1) = \{3\}$.
5. Take g^∞ such that $g(1) = 3$, $g(2) = 1$, $g(3) = 3$.
6. $f_2(1) = 3$, $f_2(2) = 1$, $f_2(3) = 3$.

Iteration 2:

2. $\phi(f_2^\infty) = (7,7,7)^T$; $u(f_2^\infty) = (-4,-5,0)^T$.
3. $A(1,f_2) = \emptyset$; $A(2,f_2) = \{3\}$; $A(3,f_2) = \emptyset$
5. Take g^∞ such that $g(1) = 3$, $g(2) = 3$, $g(3) = 3$.
6. $f_3(1) = 3$, $f_3(2) = 3$, $f_3(3) = 3$.

Iteration 3:

2. $\phi(f_3^\infty) = (7,7,7)^T$; $u(f_3^\infty) = (-4,-2,0)^T$.
3. $A(1,f_3) = \emptyset$; $A(2,f_3) = \emptyset$; $A(3,f_3) = \emptyset$.
4. f_3^∞ is an average optimal policy.

Linear programming

Iteration 1:

Policy f_1^∞ chooses in the three states the actions 3,2 and 1 respectively. Since the three states are recurrent in the Markov chain under $P(f_1)$, the variables x_{13}, x_{22} and x_{31} are basic-variables. The corresponding simplex tableau is as follows (the z-variables are artificial variables; the variables y_{11}, y_{22} and y_{33} can be omitted since the corresponding coefficients are all zeros).

		x_{11}	x_{12}	x_{21}	x_{23}	x_{32}	x_{33}	y_{12}	y_{13}	y_{21}	y_{23}	y_{31}	y_{32}
x_{13}	1/3	1	1					1	(1)	-1		-1	
z_2	0		-1	1	1	-1							
z_3	0		1	-1	-1	1							
x_{31}	1/3	1		1				1	1	-1		-1	
x_{22}	1/3			1	1			-1		1	(1)		-1
z_6	0	-1		-1		1	(1)	-1	-2	1	-1	2	1
z_0	5	10	1	6	-1	-9	-7	7	11	-7	4	-11	-4

Iteration 2:

Since the Markov chain under $P(f_2)$ has only state 3 as recurrent state (with $f_2(3) = 3$) and since $f_2(1) = 3$ and $f_2(2) = 1$, we let enter the variables y_{13}, y_{21} and x_{33} into the basis and we require that x_{13}, x_{22} and x_{31} become non-basic or basic with value 0. Then, after 3 standard pivot iterations, we obtain the tableau corresponding to f_2^∞:

		x_{11}	x_{12}	x_{21}	x_{23}	x_{32}	y_{12}	x_{13}	x_{22}	y_{23}	y_{31}	y_{32}
y_{13}	2/3	1	1	1	1			1	1	1	-1	-1
z_2	0		-1	1	1	-1						
z_3	0		1	-1	-1	1						
x_{31}	0		-1	1				-1				
y_{21}	1/3			1	1		-1		1	(1)		-1
x_{33}	1	1	2		1	1		2	1			
z_0	7	6	4	2	2	-2	0	3	3	0	0	0

Iteration 3:

The average optimal policy f_3^∞ is obtained by changing the variables y_{21} and y_{23} (this choice follows again from the analysis of the Markov chain induced by $P(f_3)$). The corresponding tableau becomes:

		x_{11}	x_{12}	x_{21}	x_{23}	x_{32}	y_{12}	x_{13}	x_{22}	y_{21}	y_{31}	y_{32}
y_{13}	1/3	1	1				1	1		-1	-1	
z_2	0		-1	1	1	-1						
z_3	0		1	-1	-1	1						
x_{31}	0		-1	1				-1				
y_{23}	1/3			1	1		-1		1	1		-1
x_{33}	1	1	2		1	1		2	1			
z_0	7	6	4	2	2	-2	0	3	3	0	0	0

REMARK 4.4.1. The final tableau is in the usual context of the simplex method not an optimal tableau. In an optimal tableau the row of the dual variables (i.e. the row at the bottom) has to be nonnegative. We can obtain such an optimal simplex tableau by changing the variables z_3 and x_{32}. Then the corresponding policy is again f_3^∞.

4.5. THE WEAK UNICHAIN CASE

Throughout this section we have the following assumption.

ASSUMPTION 4.5.1. (i) The AMD value-vector ϕ has identical components.
(ii) For any pure and stationary average optimal policy f^∞ and for an arbitrary ergodic set $E_1(f)$ in the Markov chain induced by $P(f)$, there exists a policy $g^\infty \in C_D$ such that g^∞ is also average optimal and $E_1(f)$ are the recurrent states in the Markov chain induced by $P(g)$.

If assumption 4.5.1 is satisfied, then the model is called *weakly unichained.* The weak unichain case includes the completely ergodic case, the unichain case (cf. section 4.6) but also the *communicating case* (i.e. for each pair $i,j \in E$ there exists a policy $f_o^\infty \in C_D$ and an integer $t \in \mathbb{N}$ such that $\mathbb{P}_{f_o^\infty}(X_t = j | X_1 = i) > 0$). The term communicating comes from BATHER [1973]; this concept is also used in HORDIJK [1974], chapter 8.

Let f^∞ be an average optimal policy and g^∞ the policy mentioned in asasumption 4.5.1. Then, this assumption implies that the policy f_1^∞, where

$$f_1(i) := \begin{cases} f(i) & i \in E_1(f) \\ \\ g(i) & i \in E \backslash E_1(f) \end{cases}$$

is also average optimal. Furthermore, it is obvious that the Markov chain induced by $P(f_1)$ is unichained. Since ϕ_j is independent of the initial state j, we may use instead of the AMD-value-vector ϕ a $\phi_o \in \mathbb{R}^1$ such that $\phi = \phi_o \cdot e$. From the results of section 4.2, it follows that ϕ_o is the optimal solution of the linear program

$$(4.5.1) \qquad \min\{\tilde{\phi}_o | \tilde{\phi}_o + \tilde{u}_i \geq r_{ia} + \textstyle\sum_j p_{iaj}\tilde{u}_j \qquad a \in A(i),\ i \in E\}.$$

The corresponding dual linear programming problem is

$$(4.5.2) \qquad \max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j \in E \\ \sum_i\sum_a x_{ia} = 1 & \\ x_{ia} \geq 0,\ a \in A(i), & i \in E \end{array}\right\}.$$

Below, we present an algorithm for the determination of an optimal policy and we prove its correctness.

ALGORITHM XVI for the construction of a pure and stationary average optimal policy (weak unichain case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$\max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j \in E \\ \sum_i\sum_a x_{ia} = 1 & \\ x_{ia} \geq 0,\ a \in A(i), & i \in E \end{array}\right\}.$$

step 2: Take $f_*(i)$ such that $x^*_{if_*(i)} > 0$, $i \in E_{x^*}$.

step 3: Let $E_o := E_{x^*}$.

step 4: If $E_o = E$, then f_*^∞ is an average optimal policy (STOP).
Otherwise: go to step 5.

step 5a: Choose a triple (i,a_i,j) that satisfies $i \in E\backslash E_o$, $a_i \in A(i)$, $j \in E_o$ and $p_{ia_ij} > 0$.

step 5b: Define $f_*(i) := a_i$, $E_o := E_o \cup \{i\}$; go to step 4.

THEOREM 4.5.1. *Algorithm XVI determines an average optimal policy within a finite number of steps.*

PROOF. The simplex method is finite and gives an optimal solution x^* of program (4.5.2). Let (ϕ_o,u^*) be an optimal solution of program (4.5.1). The algorithm terminates after a finite number of steps and determines a set E_o (possibly equal to E) such that $E\backslash E_o$ is closed under any policy. Similarly to proposition 4.2.2 it can be shown that E_{x^*} is closed under $P(f_*)$, where f_* is any completion of the function f_* already defined on E_o. Since the states of $E_o\backslash E_{x^*}$ are transient under $P(f_*)$ and are absorbed in E_{x^*} with probability 1, we have

(4.5.3) $\quad p^*_{ij}(f_*) = 0 \qquad i \in E_o,\ j \notin E_{x^*}.$

The complementary slackness property of linear programming (cf. corollary 1.3.1) and the choice of f_* in step 2 imply that

(4.5.4) $\quad \phi_o + u^*_i = r_i(f_*) + (P(f_*)u)_i, \qquad i \in E_{x^*}.$

From (4.5.3) and (4.5.4) it follows that

(4.5.5)
$$\begin{aligned}\phi_i(f^\infty_*) &= \textstyle\sum_j p^*_{ij}(f_*)r_j(f_*)\\ &= \textstyle\sum_j p^*_{ij}(f_*)\{\phi_o + u_j - \sum_k p_{jk}(f_*)u_k\}\\ &= \phi_o\cdot\textstyle\sum_j p^*_{ij}(f_*) + [P^*(f_*)(I-P(f_*)u]_i\\ &= \phi_o, \quad i \in E_o.\end{aligned}$$

Hence, f^∞_* is average optimal on the set E_o.
Suppose that $E_o \neq E$. Let g^∞ be a pure and stationary average optimal policy. The policy f^∞_1 defined by

$$f_1(i) := \begin{cases} f(i) & i \in E_o \\ g(i) & i \in E\backslash E_o \end{cases}$$

is also average optimal and the Markov chain induced by $P(f_1)$ has an ergodic set, say $E_1(f_1)$, in E_o. Obviously, $\phi_i(f^\infty_1) = \phi_o = \mathit{max}_{j\in E}\phi_j(f^\infty_1), i \in E_1(f_1)$. Then, assumption 4.5.1 is contradictory to the fact that $E\backslash E_o$ is closed under any policy. Consequently, we have shown that $E_o = E$. Then, (4.5.5) implies that f^∞_* is average optimal. □

<u>REMARK 4.5.1</u>. In DENARDO & FOX [1968] the so-called *general single chain* case is treated, i.e. the case in which there exists a pure and stationary average optimal policy f^∞ such that the Markov chain induced by $P(f)$ has one ergodic set plus a (perhaps empty) set of transient states. They claim that in this case an average optimal policy can be obtained by algorithm XVI. In example 4.5.1 we show that this is in general not true since the algorithm may terminate with $E_o \neq E$. However, in the general single chain case an average optimal policy can be obtained by successive application of algorithm

XVI on $E \backslash E_o$ until $E_o = E$.

EXAMPLE 4.5.1. It can easily be verified that the model of figure 4.5.1 belongs to the general single chain case. The linear program is:

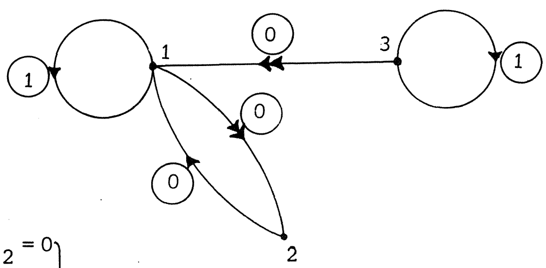


Figure 4.5.1

$$max\left\{x_{11}+x_{31} \left| \begin{array}{r} x_{12}-x_{21} \quad -x_{32} = 0 \\ -x_{12}+x_{21} \quad = 0 \\ x_{32} = 0 \\ x_{11}+x_{12}+x_{21}+x_{31}+x_{32} = 1 \\ x_{11}, x_{12}, x_{21}, x_{31}, x_{32} \geq 0 \end{array} \right.\right\}$$

x^* is an extreme optimal solution where $x^*_{11} = x^*_{12} = x^*_{21} = x^*_{32} = 0$, $x^*_{31} = 1$. Since $E_{x^*} = \{3\}$ and $E \backslash E_{x^*}$ is closed under any policy, algorithm XVI gives not an optimal policy.

4.6. THE COMPLETELY ERGODIC AND THE UNICHAIN CASE

We first discuss the *completely ergodic* AMD-model, i.e. the AMD-model under the following assumption.

ASSUMPTION 4.6.1. *For any pure and stationary policy* f^∞ *all states belong to a single ergodic set in the Markov chain induced by* $P(f)$.

This case is the classical one and the solution by linear programming is well-known. We discuss in this monograph the completely ergodic case by reason of completeness. The linear programming formulation was first presented by MANNE [1960] and DE GHELLINCK [1960]. The algorithm is similar to algorithm XVI but the steps 3 until 5 are superfluous because there are no transient states. Hence, we obtain the following algorithm.

ALGORITHM XVII for the construction of a pure and stationary average optimal policy (completely ergodic case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$(4.6.1)\qquad \max\left\{\sum_i\sum_a r_{ia}x_{ia}\;\middle|\;\begin{array}{ll}\sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j\in E\\ \sum_i\sum_a x_{ia} = 1 & \\ x_{ia}\geq 0,\ a\in A(i), & i\in E\end{array}\right\}$$

step 2: Take $f_*(i)$ such that $x^*_{if_*(i)} > 0$, $i \in E$.

LEMMA 4.6.1. *If the Markov chain induced by* $P(f)$ *has at most one ergodic set for every* $f^\infty \in C_D$*, then the Markov chain induced by* $P(\pi)$ *has also at most one ergodic set for every* $\pi^\infty \in C_S$.

PROOF. Suppose that there is a $\pi^\infty \in C_S$ such that the Markov chain induced by $P(\pi)$ has more than one ergodic set. Then we can write

$$P(\pi) = \begin{pmatrix} P_1 & 0 & 0\\ 0 & P_2 & 0\\ R_1 & R_2 & Q\end{pmatrix}\quad\text{, where } P_1 \neq 0 \text{ and } P_2 \neq 0.$$

Define $f^\infty \in C_D$ by $f(i) := a_i$ such that $\pi_{ia_i} > 0$, $i \in E$. Notice that $p_{ij}(\pi)=0$ implies $p_{ij}(f) = 0$. Hence the Markov chain induced by $P(f)$ has also at least two ergodic sets. This yields a contradiction. □

From assumption 4.6.1 and lemma 4.6.1 it follows that for any stationary policy π^∞ the Markov chain induced by $P(\pi)$ has exactly one ergodic set. Furthermore by the same argument as used in lemma 4.6.1 it can be shown that there are no transient states. Hence, the theorems 2.3.2 and 2.3.3 imply that $P^*(\pi)$ has identical rows, say $p^*(\pi)$, with $p^*(\pi) \gg 0$ and such that $p^*(\pi)$ is the unique solution of the so-called *steady-state equations:*

$$(4.6.2)\qquad \begin{cases}\sum_i(\delta_{ij}-p_{ij}(\pi))x_i = 0 & j\in E\\ \sum_i x_i = 1. & \end{cases}$$

For any $\pi^\infty \in C_S$ we define $x(\pi)$ by

$$(4.6.3)\qquad x_{ia}(\pi) := p^*_i(\pi)\cdot\pi_{ia}\qquad a\in A(i),\ i\in E$$

and for any feasible solution x of the linear program (4.6.1) we define $\pi^\infty(x)$ by

(4.6.4) $\pi_{ia}(x) := x_{ia} \big/ \sum_a x_{ia}$ $\quad a \in A(i),\ i \in E.$

THEOREM 4.6.1. *The mapping* $x_{ia}(\pi) = p_i^*(\pi)\cdot\pi_{ia}$ $\ a \in A(i),\ i \in E$, *is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the linear programming problem* (4.6.1) *with* (4.6.4) *as the inverse mapping. Furthermore, this mapping has the property that pure policies correspond to extreme feasible solutions.*

PROOF. Let π^∞ be any stationary policy. Then $x(\pi)$ defined by (4.6.3) satisfies

$$\textstyle\sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia}(\pi) = \sum_i(\delta_{ij}-p_{ij}(\pi))p_i^*(\pi) = 0 \qquad j \in E,$$

$$\textstyle\sum_i\sum_a x_{ia}(\pi) = \sum_i p_i^*(\pi) = 1 \quad \text{and} \quad x_{ia}(\pi) \ge 0 \qquad a \in A(i),\, i \in E.$$

Hence, $x(\pi)$ is a feasible solution of program (4.6.1).
Let x be an arbitrarily chosen feasible solution of (4.6.1). Then, $\pi_{ia}(x)$ is well-defined on E_x and $x_{ia} = \pi_{ia}(x)\cdot x_i$, $a \in A(i)$, $i \in E$, where $x_i := \Sigma_a x_{ia}$ and $\pi_{ia}(x)$ is arbitrarily chosen on $E\backslash E_x$. We obtain

$$0 = \textstyle\sum_i\sum_a(\delta_{ij}-p_{iaj})\pi_{ia}(x)\cdot x_i = \sum_i(\delta_{ij}-p_{ij}(\pi(x)))\cdot x_i \qquad j \in E$$

$$1 = \textstyle\sum_i\sum_a \pi_{ia}(x)\cdot x_i = \sum_i x_i,$$

implying that x is a solution of the steady-state equations. Hence, $x_i = p_i^*(\pi(x))$, $i \in E$. Therefore, it follows that $\pi(x)$ is well-defined on E and that $x = x(\pi(x))$, i.e. $\pi^\infty(x)$ is well-defined and the mapping (4.6.3) is onto. Since $\pi_{ia}(x(\pi)) = \pi_{ia}$ $\ a \in A(i)$, $i \in E$, the mapping is one-to-one and (4.6.4) is the inverse mapping.

Let f^∞ be any pure and stationary policy. Suppose that $x(f)$ is not an extreme point, i.e. $x(f) = \lambda x^1 + (1-\lambda)x^2$ where $\lambda \in (0,1)$, $x^1 \neq x^2$ and x^1, x^2 are feasible solutions of (4.6.1). Since $x^1_{ia} = x^2_{ia} = x_{ia}(f) = 0$, $a \neq f(i)$ $i \in E$, x^1 and x^2 are feasible solutions of the linear system

$$\begin{cases} x^T(I-P(f)) = 0 \\ x^Te \qquad\quad\ \ = 1. \end{cases}$$

This system has a unique solution and consequently $x^1 = x^2$, implying a contradiction. Hence, we have shown that x(f) is an extreme solution of (4.6.1). Conversely, let x be any extreme feasible solution of program (4.6.1). Since the sum of the first N components yields a zero in every column, the rank of the system of the N+1 equations is at most N. Therefore, any extreme solution has at most N positive components. Since $\Sigma_a x_{ia} > 0$, $i \in E$, x has in each state i exactly one positive component. Hence, the corresponding policy is pure. This completes the proof. □

Consider the policy improvement method for the completely ergodic case. Since $\phi(f^\infty)$ has identical components, we may replace $\phi(f^\infty)$ by $\phi_o(f^\infty)\cdot e$, where $\phi_o(f^\infty) \in \mathbb{R}^1$. Furthermore, we remark that the set A(i,f) defined by (4.4.2) becomes

$$A(i,f) = \{a \in A(i) \mid \phi_o(f^\infty) + \sum_j (\delta_{ij} - p_{iaj}) u_j(f^\infty) < r_{ia}\}.$$

Look at one iteration of the policy improvement algorithm. If $A(i,f) = \emptyset$, then $g(i) := f(i)$. Otherwise, we may take g(i) from A(i,f). By theorem 4.6.1 the vector x(f) defined by (4.6.3) is an extreme feasible solution of the linear program (4.6.1). The dual program of (4.6.1) is

$$min\{\tilde{\phi} \mid \tilde{\phi} + \sum_j (\delta_{ij} - p_{iaj}) \tilde{u}_j \geq r_{ia}\}.$$

In the simplex tableau corresponding to x(f), the column of a nonbasic $x_{ia}(f)$ has in the transformed objective function the value (cf. theorem 1.4.1 and tableau (1.4.2))

$$(4.6.5) \qquad d_{ia} = \tilde{\phi} + \sum_j (\delta_{ij} - p_{iaj}) \tilde{u}_j - r_{ia}.$$

Since $x_{if(i)}(f) > 0$, $i \in E$, it follows from the orthogonality of the corresponding primal and dual variables in the simplex tableau that $d_{if(i)} = 0$, $i \in E$. Then, we obtain

$$\tilde{\phi}\cdot e = P^*(f)(\tilde{\phi}\cdot e) = P^*(f)\{r(f) - (I-P(f))\tilde{u}\} = P^*(f)r(f) = \phi(f^\infty).$$

Since

$$\begin{aligned} \phi(f^\infty) + (I-P(f))u(f\ ) &= \phi(f^\infty) + (I-P(f))D(f)r(f) \\ &= \phi(f^\infty) + (I-P^*(f))r(f) \\ &= r(f), \end{aligned}$$

We have

$$(I-P(f))(u(f^\infty)-\tilde{u}) = 0.$$

Then

$$u(f^\infty) - \tilde{u} = P^*(f)(u(f^\infty)-\tilde{u}).$$

Because $P^*(f)$ has identical rows, $u(f^\infty)-\tilde{u}$ has identical components and consequently

$$\sum_j(\delta_{ij}-p_{iaj})\tilde{u}_j = \sum_j(\delta_{ij}-p_{iaj})u_j(f^\infty).$$

Hence, (4.6.5) can be written as

(4.6.6) $$d_{ia} = \phi_o(f^\infty) + \sum_j(\delta_{ij}-p_{iaj})u_j(f^\infty) - r_{ia}.$$

Since $a \in A(i,f)$ if and only if $d_{ia} < 0$, it follows that the set of actions from which $g(i)$ can be chosen corresponds to the possible choices for the pivot column in the simplex method. Hence, we have shown the following.

CONCLUSIONS.

1. *Any policy improvement algorithm is equivalent to a block-pivoting simplex algorithm.*
2. *The standard simplex algorithm is equivalent to a particular policy improvement algorithm.*

We continue this section under the following assumption *(unichainedness)*.

ASSUMPTION 4.6.2. *For any pure and stationary policy* f^∞*, the Markov chain induced by* $P(f)$ *has one ergodic set plus a (perhaps empty) set of transient states.*

In this case an optimal policy can be determined by the following algorithm.

ALGORITHM XVIII for the construction of a pure and stationary average optimal policy (unichain case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$(4.6.7) \qquad \max\left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = 0 & j \in E \\ \sum_i \sum_a x_{ia} = 1 & \\ x_{ia} \geq 0,\ a \in A(i), & i \in E \end{array} \right\}.$$

step 2: Take f_*^∞ such that

$$f_*(i) := \begin{cases} a_i \quad \text{where } x^*_{ia_i} > 0 & i \in E_{x^*} \\ \text{arbitrarily} & i \in E \backslash E_{x^*}. \end{cases}$$

THEOREM 4.6.2. *Algorithm XVIII provides a pure and stationary average optimal policy in the unichain case.*

PROOF. Since the Markov chain induced by $P(f_*)$ has only one ergodic set and since E_{x^*} is closed under $P(f_*)$ (the proof is similar to the proof of proposition 4.2.2), it follows that the states of $E \backslash E_{x^*}$ are transient under $P(f_*)$. Then, the proof of the theorem is similar to the proof of theorem 4.5.1. □

REMARK 4.6.1. In the unichain case there is in general no one-to-one correspondence between the feasible solutions of program (4.6.7) and the stationary policies.

4.7. ADDITIONAL CONSTRAINTS

4.7.1. INTRODUCTION

We will discuss the problem of finding an optimal policy when there are some additional constraints on the limit points of the expected state-action frequencies. Such problems may for instance occur if more than one reward function is of importance. Then we want to maximize the expected average reward with regard to one reward function while we restrict the other reward functions by some bounds.

DERMAN [1970], chapter 7, has considered the unichain case and he has solved this problem by linear programming. In DERMAN & VEINOTT [1972] an iterative algorithm, based on the Dantzig-Wolfe principle was proposed. They write "until the faces of the linear programming polytope are found, routine application of the simplex method is generally not possible". Therefore, they need the decomposition principle.

In section 4.7.2 we shall characterize this linear programming polytope and we prove some properties of the limit points of the state-action frequencies. We present a treatment of the general multichain case based on the solution of one linear program.

In general, there does not exist a stationary optimal solution. We will derive an algorithm for the construction of a memoryless optimal policy. For practical purposes, this algorithm needs too many calculations; furthermore, memoryless (i.e. Markov) policies are unusual in practice.

Fortunately, if certain conditions are satisfied, then optimal policies can be computed that are stationary. In section 4.7.4 we shall discuss these conditions.

We close the treatment of additional constraints with a description in section 4.7.5 of the unichain case. In this case a stationary optimal policy can always be found. We shall show this result by a proof different from the proof of theorem 3 on page 95 in DERMAN [1970] and we present an algorithm to perform the calculations.

4.7.2. LIMIT POINTS OF STATE-ACTION FREQUENCIES

Since the state-action frequencies depend on the initial distribution we assume that $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$ is a known initial distribution, i.e. $\beta_j \geq 0$, $j \in E$, and $\Sigma_j \beta_j = 1$.

<u>REMARK 4.7.1</u>. In contrast with the use of the vector β in the sections 4.2, 4.3 and 4.4, we allow in this section that $\beta_i = 0$ for some $i \in E$. DERMAN & VEINOTT [1972] discuss the constrained problem for a fixed starting state i.

For any policy R and any $T \in \mathbb{N}$, we denote the *expected state-action frequencies* in the first T periods by $x^T(R)$, i.e.

$$(4.7.1) \qquad x^T_{ja}(R) := \frac{1}{T} \sum_{t=1}^{T} \sum_i \beta_i \cdot \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) \qquad a \in A(j),\ j \in E.$$

By $X(R)$ we denote the set of all limit points of the vectors $\{x^T(R),\ T = 1,2,\ldots\}$. These limit points are limit points in the vector space of the vectors $x^T(R)$. Any $x^T(R)$ satisfies $\Sigma_j \Sigma_a x^T_{ja}(R) = 1$ and therefore also

$\Sigma_j\Sigma_a x_{ja}(R) = 1$ for every $x(R) \in X(R)$. Furthermore, if $x^{T_k}(R) \to x(R)$ for $k \to \infty$, then $\lim_{k\to\infty} x_{ja}^{T_k}(R) = x_{ja}(R)$ for all $a \in A(j)$, $j \in E$.
Let $C_1 := \{R \in C \,|\, |X(R)| = 1\}$. In section 4.3 we have already seen that for any stationary policy π^∞ the set $X(\pi^\infty)$ consists of one element, namely

(4.7.2) $\qquad X(\pi^\infty) = \{x(\pi)\}$, where $x_{ja}(\pi) := [\beta^T P^*(\pi)]_j \cdot \pi_{ja}$, $a \in A(j)$, $j \in E$.

Hence, C_1 contains all stationary policies.

We introduce the following notations:

$$L \;\;:= \{x(R) \in X(R) \,|\, R \in C\}$$
$$L(M) := \{x(R) \in X(R) \,|\, R \in C_M\}$$
$$L(C) := \{x(R) \in X(R) \,|\, R \in C_1\}$$
$$L(S) := \{x(R) \in X(R) \,|\, R \in C_S\}$$
$$L(D) := \{x(R) \in X(R) \,|\, R \in C_D\}.$$

THEOREM 4.7.1. $\overline{L(D)} = \overline{L(S)} = L(C) = L(M) = L$.

PROOF. (cf. DERMAN [1970] pp.93-94). It is obvious that $L(D) \subset L(S) \subset L(C) \subset L$. We first prove that $L \subset \overline{L(D)}$. Suppose the contrary. Then, there exists a policy R such that $x(R) \in L$ and $x(R) \notin \overline{L(D)}$. Since $\overline{L(D)}$ is a closed convex set, it follows from theorem 1.2.1 that there exist coefficients r_{ja} such that

(4.7.3) $\qquad \sum_j\sum_a r_{ja}x_{ja}(R) > \sum_j\sum_a r_{ja}x_{ja} \quad$ for all $x \in \overline{L(D)}$.

Theorem 4.2.3 implies that there is for the AMD-model with rewards r_{ia} a pure and stationary policy f^∞ which is optimal with respect to the utility function $\hat{\phi}$, defined in (4.2.9). Because $x(R) \in L$, there is a sequence $\{T_k,\ k = 1,2,\ldots\}$ such that

$$x_{ja}(R) = \lim_{k\to\infty} x_{ja}^{T_k}(R) \qquad a \in A(j),\ j \in E.$$

Hence,

$$\sum_j\sum_a r_{ja}x_{ja}(R) = \sum_j\sum_a r_{ja}\cdot\lim_{k\to\infty} x_{ja}^{T_k}(R)$$
$$= \lim_{k\to\infty} \frac{1}{T_k} \sum_{t=1}^{T_k}\sum_i \beta_i \cdot \sum_j\sum_a \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i)\cdot r_{ja}$$

$$= \sum_i \beta_i \cdot \lim_{k\to\infty} \frac{1}{T_k} \sum_{t=1}^{T_k} \sum_j \sum_a \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i) \cdot r_{ja}$$

$$\leq \sum_i \beta_i \cdot \limsup_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \sum_j \sum_a \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i) \cdot r_{ja} =$$

$$\beta^T \hat{\phi}(R) \leq \beta^T \hat{\phi}(f^\infty) = \sum_j \sum_a r_{ja} x_{ja}(f^\infty),$$

which contradicts (4.7.3): we have shown that $L \subset \overline{L(D)}$.
Since $L(D) \subset L(S) \subset L \subset \overline{L(D)}$, we obtain $\overline{L(S)} = \overline{L(D)}$. From

$$L(M) \subset L \subset \overline{L(S)} = \overline{L(D)}$$

and

$$L(C) \subset L \subset \overline{L(S)} = \overline{L(D)}$$

it follows that for the proof of the theorem it remains to prove that $\overline{L(D)} \subset L(M) \cap L(C)$.

Therefore, take any $x \in \overline{L(D)}$. Let $C_D = \{f_1^\infty, f_2^\infty, \ldots, f_n^\infty\}$. Then we can write

$$x_{ja} = \sum_{k=1}^{n} p_k x_{ja}(f_k) \qquad a \in A(j),\ j \in E,$$

for certain $p_k \geq 0$ such that $\sum_{k=1}^{n} p_k = 1$.
The existence of a Markov policy R satisfying

$$\sum_i \beta_i \cdot \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i) =$$

$$\sum_i \beta_i \cdot \sum_k p_k \mathbb{P}_{f_k^\infty}(X_t = j, Y_t = a \mid X_1 = i) \qquad t \in \mathbb{N},\ a \in A(j),\ j \in E,$$

is shown in theorem 2.5.1. Hence,

$$x_{ja} = \sum_k p_k x_{ja}(f_k)$$

$$= \sum_k p_k \cdot \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_i \beta_i \cdot \mathbb{P}_{f_k^\infty}(X_t = j, Y_t = a \mid X_1 = i)$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_i \beta_i \cdot \sum_k p_k \mathbb{P}_{f_k^\infty}(X_t = j, Y_t = a \mid X_1 = i)$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_i \beta_i \cdot \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i)$$

$$= x_{ja}(R) \qquad \text{for all } a \in A(j),\ j \in E.$$

Consequently, $x = x(R) \in L(M)$ and $x = \lim_{T\to\infty} x^T(R) \in L(C)$, which completes the proof of the theorem. □

REMARK 4.7.2. Theorem 4.7.1 shows that for any utility function, which is based on the limit points of the expected state-action frequencies, it is sufficient to consider only the policies of class C_1. For instance, the "weak" criterion $\phi(R)$ and the "strong" criterion $\hat{\phi}(R)$ are in fact the same optimality criterion, since $\phi(R) = \hat{\phi}(R)$ for any $R \in C_1$ (cf. theorem 4.2.3).

We are interested in the problem to find, for a given initial distribution, a policy which is optimal in the set of policies that satisfy some additional constraints. These constraints will be linear functions of the expected state-action frequencies.

Let $\Sigma_i \Sigma_a q_{iak} x_{ia}(R) \le b_k$ be the k-th constraint. Then we formulate the *constrained Markov decision problem* by

$$(4.7.4) \qquad \sup_R \left\{ \phi(\beta,r) \,\middle|\, \begin{array}{ll} \sum_i \sum_a q_{iak} x_{ia}(R) \le b_k & k = 1,2,\dots,m \\ x(R) \in X(R) & \end{array} \right\},$$

where $\phi(\beta,R) := \liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T \Sigma_j \Sigma_a \Sigma_i \beta_i \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i)\, r_{ja}$.
By the result of theorem 4.7.1 we may replace (4.7.4) by

$$(4.7.5) \qquad \sup_{R\in C_1} \{ \phi(\beta,R) \mid \sum_i \sum_a q_{iak} x_{ia}(R) \le b_k \quad k = 1,2,\dots,m\}.$$

Notice that for $R \in C_1$ $\phi(\beta,R) = \Sigma_j \Sigma_a x_{ja}(R) r_{ja}$.

In order to solve problem (4.7.5), we propose - inspired by the linear programming formulation for the unconstrained Markov decision problem, given in section 4.2 - to study the following linear programming problem:

$$(4.7.6) \qquad \max \left\{ \sum_i \sum_a r_{ia} x_{ia} \,\middle|\, \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} & & = 0, \ , j \in E \\ \sum_a x_{ja} & + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} & = \beta_j, \ j \in E \\ \sum_i \sum_a q_{iak} x_{ia} & & \le b_k, \ 1 \le k \le m \\ & x_{ia}, y_{ia} \ge 0, & a \in A(i),\ i \in E \end{array} \right\}$$

The fact that program (4.7.6) can be used to solve problem (4.7.5) is based upon the following theorem. Consider the linear system

$$(4.7.7)\quad \begin{cases} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} & = 0, \quad j \in E \\ \sum_a x_{ja} + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} & = \beta_j, \quad j \in E \\ x_{ia}, y_{ia} \geq 0,\ a \in A(i), & i \in E \end{cases}.$$

Define the set X by

$$(4.7.8)\quad X := \{x \mid \text{there exists a } y \text{ such that } (x,y) \text{ is feasible for (4.7.7)}\}.$$

THEOREM 4.7.2. L = X.

PROOF. Theorem 4.7.1 implies that it is sufficient to prove that $\overline{L(D)} = X$. From theorem 4.3.1 it follows that $L(S) \subset X$ (it can easily be checked that the proof of theorem 4.3.1 may also be used when $\beta_j = 0$ for some $j \in E$). Hence, certainly $L(D) \subset X$.
Since X is the projection of a polyhedron, X is also a polyhedron and consequently $\overline{L(D)} \subset X$. From (4.7.7) it follows that $x_{ia} \geq 0$ for all $a \in A(i)$, $i \in E$, and that $\Sigma_i\Sigma_a x_{ia} = 1$. Therefore, X is a polytope, i.e. X is a bounded polyhedron. Then from corollary 1.2.2 it follows that X is the closed convex hull of a finite number of extreme points. Hence, it is sufficient to show that any extreme point of X belongs to L(D).
Let $\bar{x}$ be an arbitrarily chosen extreme point of X, and let $\bar{X}$ be the closed convex hull of the extreme points of X that are different from $\bar{x}$. Then $\bar{x} \notin \bar{X}$ and theorem 1.2.1 implies the existence of coefficients r_{ia} $a \in A(i)$, $i \in E$ such that

$$(4.7.9)\quad \sum_i\sum_a r_{ia}\bar{x}_{ia} > \sum_i\sum_a r_{ia}x_{ia} \quad \text{for every } x \in \bar{X}.$$

Therefore it follows from (4.7.9) that any optimal solution (x^*,y^*) of the linear program

$$(4.7.10)\quad \max\left\{\sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j \in E \\ \sum_a x_{ja} + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} = \beta_j, & j \in E \\ x_{ia}, y_{ia} \geq 0,\ a \in A(i), & i \in E \end{array}\right\}$$

satisfies $x^* = \bar{x}$.
Consider the AMD-model with rewards r_{ia}, $a \in A(i)$, $i \in E$. Let f_*^∞ be any pure and stationary average optimal policy. Then $(x(f_*),y(f_*))$, defined

in (4.3.2), is by theorem 4.3.3 an optimal solution of program (4.7.10). Hence, $\bar{x} = x(f_*) \in L(D)$, which completes the proof. □

REMARK 4.7.3. Recently, we learned from VEINOTT [1973] the result of theorem 4.7.2 was already known to him in 1973.

REMARK 4.7.4. From the theorems 4.7.1 and 4.7.2 it follows that any extreme point of X is an element of L(D). The next example shows that the converse statement is not true, in general. Furthermore, this example displays that L(S) ≠ X is possible, and that X is a real subset of

$$\left\{x \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j \in E\\ \sum_i\sum_a x_{ia} = 1 & \\ x_{ia} \ge 0,\ a \in A(i), & i \in E \end{array}\right\}$$

EXAMPLE 4.7.1. Consider the model of figure 4.7.1 and write for any stationary policy π^∞ the transition matrix as

$$P(\pi) = \begin{pmatrix} 0 & \pi_1 & 1-\pi_1 \\ \pi_2 & 1-\pi_2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

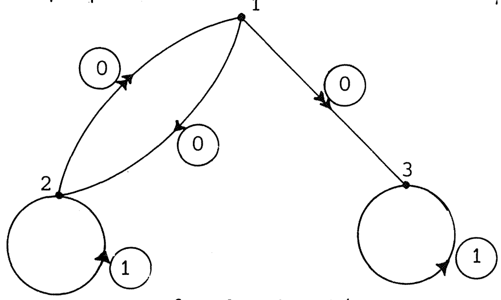


Figure 4.7.1

It can easily be verified that $P^*(\pi)$ and $x(\pi)$ are given by:

a. $\pi_1 = 1$:

$$P^*(\pi) = \begin{pmatrix} \pi_2\cdot(1+\pi_2)^{-1} & (1+\pi_2)^{-1} & 0 \\ \pi_2\cdot(1+\pi_2)^{-1} & (1+\pi_2)^{-1} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$x_{11}(\pi) = x_{22}(\pi) = (2/3)\cdot\pi_2\cdot(1+\pi_2)^{-1};$$
$$x_{21}(\pi) = (2/3)\cdot(1-\pi_2)\cdot(1+\pi_2)^{-1};$$
$$x_{12}(\pi) = 0;\ x_{31}(\pi) = 1/3.$$

b. $\pi_2 = 0$ and $\pi_1 \neq 1$:

$$P^*(\pi) = \begin{pmatrix} 0 & \pi_1 & 1-\pi_1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$x_{11}(\pi) = x_{12}(\pi) = x_{22}(\pi) = 0;$$
$$x_{21}(\pi) = (1/3)\cdot(1+\pi_1);\ x_{31} = (1/3)\cdot(2-\pi_1).$$

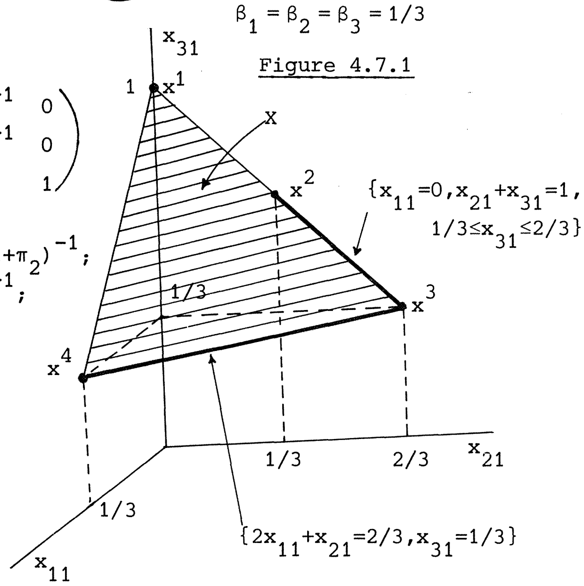


Figure 4.7.2

c. $\pi_2 \neq 0$ and $\pi_1 \neq 1$:

$$P^*(\pi) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$x_{11}(\pi) = x_{12}(\pi) = x_{21}(\pi) = x_{22}(\pi) = 0; \; x_{31}(\pi) = 1.$$

Since we always have that $x_{12} = 0$ and $x_{22} = x_{11}$, we can draw the sets L(D), L(S) and X in the 3-dimensional space with coordinates x_{11}, x_{21} and x_{31} (see figure 4.7.2).

L(D) $= \{x^1,x^2,x^3,x^4\}$, where x^i, $1 \leq i \leq 4$, is drawn in figure 4.7.2.

L(S) consists of x^1 and the points between x^2 and x^3, together with the points between x^3 and x^4 (the dark lines in the figure).

X is the convex hull of $\{x^1,x^2,x^3,x^4\}$, i.e. the polytope

$$\left\{ x \left| \begin{array}{l} x_{11}+x_{12}+x_{21}+x_{22}+x_{31} = 1; \; x_{12}=0; \; x_{11}=x_{22} \\ x_{11},x_{12},x_{21},x_{22} \geq 0; \; x_{31} \geq 1/3 \end{array} \right. \right\}.$$

In figure 4.7.2 we see that x^2 is not an extreme point of X, although $x^2 \in$ L(D). Moreover, it follows that L(S) $\neq$ X.

4.7.3. COMPUTATION OF A MARKOVIAN OPTIMAL POLICY

In this section we present an algorithm for the construction of a Markovian optimal policy. We first show that the problems (4.7.5) and (4.7.6) are strongly related.

<u>THEOREM 4.7.3.</u>

(i) *Problem* (4.7.5) *is feasible if and only if problem* (4.7.6) *is feasible.*

(ii) *The optima of the problems* (4.7.5) *and* (4.7.6) *are equal.*

(iii) *If* R *is an optimal solution of problem* (4.7.5), *then* x(R) *is an optimal solution of problem* (4.7.6).

(iv) *Let* (x,y) *be an optimal solution of problem* (4.7.6), *and let* $x = \Sigma_{k=1}^n p_k x(f_k)$, *where* $p_k \geq 0$ *such that* $\Sigma_k p_k = 1$, *and* $\{f_1,f_2,\ldots,f_n\} = C_D$. *Suppose that* $R \in C_M$ *is the policy, introduced in theorem* 2.5.1, *such that*

(4.7.11) $\qquad \sum_i \beta_i \cdot \mathbb{P}_R(X_t = j,\ Y_t = a \mid X_1 = i) =$

$$\sum_i \beta_i \cdot \sum_k p_k \mathbb{P}_{f_k^\infty}(X_t = j,\ Y_t = a \mid X_1 = i) \qquad t \in \mathbb{N},\ a \in A(j),\ j \in E.$$

Then, R *is an optimal solution of problem* (4.7.5).

PROOF. The theorems 4.7.1 and 4.7.2 imply that X = L(C). Moreover, any $R \in C_1$ satisfies $\phi(\beta,R) = \Sigma_j \Sigma_a x_{ja}(R) r_{ja}$. By these observations, the parts (i), (ii) and (iii) are straightforward.
For the proof of part (iv) we can similarly as in the proof of theorem 4.7.1 show that x = x(R), and $R \in C_1$. Consequently,

$$\phi(\beta,R) = \sum_i \sum_a r_{ia} x_{ia}(R) = \sum_i \sum_a r_{ia} x_{ia} = \text{optimum (4.7.6)}.$$

Hence, R is an optimal solution of problem (4.7.5). □

REMARK 4.7.5. To compute an optimal policy from an optimal solution (x^*,y^*) of the linear program (4.7.6), we first have to write x^* as

$$x^* = \sum_k p_k x(f_k), \quad \text{where } p_k \geq 0 \text{ and } \sum_k p_k = 1.$$

Next, we have to determine $R = (\pi^1,\pi^2,\ldots) \in C_M$ such that R satisfies (4.7.11). The decision rules π^t, $t \in \mathbb{N}$, can be obtained from DERMAN & STRAUCH [1966].

ALGORITHM XIX for the construction of an optimal Markov policy in a constrained AMD-model.

step 1: Determine an optimal solution (x^*,y^*) of the linear programming problem

(4.7.12)
$$\max\left\{ \sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{lr} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} & = 0\ ,\ j \in E \\ \sum_a x_{ja} \qquad + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} & = \beta_j,\ j \in E \\ \sum_i \sum_a q_{iak} x_{ia} & \leq b_k,\ 1 \leq k \leq m \\ & x_{ia}, y_{ia} \geq 0,\ a \in A(i),\ i \in E \end{array} \right\}$$

(if problem (4.7.12) is infeasible, then problem (4.7.5) is also infeasible).

step 2a: Suppose that $C_D = \{f_1^\infty, f_2^\infty, \dots, f_n^\infty\}$. Compute $P^*(f_k)$ by algorithm III, $k = 1,2,\dots,n$.

step 2b: Take

$$x_{ja}^k := \begin{cases} [\beta^T P^*(f_k)]_j & a = f_k(j) \\ 0 & a \neq f_k(j) \end{cases} \qquad j \in E,\ k = 1,2,\dots,n.$$

step 3: Determine p_k $(k = 1,2,\dots,n)$ as a feasible solution of the linear system

$$(4.7.13) \quad \begin{cases} \sum_k p_k x_{ja}^k = x_{ja}^* & a \in A(j),\ j \in E \\ \sum_k p_k = 1 & \\ p_k \geq 0 & k = 1,2,\dots,n \end{cases}$$

(this can be performed by the so-called phase I of the simplex method).

step 4: $R^* := (\pi^1, \pi^2, \dots)$, where

$$\pi_{ja}^t := \begin{cases} \dfrac{\sum_i \beta_i \cdot \sum_k p_k [P^{t-1}(f_k)]_{ij} \cdot \delta_{af_k(j)}}{\sum_i \beta_i \cdot \sum_k p_k [P^{t-1}(f_k)]_{ij}} & \text{if } \sum_i \beta_i \cdot \sum_k p_k [P^{t-1}(f_k)]_{ij} \neq 0 \\ \text{arbitrarily} & \text{if } \sum_i \beta_i \cdot \sum_k p_k [P^{t-1}(f_k)]_{ij} = 0. \end{cases}$$

Then, R^* is an optimal Markov policy for the constrained AMD-model.

REMARK 4.7.6. Algorithm XIX is inattractive for practical problems. The number of calculations is prohibitive. Moreover, the use of Markov policies is inefficient in practice. Therefore, in the next section we discuss the problem of finding an optimal stationary policy, if one exists.

EXAMPLE 4.7.2. We apply algorithm XIX to the model of figure 4.7.3 with additional constraints $\frac{1}{4} \leq x_{21}(R) \leq \frac{1}{2}$. Since for any policy R we have $x_{11}(R) = x_{12}(R) = x_{32}(R) = 0$, we can illustrate the points x(R) in the 2-dimensional space with coordinates x_{21} and x_{31}. It can easily be verified that any stationary policy π^∞ satisfies (see figure 4.7.4):

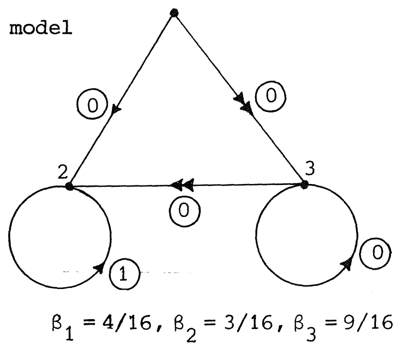


Figure 4.7.3

if $\pi_{31} \neq 1$: $x_{11}(\pi) = x_{12}(\pi) = x_{31}(\pi) = x_{32}(\pi) = 0; x_{21}(\pi) = 1$.

if $\pi_{31} = 1$: $x_{11}(\pi) = x_{12}(\pi) = x_{32}(\pi) = 0;$

$x_{21}(\pi) = (1/16)\cdot(3+4\pi_{11});$

$x_{31}(\pi) = (1/16)\cdot(13-4\pi_{11})$.

Let x^1, x^2, x^3 be the points corresponding to pure policies which are drawn in figure 4.7.4. Then

$L(D) = \{x^1, x^2, x^3\}.$

$L(S) = \{x^2\} \cup \overline{\{x^1, x^3\}}$

$L(M) = L(C) = L = X = \overline{\{x^1, x^2, x^3\}}.$

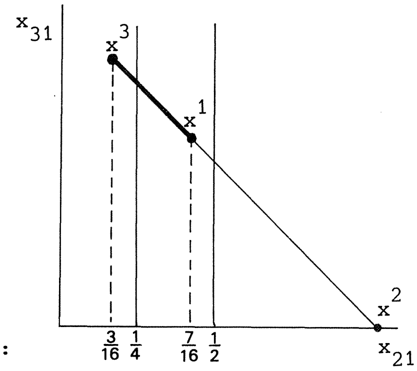


Figure 4.7.4

The formulation of program (4.7.12) becomes (if $p_{iai} = 1$, then the coefficients of the variable y_{ia} are all zeroes; therefore, we remove such variables from the formulation):

maximize x_{21}

subject to

$$\begin{array}{llllllllll}
x_{11} & +x_{12} & & & & & & & = 0 \\
-x_{11} & & & & -x_{32} & & & & = 0 \\
& -x_{12} & & & +x_{32} & & & & = 0 \\
x_{11} & +x_{12} & & & & +y_{11} & +y_{12} & & = 4/16 \\
& & x_{21} & & & -y_{11} & & -y_{32} & = 3/16 \\
& & & x_{31} & +x_{32} & & -y_{12} & +y_{32} & = 9/16 \\
& & x_{21} & & & & & & \leq 1/2 \\
& & -x_{21} & & & & & & \leq -1/4
\end{array}$$

$$x_{11}, x_{12}, x_{21}, x_{31}, x_{32}, y_{11}, y_{12}, y_{32} \geq 0$$

Algorithm XIX gives for this problem the following results.

<u>step 1</u>: $x^*_{11} = 0, x^*_{12} = 0, x^*_{21} = 1/2, x^*_{31} = 1/2, x^*_{32} = 0; y^*_{11} = 0, y^*_{12} = 1/4, y^*_{32} = 5/16.$

<u>step 2a</u>: Let f^∞_k, $k = 1,2,3,4$, be such that

$$f_1(1) = 1, f_1(2) = 1, f_1(3) = 1; \; f_2(1) = 1, f_2(2) = 1, f_2(3) = 2;$$
$$f_3(1) = 2, f_3(2) = 1, f_3(3) = 1; \; f_4(1) = 2, f_4(2) = 1, f_4(3) = 2.$$

By algorithm III we obtain

$$P^*(f_1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; ; \quad P^*(f_2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \; ;$$

$$P^*(f_3) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; ; \quad P^*(f_4) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

<u>step 2b</u>: $x^1_{11} = x^1_{12} = x^1_{32} = 0$; $x^1_{21} = 7/16$; $x^1_{31} = 9/16$.
$x^2_{11} = x^2_{12} = x^2_{32} = 0$; $x^2_{21} = 1$; $x^2_{31} = 0$.
$x^3_{11} = x^3_{12} = x^3_{32} = 0$; $x^3_{21} = 3/16$; $x^3_{31} = 13/16$.
$x^4_{11} = x^4_{12} = x^4_{32} = 0$; $x^4_{21} = 1$; $x^4_{31} = 0$.

<u>step 3</u>: $p_1 = 8/9$; $p_2 = 1/9$; $p_3 = 0$; $p_4 = 0$.
<u>step 4</u>: Since

$$P^t(f_1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } P^t(f_2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} , \; t \in \mathbb{N},$$

we get $R^* = (\pi^1, \pi^2, \ldots)$, where

$$\pi^t_{11} = 1 \quad t \in \mathbb{N}; \quad \pi^t_{21} = 1 \quad t \in \mathbb{N};$$

$$\pi^t_{31} = \begin{cases} 8/9 & t = 1 \\ 1 & t \geq 2; \end{cases} \quad \pi^t_{32} = \begin{cases} 1/9 & t = 1 \\ 0 & t \geq 2. \end{cases}$$

4.7.4. COMPUTATION OF A STATIONARY OPTIMAL POLICY (GENERAL CASE)

Suppose that we have obtained an optimal solution (x^*, y^*) of problem (4.7.12). Then we define the stationary policy $(\pi^*)^\infty$ by

$$(4.7.14) \quad \pi^*_{ia} := \begin{cases} x^*_{ia} / \sum_a x^*_{ia} & a \in A(i),\ i \in E_{x^*} \\ y^*_{ia} / \sum_a y^*_{ia} & a \in A(i),\ i \in E_{y^*} \backslash E_{x^*} \\ \text{arbitrarily} & \text{elsewhere.} \end{cases}$$

Then, $x_{ja}(\pi^*) = [\beta^T P^*(\pi^*)]_j \cdot \pi^*_{ja}$ $\quad a \in A(j),\ j \in E.$

<u>REMARK 4.7.7</u>. Since it is possible that $\beta_j = 0$ for some j, it is also possible that $E_{x^*} \cup E_{y^*} \neq E$. Therefore (4.7.14) differs from (4.3.1).

THEOREM 4.7.4. *If* $x^* = x(\pi^*)$, *then* $(\pi^*)^\infty$ *is an optimal solution of problem* (4.7.5).

PROOF. Since $x^* = x(\pi^*)$ it is obvious that $(\pi^*)^\infty$ is a feasible solution of (4.7.5). Moreover, by theorem 4.7.3,

$$\phi(\beta,(\pi^*)^\infty) = \sum_j\sum_a x^*_{ja} r_{ja} = \text{optimum } (4.7.6) = \text{optimum } (4.7.5),$$

i.e. $(\pi^*)^\infty$ is an optimal solution of problem (4.7.5). □

If we compute $P^*(\pi^*)$, which can be done by algorithm III, then we can check whether $x^*_{ja} = [\beta^T P^*(\pi^*)]_j \cdot \pi^*_{ja}$ $a \in A(j)$, $j \in E$. However, in certain cases we may decide that $x^* = x(\pi^*)$ without the computation of $P^*(\pi^*)$. In the following theorem we present some sufficient conditions for the property that $x^* = x(\pi^*)$.

THEOREM 4.7.5.

(i) *If the Markov chain under* $P(\pi^*)$ *has one ergodic set plus a (perhaps empty) set of transient states, then* $x^* = x(\pi^*)$.

(ii) *if* $y^*_{ia}/\Sigma_a y^*_{ia} = \pi^*_{ia}$ $a \in A(i)$, $i \in E_{x^*} \cap E_{y^*}$, *then* $x^* = x(\pi^*)$.

PROOF.

(i) From remark 4.3.1 it follows that x^* is a stationary probability distribution of the Markov chain induced by $P(\pi^*)$. Then theorem 2.3.3 implies that $x^*_i = p^*_{ii}(\pi^*)$, $i \in E$. Since the Markov chain under $P(\pi^*)$ has only one ergodic set, we have $x^*_{ia} = [\beta^T P^*(\pi^*)]_i \cdot \pi^*_{ia} = x_{ia}(\pi^*)$ $a \in A(i)$, $i \in E$.

(ii) Since $y^*_{ia}/y^*_i = \pi^*_{ia}$ $a \in A(i)$, $i \in E_{y^*}$, we have

$$\beta_j = \sum_a x^*_{ja} + \sum_i\sum_a(\delta_{ij}-p_{iaj})y^*_{ia} = x^*_j + \sum_i\sum_a(\delta_{ij}-p_{iaj})\pi^*_{ia}\cdot y^*_i$$

$$= x^*_j + \sum_i y^*_i\cdot(\delta_{ij}-p_{ij}(\pi^*)), \quad j \in E.$$

Therefore, (x^*,y^*) satisfies

$$\begin{cases}(x^*)^T = (x^*)^T P(\pi^*) \\ (x^*)^T = \beta^T - (y^*)^T(I-P(\pi^*)).\end{cases}$$

Consequently,

$$(x^*)^T = (x^*)^T P^*(\pi^*) = \beta^T P^*(\pi^*) - (y^*)^T (I-P(\pi^*))P^*(\pi^*)$$

$$= \beta^T P^*(\pi^*).$$

Hence,

$$x^*_{ia} = [\beta^T P^*(\pi^*)]_i \cdot \pi^*_{ia} = x_{ia}(\pi^*) \quad a \in A(i),\ i \in E. \quad \square$$

The next example shows that in general $(\pi^*)^\infty$ is not an optimal solution of problem (4.7.5) although in this example there exists a stationary optimal solution.

EXAMPLE 4.7.3. Consider the model of example 4.7.2 with the additional constraint $x_{21}(R) \le 1/4$. The optimal solution of the linear program is:

$$x^*_{11} = 0,\ x^*_{12} = 0,\ x^*_{21} = 1/4,\ x^*_{31} = 3/4,\ x^*_{32} = 0;$$

$$y^*_{11} = 0,\ y^*_{12} = 1/4,\ y^*_{32} = 1/16;\ \text{optimum} = 1/4.$$

The policy $(\pi^*)^\infty$ satisfies $\pi^*_{12} = \pi^*_{21} = \pi^*_{31} = 1$.
$(\pi^*)^\infty$ is not optimal, since

$$\phi(\beta,(\pi^*)^\infty) = \beta^T P^*(\pi^*) r(\pi^*) = 3/16 < 1/4 = \text{optimum value.}$$

Consider the stationary policy $\hat{\pi}^\infty$, where $\hat{\pi}_{11} = 1/4$, $\hat{\pi}_{12} = 3/4$, $\hat{\pi}_{21} = \hat{\pi}_{31} = 1$. Since

$$x_{11}(\hat{\pi}) = x_{12}(\hat{\pi}) = x_{32}(\hat{\pi}) = 0,\ x_{21}(\hat{\pi}) = 1/4 \ \text{and}\ x_{32}(\hat{\pi}) = 3/4,$$

we have a feasible solution $\hat{\pi}^\infty$ of problem (4.7.5) with $\beta^T\phi\ (\hat{\pi}^\infty) = 1/4 =$ optimum value. Hence, in this example there exists a stationary optimal solution.

In example 4.7.3, we have $y^*_{ia}/y^*_i \neq \pi^*_{ia}$ for some $a \in A(i)$, $i \in E_{x^*} \cap E_{y^*}$. However, if we can find for the same x^* another y, say $\tilde{y}$, such that the new point $(x^*,\tilde{y})$ is feasible for (4.7.12) and satisfies

(4.7.15) $\quad \tilde{y}_{ia}/\tilde{y}_i = \pi^*_{ia} \qquad a \in A(i),\ i \in E_{x^*} \cap E_{\tilde{y}},$

then, by the same arguments as in theorem 4.7.5, the stationary policy $\tilde{\pi}^\infty$ defined by

$$(4.7.16) \qquad \tilde{\pi}_{ia} = \begin{cases} \pi^*_{ia} & a \in A(i),\ i \notin E_{\tilde{y}} \setminus E_{x^*} \\ \tilde{y}_{ia}/\tilde{y}_i & a \in A(i),\ i \in E_{\tilde{y}} \setminus E_{x^*} \end{cases}$$

is an optimal policy of problem (4.7.5).

The claim that (4.7.15) is satisfied is equivalent to the requirement that

$$\tilde{y}_{ia} = \tilde{y}_i \cdot \pi^*_{ia} \qquad a \in A(i),\ i \in E_{x^*}.$$

Hence, to find a $\tilde{y}$ such that (4.7.15) is satisfied is equivalent to the determination of a feasible solution of the linear system

$$(4.7.17) \qquad \begin{cases} \sum_{i \notin E_{x^*}} \sum_a (\delta_{ij} - p_{iaj}) \cdot \tilde{y}_{ia} + \sum_{i \in E_{x^*}} (\delta_{ij} - p_{ij}(\pi^*)) \cdot \tilde{y}_i = \beta_j - x^*_j,\ j \in E \\ \tilde{y}_{ia} \geq 0,\ a \in A(i),\ i \in E \setminus E_{x^*};\ \tilde{y}_i \geq 0,\ i \in E_{x^*} \end{cases}$$

The feasibility of system (4.7.17) can be checked by the so-called phase I of the simplex method. Hence, we have shown the following result.

<u>THEOREM 4.7.6</u>. *If* $\tilde{y}$ *is a feasible solution of* (4.7.17), *then* $\tilde{\pi}^\infty$ *is an optimal solution of problem* (4.7.5), *where* $\tilde{\pi}^\infty$ *is defined by* (4.7.16).

<u>EXAMPLE 4.7.4</u>. We consider the same model as in example 4.7.3. The optimal solution (x^*, y^*) does not satisfy $y^*_{ia}/y^*_i = \pi^*_{ia}$, $a \in A(i)$, $i \in E_{x^*} \cap E_{y^*}$. Hence, we introduce system (4.7.17):

$$\begin{cases} \tilde{y}_{11} + \tilde{y}_{12} = 4/16 \\ -\tilde{y}_{11} = -1/16 \\ \quad - \tilde{y}_{12} = -3/16 \\ \tilde{y}_{11}, \tilde{y}_{12} \geq 0. \end{cases}$$

This system has a feasible solution, namely $\tilde{y}_{11} = 1/16$, $\tilde{y}_{12} = 3/16$. Hence, the stationary policy $\tilde{\pi}^\infty$, where $\tilde{\pi}_{11} = 1/4$, $\tilde{\pi}_{12} = 3/4$, $\tilde{\pi}_{21} = \tilde{\pi}_{31} = 1$, is an optimal solution of (4.7.5).

<u>THEOREM 4.7.7</u>. *If the linear system* (4.7.17) *is infeasible and if every optimal solution* (x,y) *of problem* (4.7.12) *satisfies* $x = x^*$, *then problem* (4.7.5) *has no optimal solution which belongs to the class of stationary policies.*

<u>PROOF</u>. Suppose that (4.7.5) has an optimal stationary policy, say π^∞. Then $(x(\pi),y(\pi))$ is a feasible solution of problem (4.7.12) and satisfies

$$\sum_i \sum_a r_{ia} x_{ia}(\pi) = \sum_i \sum_a r_{ia} (\beta^T P^*(\pi))_i \cdot \pi_{ia} = \beta^T P^*(\pi) r(\pi)$$

$$= \text{optimum } (4.7.5).$$

Hence, $(x(\pi),y(\pi))$ is an optimal solution of problem (4.7.12). Consequently, $x(\pi) = x^*$. Then, however, $y(\pi)$ is a feasible solution of (4.7.17), which is contradictory to the assumption that (4.7.17) is infeasible. □

<u>REMARK 4.7.8</u>. If the conditions of theorem 4.7.7 hold and consequently no stationary optimal policy exists, then we can use algorithm XIX for the construction of an optimal (Markov) policy.

<u>EXAMPLE 4.7.5</u>. Consider the model of example 4.7.2 with the same constraint $1/4 \le x_{21}(R) \le 1/2$. We have observed that (x^*,y^*) is an optimal solution of problem (4.7.12), where $x^*_{11} = 0$, $x^*_{12} = 0$, $x^*_{21} = 1/2$, $x^*_{31} = 1/2$, $x^*_{32} = 0$ and $y^*_{11} = 0$, $y^*_{12} = 1/4$, $y^*_{32} = 5/16$. It can easily be verified that x^* is unique and that the linear system (4.7.17) i.e.

$$\begin{cases} \tilde{y}_{11} + \tilde{y}_{12} = 4/16 \\ -\tilde{y}_{11} = -5/16 \\ -\tilde{y}_{12} = 1/16 \\ \tilde{y}_{11}, \tilde{y}_{12} \ge 0, \end{cases}$$

is infeasible. Hence, this example has no stationary optimal policy. An optimal Markov policy for this problem was computed in example 4.7.2.

If the linear system (4.7.17) is infeasible and x^* is not unique, then it is possible that problem (4.7.5) has a stationary optimal solution, even if (x^*,y^*) is an extreme point of (4.7.12). Hence, we can compute every optimal extreme point of the linear program (4.7.17) and in each of the obtained points we can perform the analysis described above in order to find a stationary optimal policy.

EXAMPLE 4.7.6. Consider the model described in example 4.7.1 and add the constraint $x_{21}(R) \geq 1/9$. The formulation of problem (4.7.17) is:

$$\begin{array}{llcl}
\textit{maximize} & x_{21} + x_{31} & & \\
\textit{subject to} & x_{11} + x_{12} - x_{22} & = & 0 \\
 & -x_{11} + x_{22} & = & 0 \\
 & -x_{12} & = & 0 \\
 & x_{11} + x_{12} + y_{11} + y_{12} - y_{22} & = & 1/3 \\
 & x_{21} + x_{22} - y_{11} + y_{22} & = & 1/3 \\
 & x_{31} - y_{12} & = & 1/3 \\
 & -x_{21} & \leq & -1/9 \\
 & x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, y_{11}, y_{12}, y_{22} & \geq & 0
\end{array}$$

(x^*,y^*), where $x^*_{11} = 0$, $x^*_{12} = 0$, $x^*_{21} = 1/9$, $x^*_{22} = 0$, $x^*_{31} = 8/9$ and $y^*_{11} = 0$, $y^*_{12} = 5/9$, $y^*_{22} = 2/9$, is an extreme optimal solution, but x^* is not unique. The linear system (4.7.17) is infeasible, namely:

$$\begin{cases}
\tilde{y}_{11} + \tilde{y}_{12} = 1/3 \\
-\tilde{y}_{11} = 2/9 \\
-\tilde{y}_{12} = -5/9 \\
\tilde{y}_{11}, \tilde{y}_{12} \geq 0.
\end{cases}$$

It can easily be verified that $(\hat{x},\hat{y})$, where $\hat{x}_{11} = 0$, $\hat{x}_{12} = 0$, $\hat{x}_{21} = 2/3$, $\hat{x}_{22} = 0$, $\hat{x}_{31} = 1/3$ and $\hat{y}_{11} = 1/3$, $\hat{y}_{12} = 0$, $\hat{y}_{22} = 0$ is also an extreme optimal solution of program (4.7.12). Then theorem 4.7.5 (ii) implies that $\hat{\pi}^\infty$ is an optimal solution of problem (4.7.5), where $\hat{\pi}_{11} = \hat{\pi}_{21} = \hat{\pi}_{31} = 1$.

THEOREM 4.7.8. *Let* (x^*,y^*) *be an optimal solution of problem* (4.7.12). *Consider the nonlinear system*

$$(4.7.18) \quad \begin{cases}
\sum_i \sum_a r_{ia} x_{ia} = \sum_i \sum_a r_{ia} x^*_{ia} \\
\sum_i \sum_a q_{iak} x_{ia} \leq b_k \quad 1 \leq k \leq m \\
\sum_a x_{ja} + \sum_{i \notin E_x} \sum_a (\delta_{ij} - p_{iaj}) y_{ia} + \sum_{i \in E_x} (\delta_{ij} - \sum_a p_{iaj} \, x_{ia} / \sum_a x_{ia}) y_i = \beta_j, \; j \in E \\
\sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = 0, \quad j \in E \\
x_{ia} \geq 0, a \in A(i),\ i \in E;\ y_{ia} \geq 0, a \in A(i),\ i \in E \backslash E_x;\ y_i \geq 0,\ i \in E_x
\end{cases}$$

(i) *If* $(\tilde{x},\tilde{y})$ *is a feasible solution of* (4.7.18), *then the policy* $\tilde{\pi}^{\infty}$ *defined by*

$$\tilde{\pi}_{ia} := \begin{cases} \tilde{x}_{ia}/\sum_a \tilde{x}_{ia} & a \in A(i),\ i \in E_{\tilde{x}} \\ \tilde{y}_{ia}/\sum_a \tilde{y}_{ia} & a \in A(i),\ i \in E_{\tilde{y}} \backslash E_{\tilde{x}} \\ \textit{arbitrarily} & \textit{elsewhere} \end{cases}$$

is an optimal solution of problem (4.7.5).

(ii) *If* (4.7.18) *is infeasible, then problem* (4.7.5) *has no stationary optimal policy.*

PROOF.

(i) Theorem 4.7.6 implies that $\tilde{x} = x(\tilde{\pi})$. Hence, $\tilde{\pi}^{\infty}$ is a feasible solution of problem (4.7.5) with as value of the objective function

$$\phi(\beta,(\tilde{\pi})^{\infty}) = \sum_j \sum_a r_{ja} x_{ja}(\tilde{\pi}) = \sum_j \sum_a r_{ja} \tilde{x}_{ja} = \sum_j \sum_a r_{ja} x^*_{ja}$$
$$= \text{optimum } (4.7.12).$$

Hence, $\tilde{\pi}^{\infty}$ is an optimal solution of problem (4.7.5).

(ii) Suppose that $\hat{\pi}^{\infty}$ is a stationary optimal solution of problem (4.7.5). Then $(\hat{x},\hat{y})$ such that

$$\hat{x}_{ia} = x_{ia}(\hat{\pi}) \qquad a \in A(i),\ i \in E,$$
$$\hat{y}_{ia} = y_{ia}(\hat{\pi}) \qquad a \in A(i),\ i \in E \backslash E_{\hat{x}}$$
$$\hat{y}_i = \sum_a y_{ia}(\hat{\pi}) \qquad i \in E_{\hat{x}},$$

where $x(\hat{\pi})$ and $y(\hat{\pi})$ are defined by (4.3.2), is a feasible solution of (4.7.18). This implies a contradiction. □

REMARK 4.7.9. In general, it is a difficult problem to find a feasible solution of problem (4.7.18). However, computational results indicate that it is mostly not necessary to solve problem (4.7.18) in order to obtain a stationary optimal solution of (4.7.5), if one exists. Below we present an algorithm for the construction of a stationary policy. This algorithm is based on the theorems 4.7.4-4.7.7. We have tested 400 problems and the algorithm has always given an optimal stationary policy, if one exists. Furthermore, if the stationary policy is nonoptimal, then this policy may be considered as an approximate solution of problem (4.7.5). For this approximation we know the deviation to the optimal value and also we know which constraints are violated.

ALGORITHM XX for the construction of a stationary policy in a constrained AMD-model (multichain case).

step 1: Use the simplex method to compute an optimal solution (x^*,y^*) of the linear programming problem

$$(4.7.19)\qquad \max\left\{ \sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{lr} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} & = 0\ ,\ j \in E \\ \sum_a x_{ja} \qquad + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} & = \beta_j,\ j \in E \\ \sum_i\sum_a q_{iak}x_{ia} & \le b_k,\ 1\le k\le m \\ x_{ia},y_{ia} & \ge 0,\ a \in A(i)\ ,\ i \in E \end{array} \right\}$$

(if this linear program is infeasible, then the constrained Markov decision problem (4.7.5) is also infeasible).

step 2: Determine the stationary policy $(\pi^*)^\infty$ such that

$$\pi^*_{ia} := \begin{cases} x^*_{ia}/\sum_a x^*_{ia} & a \in A(i),\ i \in E_{x^*}. \\ y^*_{ia}/\sum_a y^*_{ia} & a \in A(i),\ i \in E_{y^*}\backslash E_{x^*}. \\ \text{arbitrarily} & \text{elsewhere.} \end{cases}$$

step 3a: If $y^*_{ia}/\Sigma_a y^*_{ia} = \pi^*_{ia}$ for all $a \in A(i)$, $i \in E_{x^*}\cap E_{y^*}$, then $(\pi^*)^\infty$ is an optimal solution of problem (4.7.5) (STOP).

step 3b: Go to step 4a or to step 4b (comment is given in remark 4.7.10).

step 4a: Compute an optimal solution $(\tilde{y},\tilde{z})$ of the linear program

$$\min\left\{ \sum_j z_j \;\middle|\; \begin{array}{l} \sum_{i\in E\backslash E_{x^*}} \sum_a (\delta_{ij}-p_{iaj})y_{ia} + \sum_{i\in E_{x^*}} (\delta_{ij}-p_{ij}(\pi^*))y_i+z_j = \beta_j-x^*_j,\ j \in E \\ y_{ia} \ge 0,\ a \in A(i),\ i \in E\backslash E_{x^*};\ y_i \ge 0,\ i \in E_{x^*};\ z_j \ge 0,\quad j \in E \end{array} \right\}.$$

If $\Sigma_j\tilde{z}_j = 0$, then $\tilde{\pi}^\infty$, where

$$\tilde{\pi}_{ia} := \begin{cases} \tilde{y}_{ia}/\sum_a\tilde{y}_{ia} & a \in A(i),\ i \in E_{\tilde{y}}\backslash E_{x^*} \\ \pi^*_{ia} & \text{elsewhere} \end{cases}$$

is an optimal solution of problem (4.7.5) (STOP).

Otherwise, go to step 5.

step 4b: Compute $x_{ia}(\pi^*) := [\beta^T P^*(\pi^*)]_i \cdot \pi^*_{ia}$, $a \in A(i)$, $i \in E$ (the computation of the stationary matrix $P^*(\pi^*)$ can be performed by algorithm III).
If $x^* = x(\pi^*)$, then $(\pi^*)^\infty$ is an optimal solution of problem (4.7.5) (STOP).

Otherwise: if $\Sigma_i \Sigma_a q_{iak} x_{ia}(\pi^*) \le b_k$ $k = 1,2,\dots,m$ and
$\Sigma_i \Sigma_a r_{ia} x_{ia}(\pi^*) = \Sigma_i \Sigma_a r_{ia} x^*_{ia}$, then $(\pi^*)^\infty$ is an optimal solution of problem (4.7.5) (STOP).

Otherwise, go to step 5.

step 5: Put $(\pi^*)^\infty$ on the list L_1 of stationary policies and x^* on the list L_2 of solutions that have been analysed.

step 6: If there exists an extreme optimal solution $(\hat{x},\hat{y})$ of program (4.7.19) such that $\hat{x} \notin L_2$, then:

$$(x^*,y^*) := (\hat{x},\hat{y}) \text{ and go to step 2}$$

(the determination of all extreme optimal solutions can be performed by algorithm I).

Otherwise: go to step 7.

step 7: Any stationary policy $(\pi^*)^\infty$ from the list L_1 may be viewed as an approximate solution of problem (4.7.5).

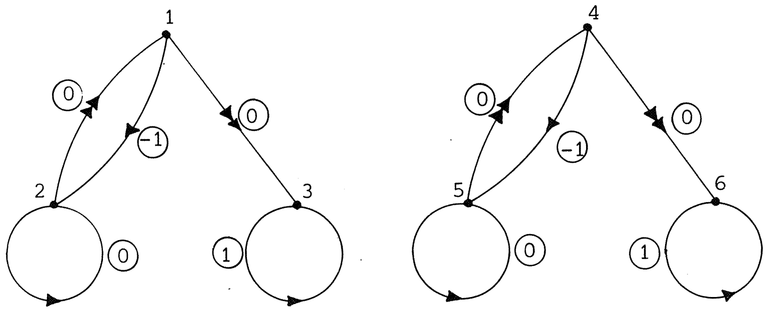
REMARK 4.7.10. If the condition $y^*_{ia}/\Sigma_a y^*_{ia} = \pi^*_{ia}$, $a \in A(i)$, $i \in E_{x^*} \cap E_{y^*}$ is not satisfied in step 3a, then we have to decide for a continuation in step 4a or step 4b. When $|E_{x^*}|$ is small with respect to $|E|$, then the linear program of step 4a has many variables. In this case we propose to perform step 4b. When $|E_{x^*}|$ is (nearly) equal to $|E|$, then we propose to continue in step 4a.

REMARK 4.7.11. Suppose that there exists an optimal stationary policy π^∞ such that $x(\pi)$ is an extreme point of $\tilde{X}$, where

$$\tilde{X} = \{x \in X \mid (x,y) \text{ is an optimal solution of problem (4.7.5) for some } y\}.$$

Then, algorithm XX will find an optimal stationary policy. Unfortunately, it is possible that $x(\pi)$ is not an extreme point of $\tilde{X}$ for every optimal stationary policy π^∞. In example 4.7.7 we show this phenomenon.

EXAMPLE 4.7.7. Consider the model drawn in figure 4.7.5, with the constraints $x_{31}(R) \le 5/12$, $x_{61}(R) \le 5/12$, $x_{31}(R)+x_{61}(R) \le 2/3$. It can easily be verified that the set of optimal x-vectors is given by



$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 1/6$

Figure 4.7.5

$$\left\{ x \,\middle|\, \begin{matrix} x_{11} = x_{12} = x_{22} = x_{41} = x_{42} = x_{52} = 0;\; x_{21} + x_{31} = x_{51} + x_{61} = 1/2; \\ x_{31} \le 5/12,\; x_{61} \le 5/12;\; x_{31} + x_{61} = 2/3;\; x_{21}, x_{31}, x_{51}, x_{61} \ge 0; \end{matrix} \right\} = \tilde{X}.$$

By the dependency of x_{21} and x_{51} on x_{31} and x_{61} respectively, we can draw the set $\tilde{X}$ in the 2-dimensional space with the coordinates x_{31} and x_{61} (see figure 4.7.6). Consider the policy f^∞, where $f(1) = 2$, $f(2) = 1$, $f(3) = 1$, $f(4) = 2$, $f(5) = 1$, $f(6) = 1$. Then, $x(f)$ satisfies $x_{11}(f) = x_{12}(f) = x_{22}(f) = x_{41}(f) = x_{42}(f) = x_{52}(f) = 0$, $x_{21}(f) = x_{51}(f) = 1/6$, $x_{31}(f) = x_{61}(f) = 1/3$. Hence, f^∞ is an optimal solution of problem (4.7.5), but $x(f)$ is not an extreme point of $\tilde{X}$. Moreover, it can be verified that $L(S) \cap \tilde{X} = \{x(f)\}$.

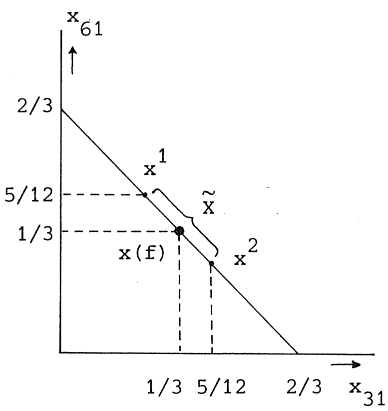


Figure 4.7.6

REMARK 4.7.12. If X = L(S), then algorithm XX will find a stationary optimal solution as soon as step 4a is visited. In theorem 4.7.9, we present a sufficient condition for the equality of the sets X and L(S). This condition is always satisfied in the unichain case as will be shown in section 4.7.5.

LEMMA 4.7.1. *For every triple* (j,a,R), *where* $j \in E$, $a \in A(j)$ and $R \in C_1$, *we have*

$$x_{ja}(R) = \lim_{\alpha\uparrow 1}(1-\alpha)\cdot\sum_{t=1}^{\infty}\alpha^{t-1}\cdot\sum_i\beta_i\cdot\mathbb{P}_R(X_t=j, Y_t=a \mid X_1=i).$$

PROOF. For the proof of this lemma we use the same arguments as in HORDIJK [1971]. Let $R \in C_1$ and suppose that $x(R) = \lim_{T\to\infty} x^T(R)$. Take a fixed pair (j,a), where $j \in E$, $a \in A(j)$. Then,

$$x_{ja}(R) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} w_t,$$

where

$$w_t := \sum_i\beta_i\cdot\mathbb{P}_R(X_t=j, Y_t=a \mid X_1=i), \quad t \in \mathbb{N}.$$

Since $|w_t|$ is bounded by 1 for all t, the power series $\Sigma_{t=1}^{\infty} w_t\alpha^{t-1}$ has radius of convergence at least 1. The series $\Sigma_{t=1}^{\infty}\alpha^{t-1}$ has radius of convergence 1. Hence, for $\alpha \in [0,1)$, we may write

$$(1-\alpha)^{-1}\cdot\sum_{t=1}^{\infty} w_t\alpha^{t-1} = (\sum_{t=1}^{\infty}\alpha^{t-1})\cdot(\sum_{t=1}^{\infty} w_t\alpha^{t-1}) = \sum_{t=1}^{\infty}(\sum_{s=1}^{t} w_s)\alpha^{t-1}.$$

From $(1-\alpha)^{-2} = \Sigma_{t=1}^{\infty} t\alpha^{t-1}$ for $0 \le \alpha < 1$, we obtain

$$x_{ja}(R) - (1-\alpha)\cdot\sum_{t=1}^{\infty}\alpha^{t-1}\cdot\sum_i\beta_i\cdot\mathbb{P}_R(X_t=j, Y_t=a \mid X_1=i) =$$

$$(1-\alpha)^2\sum_{t=1}^{\infty}\{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\}t\alpha^{t-1}.$$

Choose $\varepsilon > 0$ arbitrarily small. Since $x_{ja}(R) = \lim_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^{T} w_t$, there exists an integer T_ε such that

$$|x_{ja}(R) - \frac{1}{T}\sum_{t=1}^{T} w_t| \le \frac{1}{2}\varepsilon \quad \text{for all } T > T_\varepsilon.$$

Hence,

$$|(1-\alpha)^2\sum_{t=1}^{T_\varepsilon}\{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\}t\alpha^{t-1}| \le$$

$$(1-\alpha)^2 M\cdot\sum_{t=1}^{T_\varepsilon} T_\varepsilon\alpha^{t-1} \le \frac{1}{2}\varepsilon \quad \text{for } \alpha \text{ sufficiently near to 1 and}$$

$$M \ge \max_{1\le t\le T_\varepsilon} |x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s|,$$

and

$$|(1-\alpha)^2 \sum_{t=T_\varepsilon+1}^{\infty}\{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\} t\alpha^{t-1}| \le$$

$$(1-\alpha)^2 \sum_{t=T_\varepsilon+1}^{\infty} \frac{\varepsilon}{2} t\alpha^{t-1} \le \frac{\varepsilon}{2}(1-\alpha)^2 \sum_{t=1}^{\infty} t\alpha^{t-1} = \frac{1}{2}\varepsilon.$$

Hence

$$x_{ja}(R) = \lim_{\alpha\uparrow 1}(1-\alpha)\cdot\sum_{t=1}^{\infty}\alpha^{t-1}\cdot\sum_i \beta_i \cdot \mathbb{P}_R(X_t = j, Y_t = a \mid X_1 = i),$$

completing the proof. □

THEOREM 4.7.9. *If* $x(\pi)$ *is continuous in* π, *then* $X = L(S)$.

PROOF. Theorem 4.7.1 implies that it is sufficient to show that $L(C) \subset L(S)$. Take any $x(R) \in L(C)$. From theorem 3.4.8 it follows that for any $\alpha \in [0,1)$ there exists a stationary policy π^α such that $x^\alpha(R) = x^\alpha(\pi^\alpha)$, where $x^\alpha(\cdot)$ is defined by

$$x^\alpha_{ja}(\tilde{R}) := \sum_{t=1}^{\infty}\alpha^{t-1}\cdot\sum_i \beta_i \cdot \mathbb{P}_{\tilde{R}}(X_t = j, Y_t = a \mid X_1 = i) \quad j\in E,\ a\in A(j), \tilde{R}\in C.$$

Choose a fixed pair (j,a), $j \in E$, $a \in A(j)$. Introduce a reward function by

$$r_{ib} := \begin{cases} 1 & i = j \quad b = a \\ 0 & \text{elsewhere} \end{cases} \qquad b \in A(i),\ i \in E.$$

Then,

$$\beta^T v^\alpha(\pi^\alpha) = x^\alpha_{ja}(\pi^\alpha) \quad \text{and} \quad \beta^T\phi(\pi^\alpha) = x_{ja}(\pi^\alpha), \quad a \in [0,1).$$

Hence, we can write by lemma 4.7.1

$$(4.7.20) \qquad x_{ja}(R) = \lim_{\alpha\uparrow 1}(1-\alpha)x^\alpha_{ja}(R) = \lim_{\alpha\uparrow 1}(1-\alpha)x^\alpha_{ja}(\pi^\alpha)$$

$$= \lim_{\alpha\uparrow 1}(1-\alpha)\cdot\beta^T v^\alpha(\pi^\alpha).$$

Consider a sequence $\{\alpha_k,\ k = 1,2,\ldots\}$ such that $\alpha_k \uparrow 1$ and $\pi^{\alpha_k} \to \pi$. Since for any $i \in E$ the sequence $\{(1-\alpha_k)v_i^{\alpha_k}(\pi^{\alpha_k}),\ k = 1,2,\ldots\}$ is dominated by the sequence $\{(1-\alpha_k)v_i^{\alpha_k},\ k = 1,2,\ldots\}$ and since $\lim_{k\to\infty}(1-\alpha_k)v_i^{\alpha_k} = \phi_i$ (cf. (2.5.7)), there exists a limit point, say x, of the sequence of vectors $\{(1-\alpha_k)v^{\alpha_k}(\pi^{\alpha_k}),\ k = 1,2,\ldots\}$. Therefore, we may assume that

$$(4.7.21) \qquad x_i = \lim_{k\to\infty}(1-\alpha_k)v_i^{\alpha_k}(\pi^{\alpha_k}), \quad i \in E.$$

From (4.7.20) and (4.7.21) it follows that

$$\beta^T x = \sum_i \beta_i \cdot \lim_{k\to\infty}(1-\alpha_k) v_i^{\alpha_k}(\pi^{\alpha_k}) \tag{4.7.22}$$

$$= \lim_{k\to\infty}(1-\alpha_k)\beta^T v^{\alpha_k}(\pi^{\alpha_k}) = x_{ja}(R).$$

The continuity of $x(\tilde{\pi})$ as function of $\tilde{\pi}$ implies

$$x_{ja}(\pi) = \lim_{k\to\infty} x_{ja}(\pi^{\alpha_k})$$

$$= \lim_{k\to\infty}(1-\alpha_k)\sum_{t=1}^{\infty}\alpha_k^{t-1}\beta^T P^*(\pi^{\alpha_k}) r(\pi^{\alpha_k})$$

$$= \lim_{k\to\infty}\beta^T P^*(\pi^{\alpha_k})(1-\alpha_k)\sum_{t=1}^{\infty}\alpha_k^{t-1}P^{t-1}(\pi^{\alpha_k}) r(\pi^{\alpha_k})$$

$$= \lim_{k\to\infty}(x(\pi^{\alpha_k}))^T(1-\alpha_k) v^{\alpha_k}(\pi^{\alpha_k})$$

$$= (x(\pi))^T x = \beta^T P^*(\pi) x. \tag{4.7.23}$$

Since for every $\alpha \in [0,1)$ $v^{\alpha}(\pi^{\alpha}) = r(\pi^{\alpha}) + \alpha P(\pi^{\alpha}) v^{\alpha}(\pi^{\alpha})$, it follows from (4.7.21) that

$$x = P(\pi) x.$$

Consequently,

$$x = P^*(\pi) x. \tag{4.7.24}$$

Then the relations (4.7.22), (4.7.23) and (4.7.24) imply that

$$x_{ja}(R) = \beta^T x = \beta^T P^*(\pi) x = x_{ja}(\pi).$$

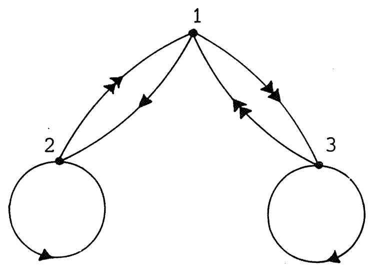
Since π is independent of the choice of the pair (j,a), we have proved that $x(R) \in L(S)$. This yields the theorem. □

<u>REMARK 4.7.13</u>. It will be shown in section 7.4.5 that unichainedness implies continuity of $x(\pi)$, and consequently $X = L(S)$. If we relax the unichainedness to communicating (i.e. for each pair $i,j \in E$ there exists a policy $f^{\infty} \in C_D$ and an integer $t \in \mathbb{N}$ such that $\mathbb{P}_{f^{\infty}}(X_t = j \mid X_1 = i) > 0$), then $X \neq L(S)$, in general. Below we give an example.

EXAMPLE 4.7.8. Consider the model corresponding to figure 4.7.7. This model is obviously communicating. It can easily be verified that

$$X = \left\{ x \;\middle|\; \begin{array}{l} x_{11}=x_{22};\ x_{12}=x_{32};\ x_{11}+x_{12}+x_{21}+x_{22}+ \\ x_{31}+x_{32}=1;\ x_{11},x_{12},x_{21},x_{22},x_{31},x_{32}\geq 0 \end{array} \right\}.$$

Take $\tilde{x}$ such that $\tilde{x}_{11}=\tilde{x}_{22}=\tilde{x}_{12}=\tilde{x}_{32}=0$, $\tilde{x}_{21}=1/4$, $\tilde{x}_{31}=3/4$. Suppose that $\tilde{x}=x(\tilde{\pi})$ for some stationary policy $\tilde{\pi}^{\infty}$. From $\tilde{x}_{21}>0$, $\tilde{x}_{22}=0$ it follows that $\tilde{\pi}_{21}=1$. Hence, state 2 is absorbing in the Markov chain induced by $P(\tilde{\pi})$. Consequently, $x_{21}(\tilde{\pi}) \geq \beta_2 = 1/3 > 1/4 = \tilde{x}_{21}$, implying a contradiction. Therefore, in this model $X \neq L(S)$.



$\beta_1=\beta_2=\beta_3=1/3$

Figure 4.7.7

We close this section with the presentation of some numerical results obtained by algorithm XX. We have solved 400 test problems. These problems can be divided in 8 classes of 50 problems as indicated in table 4.7.1 (ℓ = the number of actions in each state; m = the number of constraints)

	A	B	C	D	E	F	G	H
ℓ	2	2	2	2	2	4	4	4
m	1	2	3	4	5	1	3	5

Table 4.7.1

All problems have been generated as follows:

(i) the number of states is 10, i.e. $E = \{1,2,...,10\}$

(ii) for each pair (i,a), where $i \in E$ and $a \in A(i)$, the transition probabilities are such that $p_{iaj} \neq 0$ for exactly one j which is randomly chosen from E.

(iii) the reward r_{ia} is a random choice from $\{0,1,...,10\}$, $a \in A(i)$, $i \in E$.

(iv) the coefficients q_{iak} are randomly chosen from $\{-10,-9,...,+10\}$ $i \in E$, $a \in A(i)$, $k = 1,2,...,m$.

(v) $b_k = 0$ $k = 1,2,...,m$.

The numerical results are summarized in table 4.7.2 and give rise to the following statements:

1. 8% of the problems is infeasible and in 16% the algorithm does not find a stationary optimal policy. We have analysed that all these problems do not have stationary optimal policies. Hence, for every problem which has a stationary optimal policy algorithm XX gives one.
2. 70% of the 306 problems for which a stationary optimal policy was found, this policy was found in step 4 of the algorithm.
3. For only 9 problems the stationary optimal policy was obtained by the analysis of more than one extreme optimal solution. Hence, in 97% of the problems for which a stationary optimal policy was found, this policy was obtained from the first analysed optimal solution of program (4.7.19).

Class	k	m	Total number of problems	Infeasibility (step 1)	Policy obtained from the first analysed LP-solution		Policy obtained from second, third etc. LP-solution	No stationary optimal policy
					Termination in step 3	Termination in step 4		
A	2	1	50	1	20	22	2	5
B	2	2	50	2	13	25	1	9
C	2	3	50	4	6	29	-	11
D	2	4	50	11	5	21	2	11
E	2	5	50	13	3	25	-	9
F	4	1	50	-	22	26	-	2
G	4	3	50	-	12	29	4	5
H	4	5	50	-	4	35	-	11
Total			400	31	85	212	9	63

4.7.5. COMPUTATION OF A STATIONARY OPTIMAL POLICY (UNICHAIN CASE)

Throughout this section we use the following assumption.

<u>ASSUMPTION 4.7.1</u>. *For any pure and stationary policy* f^{∞}, *the Markov chain induced by* $P(f)$ *has one ergodic set plus a (perhaps empty) set of transient states.*

THEOREM 4.7.10. $X = L(S)$.

PROOF. By theorem 4.7.9 it is sufficient to show that $x(\pi)$ is continuous in π. Let $\lim_{k\to\infty}\pi(k) = \pi(0)$, where $\pi(k)^\infty \in C_S$, $k \in \mathbb{N}_0$. By lemma 4.6.1 and assumption 4.7.1, the Markov chain under $P(\pi(k))$ has at most one ergodic set for every $k \in \mathbb{N}_0$. Theorem 2.3.3 implies that $x(\pi(k))$ is the unique solution of the linear system

$$(4.7.25)\qquad \begin{cases} \sum_i (\delta_{ij} - p_{ij}(\pi(k)))x_i = 0, & j \in E \\ \sum_i x_i = 1 \end{cases}$$

for every $k \in \mathbb{N}_0$. Since $\pi(k) \to \pi(0)$ for $k \to \infty$, we also have $P(\pi(k)) \to P(\pi(0))$ for $k \to \infty$. Consequently, any limit point of $\{x(\pi(k)),\ k = 1,2,\ldots\}$ is a solution of (4.7.25) with $k = 0$. Hence, $x(\pi(0)) = \lim_{k\to\infty} x(\pi(k))$, i.e. $x(\pi)$ is continuous in π. □

ALGORITHM XXI for the construction of a stationary optimal policy in a constrained AMD-model (unichain case).

step 1: Use the simplex method to determine an optimal solution x^* of the linear programming problem

$$(4.7.26)\qquad \max\left\{ \sum_i\sum_a r_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j \in E \\ \sum_i\sum_a x_{ia} = 1 & \\ \sum_i\sum_a q_{iak}x_{ia} \le b_k, & k = 1,2,\ldots,m \\ x_{ia} \ge 0,\ a \in A(i), & i \in E \end{array} \right\}$$

(if this linear program is infeasible, then the constrained Markov decision problem is also infeasible).

step 2: Take $(\pi^*)^\infty$ such that

$$\pi^*_{ia} := \begin{cases} x^*_{ia}/\sum_a x^*_{ia} & a \in A(i),\ i \in E_{x^*} \\ \text{arbitrarily} & \text{elsewhere.} \end{cases}$$

THEOREM 4.7.11. *The policy* $(\pi^*)^\infty$ *obtained by algorithm XXI is an optimal solution of problem* (4.7.5).

PROOF. From the definition of π^* it follows that

$$(4.7.27)\qquad \begin{cases} \sum_i (\delta_{ij} - p_{ij}(\pi^*))(\sum_a x^*_{ia}) = 0, & j \in E \\ \sum_i (\sum_a x^*_{ia}) = 1. \end{cases}$$

Similarly as in the proof of theorem 4.7.10, we can show that (4.7.27) implies that $x^* = x(\pi^*)$. Hence, $(\pi^*)^\infty$ is a feasible solution of problem (4.7.5). Moreover,

$$\phi(\beta,(\pi^*)^\infty) = \sum_i \sum_a r_{ia} x^*_{ia} = \text{optimum } (4.7.26).$$

From theorem 4.7.10 it follows that there exists a stationary optimal solution of problem (4.7.5), say $\tilde{\pi}^\infty$. Let $\tilde{x} = x(\tilde{\pi})$. Then, $\tilde{x}$ is a feasible solution of program (4.7.26) and consequently,

$$\text{optimum } (4.7.5) = \phi(\beta,(\tilde{\pi})^\infty)) = \Sigma_i \Sigma_a r_{ia} \tilde{x}^*_{ia} \le \sum_i \sum_a r_{ia} \tilde{x}^*_{ia} = \phi(\beta,(\pi^*)^\infty).$$

Hence, $(\pi^*)^\infty$ is an optimal solution of problem (4.7.5). □

CHAPTER 5

BIAS OPTIMALITY

5.1. INTRODUCTION AND SUMMARY

The use of the expected average reward as utility function is sometimes unsatisfactory. For any stationary policy π^∞, rewards that are earned when the process is in a state which is transient under $P(\pi)$ do not influence the outcome of the average reward $\phi(\pi^\infty)$. Therefore, the average reward criterion is in some sense too little selective. The concept of *bias optimality* is a more selective criterion. This criterion was introduced by BLACKWELL [1962] (actually Blackwell used the term "*nearly optimal*"). A first algorithm to compute a bias optimal policy was presented in VEINOTT [1966]. DENARDO [1970a] has refined this method to a three-step procedure which can be executed by linear programming as well as by policy improvement.

In chapter 2 we have presented the definition of a bias optimal policy: $R^* \in C$ is said to be a *bias optimal policy* if

$$(5.1.1.) \qquad \lim_{\alpha \uparrow 1} \{v_i^\alpha(R^*) - v_i^\alpha\} = 0, \qquad i \in E.$$

Corollary 2.5.2 implies the existence of a pure and stationary bias optimal policy.

In section 5.2 we present some equivalent statements for the concept of bias optimality. One of these statements gives rise to an algorithm for the computation of a bias optimal policy.

Then, in section 5.3, we present some theorems which lead to another algorithm. This algorithm is a modification of the algorithm presented in

DENARDO [1970a]. The algorithm can be divided into three parts and in each part a linear program has to be solved. For the determination of an average optimal policy - which has to be performed in the parts 1 and 2 for two different AMD-models - we use the results of chapter 4. Furthermore, we show that Denardo's search procedure of the third part can be cancelled and that a bias optimal policy can be obtained directly from the solution of the third linear program. Some of the material of this section can also be found in KALLENBERG [1981b].

We close this chapter by section 5.4 in which we discuss the weak unichain case, the completely ergodic case and the unichain case. For these models the algorithm can be simplified.

5.2. SOME THEOREMS

We assume in this chapter that $\Sigma_j p_{iaj} = 1$ for every pair (i,a), $a \in A(i)$, $i \in E$. If this assumption is not satisfied, then we can change the model into the extended model, with state space $E \cup \{0\}$, as described in definition 3.2.2. From definition 3.2.2 and the analysis on page 30 it follows that $v_i^\alpha(R) = \tilde{v}_i^\alpha(R)$ $i \neq 0$, for every $R \in C$ and all $\alpha \in [0,1)$, where $\tilde{v}^\alpha(R)$ denotes the expected discounted reward in the extended model.

<u>LEMMA 5.2.1</u>. *For any policy* R *we have*

$$\hat{\phi}_i(R) \geq \limsup_{\alpha \uparrow 1} (1-\alpha) v_i^\alpha(R), \quad i \in E.$$

<u>PROOF</u>. The proof is similar to the proof of lemma 4.7.1. Take any $R \in C$ and $i \in E$. Let $w_t := \Sigma_j \Sigma_a \mathbb{P}_R(X_t=j, Y_t = a \mid X_1 = i) \cdot r_{ja}$, $t \in \mathbb{N}$. Then, for $\alpha \in [0,1)$ we may write

$$(1-\alpha)^{-1} v_i^\alpha(R) = \sum_{t=1}^\infty (\sum_{s=1}^t w_s) \alpha^{t-1} \text{ and } \hat{\phi}_i(R) = (1-\alpha)^2 \hat{\phi}_i(R) \sum_{t=1}^\infty t \, \alpha^{t-1}.$$

Hence, we have

$$\hat{\phi}_i(R) - (1-\alpha) v_i^\alpha(R) = (1-\alpha)^2 \sum_{t=1}^\infty \{\hat{\phi}_i(R) - \frac{1}{t} \sum_{s=1}^t w_s\} \, t \, \alpha^{t-1}.$$

Choose $\varepsilon > 0$ arbitrarily. Since $\hat{\phi}_i(R) = \limsup_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T w_t$, there exists an integer T_ε such that

$$\hat{\phi}_i(R) > \frac{1}{T} \sum_{t=1}^T w_t - \frac{1}{2}\varepsilon \quad \text{for every } T \geq T_\varepsilon.$$

Therefore,

$$(1-\alpha)^2 \sum_{t=1}^{T_\varepsilon -1} \{\hat{\phi}_i(R) - \tfrac{1}{t} \sum_{s=1}^{t} w_s\}\, t\,\alpha^{t-1} \geq$$

$$(1-\alpha)^2 \min_{1\le t\le T_\varepsilon -1} \{\hat{\phi}_i(R) - \tfrac{1}{t} \sum_{s=1}^{t} w_s\} \sum_{t=1}^{T_\varepsilon -1} t\,\alpha^{t-1} \geq$$

$$(1-\alpha)^2 (-M) \sum_{t=1}^{T_\varepsilon -1} T_\varepsilon\, \alpha^{t-1} > \tfrac{1}{2}\varepsilon \text{ for } \alpha \text{ sufficiently near enough to 1,}$$

where $M > 0$ satisfies $\min_{1\le t\le T_\varepsilon -1} \{\hat{\phi}_i(R) - \frac{1}{t}\Sigma_{s=1}^{t} w_s\} \geq -M$.
Furthermore,

$$(1-\alpha)^2 \sum_{t=T_\varepsilon}^{\infty} \{\hat{\phi}_i(R) - \tfrac{1}{t} \sum_{s=1}^{t} w_s\}\, t\,\alpha^{t-1} \geq (1-\alpha)^2 \sum_{t=T_\varepsilon}^{\infty} (-\tfrac{1}{2}\varepsilon)\, t\,\alpha^{t-1} \geq -\tfrac{1}{2}\varepsilon.$$

Then, for α sufficiently near enough to 1, we have

$$\hat{\phi}_i(R) \geq (1-\alpha)v_i^\alpha(R) - \varepsilon,$$

implying that $\hat{\phi}_i(R) \geq \limsup_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha(R)$. □

<u>THEOREM 5.2.1</u>. *If* R *is bias optimal, then* R *is also an average optimal policy; if* R *is average optimal, then* R *is not a bias optimal policy, in general.*

<u>PROOF</u>. Let f be a pure and stationary Blackwell optimal policy. Then, $v^\alpha(f^\infty) = v^\alpha$ for α sufficiently near to 1, and, by theorem 2.5.4, $\phi(f^\infty) = \phi$. Hence, using (2.5.7), we obtain

(5.2.1) $\quad \lim_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha = \lim_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha(f^\infty) = \phi_i(f^\infty) = \phi_i,\ i \in E.$

Since R is a bias optimal policy, we have $\lim_{\alpha\uparrow 1}\{v^\alpha(R) - v_i^\alpha\} = 0$, $i \in E$. Therefore, certainly $\lim_{\alpha\uparrow 1}(1-\alpha)\{v_i^\alpha(R) - v_i^\alpha\} = 0$. The existence of $\lim_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha$, $i \in E$, implies that $\lim_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha(R) = \phi_i$, $i \in E$. From lemma 5.2.1 it follows that

$$\hat{\phi}_i(R) \geq \limsup_{\alpha\uparrow 1}(1-\alpha)v_i^\alpha(R) = \phi_i,\ i \in E.$$

Then, the results of section 4.2 imply that R is average optimal.

The policy $f_*^\infty \in C_D$ such that $f_*(1) = f_*(2) = 1$ is an average optimal policy for the model of figure 5.2.1. Since $v_1^\alpha(f_*^\infty) = 0$ for all $\alpha \in [0,1)$ and $v_1^\alpha = 1$ for all $\alpha \in [0,1)$, f_*^∞ is not a bias optimal policy. □

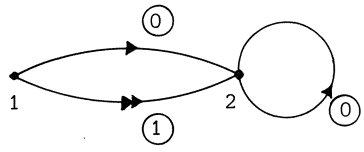


Figure 5.2.1

THEOREM 5.2.2. *Let* $f_*^\infty \in C_D$. *Then, the following four statements are equivalent*:

(i) f_*^∞ *is bias optimal.*

(ii) $\lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty)\} \geq 0$ *for each* $f^\infty \in C_D$.

(iii) $u_i(f_*^\infty) = \max\{u_i(f^\infty) = \phi_i\}$, $i \in E$, and $\phi(f_*^\infty) = \phi$.

(iv) $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T}\sum_{s=1}^{t}\{v^s(f_*^\infty) - v^s(f^\infty)\} \geq 0$ *for each* $f^\infty \in C_D$.

PROOF.

(i) ⇒ (ii): Suppose that f_*^∞ is a bias optimal policy. Take any $f^\infty \in C_D$. Since $v^\alpha(f^\infty) \leq v^\alpha$ for all $\alpha \in [0,1)$, we obtain

$$\lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty)\} \geq \lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0.$$

(ii) ⇒ (iii): From (2.5.7), it follows that

$$\lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty)\} = \lim_{\alpha\uparrow 1}\{\frac{\phi(f_*^\infty)-\phi(f^\infty)}{1-\alpha} + u(f_*^\infty) - u(f^\infty)\}.$$

Consequently, $\lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty)\} \geq 0$ implies that

$$\phi(f_*^\infty) \geq \phi(f^\infty), \text{ and } u_i(f_*^\infty) \geq u_i(f^\infty) \text{ if } f^\infty \text{ satisfies } \phi_i(f^\infty) = \phi_i(f_*^\infty).$$

Hence, $\phi(f_*^\infty) = \max_{f^\infty \in C_D} \phi(f^\infty) = \phi$. Then, we can write

$$u_i(f_*^\infty) = \max\{u_i(f^\infty) \mid \phi_i(f^\infty) = \phi_i\},\ i \in E.$$

(iii) ⇒ (iv): Let f_*^∞ be such that $u_i(f_*^\infty) = \max\{u_i(f^\infty) \mid \phi_i(f^\infty) = \phi_i\}$, $i \in E$, and $\phi(f_*^\infty) = \phi$.

We have for any $f^\infty \in C_D$

$$\sum_{s=1}^{t}\{v^s(f_*^\infty) - v^s(f^\infty)\} = \sum_{s=1}^{t}\{P^{s-1}(f_*)r(f_*) - P^{s-1}(f)r(f)\} =$$

$$\sum_{s=1}^{t}\{(P^{s-1}(f_*) - P^*(f_*))r(f_*) - (P^{s-1}(f) - P^*(f))r(f)\} +$$

$$t\{\phi(f_*^\infty) - \phi(f^\infty)\}.$$

Then, we get

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}\{v^s(f_*^\infty) - v^s(f^\infty)\} = \frac{T+1}{2}\{\phi(f_*^\infty) - \phi(f^\infty)\} + \tag{5.2.2}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}\{(P^{s-1}(f_*) - P^*(f_*))r(f_*) - (P^{s-1}(f) - P^*(f))r(f)\}.$$

Since $\phi(f_*^\infty) = \phi \geq \phi(f^\infty)$ and

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}(P^{s-1}(f) - P^*(f))r(f) = u(f^\infty)$$

(cf. theorem 2.4.1(iv)) for all $f^\infty \in C_D$, it follows from (5.2.2) that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}\{v^s(f_*^\infty) - v^s(f^\infty)\} \geq 0 \quad \text{for each } f^\infty \in C_D.$$

(iv) $\Rightarrow$ (i): Let f_0^∞ be any Blackwell optimal policy. From (5.2.2) it follows that

$$0 \leq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}\{v^s(f_*^\infty) - v^s(f^\infty)\} =$$

$$\lim_{T\to\infty}[\tfrac{T+1}{2}\{\phi(f_*^\infty) - \phi\} + \{u(f_*^\infty) - u(f_0^\infty)\} + \{\varepsilon(f_*,T) - \varepsilon(f_0,T)\}]$$

where $\lim_{T\to\infty}\varepsilon(f_*,T) = \varepsilon(f,T) = 0$. Therefore, $\phi(f_*) = \phi$ and $u(f_*) \geq u(f_0^\infty)$.

Hence, we have

$$\lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = \lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha(f_0)\} =$$

$$\lim_{\alpha\uparrow 1}\{\frac{\phi(f_*^\infty) - \phi(f_0^\infty)}{1-\alpha} + u(f_*^\infty) - u(f_0^\infty)\} = u(f_*^\infty) - u(f_0^\infty) \geq 0. \quad \square$$

REMARK 5.2.1. DENARDO & MILLER [1968] have proved the equivalence of the first three statements. This equivalence was conjectured by VEINOTT [1966]. In HORDIJK & SLADKY [1977] the equivalence is shown for a countable state space under a condition of Lyapunov function type. For a finite state space this condition is equivalent to the assumption that a fixed state can be reached from each initial state under any stationary policy.

DEFINITION 5.2.1. Let f_*^∞ be a pure and stationary bias optimal policy. Then, $u := u(f_*^\infty)$ is called the *bias-value-vector*.

REMARK 5.2.2. From statement (iii) in theorem 5.2.2 and the results of chapter 4, it follows that a pure and stationary bias optimal policy can be obtained from the algorithm stated below. This algorithm may be very attractive if the linear program (4.2.11) has only a few extreme optimal solutions.

ALGORITHM XXII for the construction of a pure and stationary bias optimal policy by analysing the average optimal policies.

step 1: Determine by algorithm II all extreme optimal solutions, say (x^k,y^k) $k = 1,2,\dots,K$, of the linear programming problem (4.2.11).

step 2: Compute $u(f_k^\infty) := \{[I-P(f_k)+P^*(f_k)]^{-1} - P^*(f_k)\}r(f_k)$, $k \in F_*$, where

$$F_* := \{k \mid \pi^\infty(x^k,y^k), \text{ defined by (4.3.1), belongs to } C_D\},$$

and let $f_k^\infty := \pi^\infty(x^k,y^k)$, $k \in F_*$.

step 3: Take $f_*^\infty \in F_*$ such that $u(f_*^\infty) \geq u(f_k^\infty)$, $k \in F_*$.

THEOREM 5.2.3. *The pure and stationary policy f_*^∞ determined by algorithm XXII is a bias optimal policy.*

PROOF. From the construction of the policy f_*^∞ it follows that $u(f_*^\infty) = \max\{u(f^\infty) \mid f^\infty \in F_*\}$. Hence, theorem 5.2.2 implies that it is sufficient to prove that $f^\infty \in F_*$ if and only if f^∞ is average optimal. The identity of F_* and the set of pure and stationary average optimal policies is a consequence of the theorems 4.3.3 and 4.3.4. □

5.3. LINEAR PROGRAMMING APPROACH (GENERAL CASE)

In order to compute a bias optimal policy, we first solve the linear program for the computation of a pure and stationary average optimal policy (see algorithm XIV). Therefore, we have to compute optimal solutions (ϕ^*,u^*) and (x^*,y^*) of the pair of dual linear programs

$$(5.3.1) \qquad min\left\{\sum_j \beta_j \hat{\phi}_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij}-p_{iaj})\hat{\phi}_j \geq 0 & , a \in A(i),\ i \in E \\ \hat{\phi}_i + \sum_j (\delta_{ij}-p_{iaj})\hat{u}_j \geq r_{ia}, & a \in A(i),\ i \in E \end{array}\right\}$$

and

$$(5.3.2) \qquad max\left\{\sum_i \sum_a r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij}-p_{iaj}) x_{ia} = 0\ , & j \in E \\ \sum_a x_{ja} + \sum_i \sum_a (\delta_{ij}-p_{iaj}) y_{ia} = \beta_j, & j \in E \\ x_{ia}, y_{ia} \geq 0,\ a \in A(i), & i \in E \end{array}\right\}$$

respectively, where $\beta_j > 0$, $j \in E$, are given numbers with $\Sigma_j \beta_j = 1$.

After the solution of the linear program (5.3.1), we can determine

$$\bar{A}(i) := \{a \in A(i) \mid \sum_j (\delta_{ij}-p_{iaj})\phi_j^* = 0\}, \qquad i \in E$$

and

$$\tilde{A}(i) := \{a \in \bar{A}(i) \mid \phi_i^* + \sum_j (\delta_{ij}-p_{iaj})u_j^* = r_{ia}\}, \qquad i \in E.$$

Moreover, theorem 4.2.2 implies that $\phi^* = \phi$, where ϕ is the AMD-value-vector.

For any $f^\infty \in C_D$ we may consider the Markov chain induced by $P(f)$. For this Markov chain we introduce the following notations:

$R(f)$: the set of recurrent states.

$T(f)$: the set of transient states.

$n(f)$: the number of ergodic sets.

Furthermore, we define

$$\tilde{E} := \{i \in E \mid \tilde{A}(i) \neq \emptyset\}$$

LEMMA 5.3.1. *Let* f^∞ *be any pure and stationary average optimal policy. Then,*

(i) $f(i) \in \bar{A}(i)$, $i \in E$.

(ii) $f(i) \in \tilde{A}(i)$, $i \in R(f)$.

(iii) $u_i(f^\infty) = u_i^* - (P^*(f)u^*)_i$, $i \in R(f)$.

(iv) $u_i(f^\infty) \leq u_i^* - (P^*(f)u^*)_i$, $i \in T(f)$.

PROOF.

(i) Since $P(f)\phi = P(f)P^*(f)r(f) = P^*(f)r(f) = \phi$, we have $f(i) \in \bar{A}(i)$, $i \in E$.

(ii) From theorem 4.3.3 it follows that $(x(f),y(f))$, defined by (4.3.2), is an optimal solution of program (5.3.2). Proposition 4.3.3 implies that $R(f) = E_{x(f)}$. From the complementary slackness property of linear programming, we obtain $f(i) \in \tilde{A}(i)$, $i \in E_{x(f)} = R(f)$.

(iii) Since $d_{ij}(f) = 0$, $i \in R(f)$, $j \in T(f)$ (see formula (2.4.3)), it follows from part (ii) that

$$[D(f)\{\phi + (I-P(f))u^*\}]_i = [D(f)r(f)]_i = u_i(f^\infty), \quad i \in R(f).$$

Hence

$$u_i(f^\infty) = [D(f)P^*(f)r(f) + D(f)(I-P(f))u^*]_i, \qquad i \in R(f).$$

Then, by theorem 2.4.1, we get

$$u_i(f^\infty) = u_i^* - (P^*(f)u^*)_i, \qquad i \in R(f).$$

(iv) Since $d_{ij}(f) \geq 0$, $i,j \in T(f)$ (see formula (2.4.3) and theorem 2.3.1), we obtain

$$d_{ij}(f)\{\phi_j + \textstyle\sum_k(\delta_{jk} - p_{jk}(f))u_k^*\} \geq d_{ij}(f)r_j(f), \qquad i,j \in T(f).$$

Part (ii) of the theorem implies that

$$d_{ij}(f)\{\phi_j + \textstyle\sum_k(\delta_{jk} - p_{jk}(f))u_k^*\} = d_{ij}(f)r_j(f), \quad i \in T(f),\ j \in R(f).$$

Hence, we have, using theorem 2.4.1,

$$u_i(f^\infty) = [D(f)r(f)]_i \leq [D(f)\{P^*(f)r(f) + (I-P(f))u^*\}]_i =$$

$$u_i^* - (P^*(f)u^*)_i, \quad i \in T(f). \quad \square$$

In the second part of the algorithm, we try to find the bias-value-vector u for the states that are recurrent under at least one bias optimal policy. Lemma 5.3.1 implies that the states of $E \setminus \tilde{E}$ are transient under all average optimal policies and that in the recurrent states i the chosen actions belong to $\tilde{A}(i)$. Hence, in the second part of the algorithm we restrict ourselves to the states of $\tilde{E}$ and the actions of $\tilde{A}(i)$, $i \in \tilde{E}$.

We want to solve a second Markov decision problem with state space $\tilde{E}$ and action sets $\tilde{A}(i)$, $i \in \tilde{E}$. Therefore, for every $i \in \tilde{E}$ we remove the action a_i from $\tilde{A}(i)$ when $\Sigma_{j \in E \setminus \tilde{E}}\, p_{ia_ij} > 0$. Hence, using the procedure stated below, we obtain a subspace $\tilde{E}$ of E and subsets $\tilde{A}(i)$ of $A(i)$, $i \in \tilde{E}$, such that $\tilde{E}$ is closed under any policy which takes actions only from $\tilde{A}(i)$, $i \in \tilde{E}$.

Procedure

step 1: If $p_{iaj} = 0$ for all $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$: STOP.
Otherwise, go to step 2.

step 2: Take $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$ such that $p_{iaj} > 0$;
$\tilde{A}(i) := \tilde{A}(i)\backslash\{a\}$;
If $\tilde{A}(i) = \emptyset$, then $\tilde{E} := \tilde{E}\backslash\{i\}$;
Go to step 1.

For any policy f^∞ such that $f(i) \in \tilde{A}(i)$, $i \in \tilde{E}$, we denote by $\tilde{f}^\infty$ the restriction to $\tilde{E}$; similarly, we denote by $\tilde{\phi}$, $\tilde{u}^*$, $r(\tilde{f})$, $P(\tilde{f})$ and $P^*(\tilde{f})$ the restriction to $\tilde{E}$ of ϕ, u^*, $r(f)$, $P(f)$ and $P^*(f)$ respectively.

LEMMA 5.3.2. *Let f^∞ be any pure and stationary average optimal policy. Suppose that the sets $\tilde{E}$ and $\tilde{A}(i)$ are the sets obtained by the above procedure. Then,*

(i) $R(f) \subset \tilde{E}$ *and* $f(i) \in \tilde{A}(i)$, $i \in R(f)$.

(ii) *The policy f_1^∞ defined such that*

$$f_1(i) := \begin{cases} a_i \in \tilde{A}(i) & i \in \tilde{E}\backslash R(f) \\ f(i) & \textit{elsewhere} \end{cases}$$

satisfies: 1. $\phi_i(\tilde{f}_1^\infty) = \phi_i(f^\infty) = \phi_i$, $\quad i \in \tilde{E}$.

2. $u_i(\tilde{f}_1^\infty) = u_i(f^\infty) = \tilde{u}_i^* - (P^*(\tilde{f}_1)\tilde{u}^*)_i$, $\quad i \in R(f)$.

PROOF.

(i) Lemma 5.3.1 implies that $R(f) \subset \tilde{E}$ and $f(i) \in \tilde{A}(i)$, $i \in R(f)$, where $\tilde{E}$ and $\tilde{A}(i)$, $i \in R(f)$, are the sets before the above procedure is applied. Since $E\backslash\tilde{E} \subset T(f)$, it follows that if $f(i) \in \tilde{A}(i)$ is removed during the procedure, then $i \in T(f)$. Consequently, after the performance of the procedure, we still have that $R(f) \subset \tilde{E}$ and $f(i) \in \tilde{A}(i)$, $i \in R(f)$.

(ii) Since $p_{i\cdot}(f_1) = p_{i\cdot}(f)$ for every $i \in R(f)$, it follows that the ergodic sets under $P(f)$ are also ergodic sets under $P(f_1)$ (possibly there are some additional ergodic sets under $P(f_1)$). Hence, $p_{i\cdot}^*(f_1) = p_{i\cdot}^*(f)$ for every $i \in R(f)$, and consequently (see formula (2.4.3)) $d_{i\cdot}(f_1) = d_{i\cdot}(f)$ for every $i \in R(f)$. Then, using lemma 5.3.1(iii), we can write

$$u_i(\tilde{f}_1^\infty) = (D(\tilde{f}_1)r(\tilde{f}_1))_i = (D(f_1)r(f_1))_i = (D(f)r(f))_i = u_i(f^\infty) =$$

$$= u_i^* - (P^*(f)u^*)_i = \tilde{u}_i^* - (P^*(\tilde{f}_1)\tilde{u}^*)_i, \quad i \in R(f).$$

Furthermore, we have since $f_1(i) \in \tilde{A}(i)$, $i \in \tilde{E}$:

$$(I-P(\tilde{f}_1))\tilde{\phi} = 0 \quad \text{and} \quad \tilde{\phi} + (I-P(\tilde{f}_1))\tilde{u}^* = r(\tilde{f}).$$

Hence,

$$\phi_i(\tilde{f}_1^\infty) = (P^*(\tilde{f}_1)r(\tilde{f}_1))_i = (P^*(\tilde{f}_1)\tilde{\phi})_i = \tilde{\phi}_i = \phi_i, \quad i \in \tilde{E}.$$

This completes the proof of the lemma. □

Consider an average optimal policy $f^\infty \in C_D$. Lemma 5.3.2 implies that for the maximization of $u_i(f^\infty)$, $i \in R(f)$, we may replace f^∞ by f_1^∞. Because we want to find in this second part of the algorithm the bias-value-vector u in the states that are recurrent under at least one bias optimal policy, we may restrict ourselves to the action sets $\tilde{A}(i)$, $i \in \tilde{E}$. For any policy f^∞ such that $f(i) \in \tilde{A}(i)$, $i \in \tilde{E}$, we have

$$(5.3.3) \qquad \phi(\tilde{f}^\infty) = P^*(\tilde{f})r(\tilde{f}) = P^*(\tilde{f})\{\tilde{\phi} + (I-P(\tilde{f}))\tilde{u}^*\} = P^*(\tilde{f})\tilde{\phi} = \tilde{\phi}$$

and

$$(5.3.4) \qquad u(\tilde{f}^\infty) = D(\tilde{f})r(\tilde{f}) = D(\tilde{f})\{\tilde{\phi} + (I-P(\tilde{f}))\tilde{u}^*\} = \tilde{u}^* - P^*(\tilde{f})\tilde{u}^*.$$

From lemma 5.3.2 it also follows that maximizing $u_i(f^\infty)$ is equivalent to maximizing $- P^*(\tilde{f})\tilde{u}^*$. Notice that the maximum value of $- P^*(\tilde{f})\tilde{u}^*$ is the AMD-value-vector, say ψ, of the Markov decision problem with state space $\tilde{E}$, action sets $\tilde{A}(i)$, $i \in \tilde{E}$, transition probabilities $\tilde{p}_{iaj} := p_{iaj}$, $a \in \tilde{A}(i)$, $i,j \in \tilde{E}$ and rewards $\tilde{r}_{ia} := -\tilde{u}_i^*$, $a \in \tilde{A}(i)$, $i \in \tilde{E}$.

From theorem 4.2.2 it follows that if (ψ^*,v^*) is an optimal solution of the linear program

$$(5.3.5) \qquad \min\left\{\sum_j \beta_j \tilde{\psi}_j \,\middle|\, \begin{array}{ll} \sum_j (\delta_{ij}-p_{iaj})\tilde{\psi}_j \geq 0, & a \in \tilde{A}(i),\ i \in \tilde{E} \\ \tilde{\psi}_i + \sum_j (\delta_{ij}-p_{iaj})\tilde{v}_j \geq -\tilde{u}_i^*, & a \in \tilde{A}(i),\ i \in \tilde{E} \end{array}\right\},$$

then $\psi^* = \psi$.

Theorem 4.2.4 implies that an average optimal policy for this second AMD-model can be found by the following rule:
Let (t^*,s^*) be an extreme optimal solution of the linear program dual to program (5.3.5), i.e. the linear programming problem

$$\max\left\{ \sum_i (-\tilde{u}_i^*) \sum_a t_{ia} \;\middle|\; \begin{array}{lr} \sum_i\sum_a(\delta_{ij}-p_{iaj})t_{ia} & = 0 \ , \ j\in\tilde{E} \\ \sum_a t_{ja} \qquad + \sum_i\sum_a(\delta_{ij}-p_{iaj})s_{ia} & = \beta_j, \ j\in\tilde{E} \\ t_{ia}, s_{ia} \geq 0, \ a \in \tilde{A}(i), \ i\in\tilde{E} & \end{array} \right\}.$$

Then any policy $\tilde{f}_*^\infty$, where

$$\text{(5.3.6)} \qquad \tilde{f}_*(i) := a_i \in \tilde{A}(i) \text{ such that} \begin{cases} t^*_{ia_i} > 0 & i \in E_{t^*} \\ s^*_{ia_i} > 0 & i \in \tilde{E}\backslash E_{t^*}, \end{cases}$$

is an average optimal policy.

<u>THEOREM 5.3.1</u>. *Let* $\tilde{f}_*^\infty$ *be any average optimal policy for the AMD-model* $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$ *and let* f_*^∞ *be a policy for the Markov decision problem* (E,A,p,r) *such that* $f_*(i) = \tilde{f}_*(i)$, $i \in \tilde{E}$. *Then,*

$$u_i(f_*^\infty) = u_i$$

for every state i *which is recurrent under at least one bias optimal policy.*

<u>PROOF</u>. Let g^∞ be any bias optimal policy for the Markov decision problem (E,A,p,r). Define the policy g_1^∞ by

$$g_1(i) := \begin{cases} a_i \in \tilde{A}(i) & i \in \tilde{E}\backslash R(g) \\ g(i) & \text{elsewhere.} \end{cases}$$

Let $\tilde{g}_1^\infty$ be the restriction of g_1^∞ to $\tilde{E}$. Then, by lemma 5.3.2,

$$\text{(5.3.7)} \qquad u_i = u_i(g^\infty) = u_i(\tilde{g}_1^\infty) = \tilde{u}_i^* - (P^*(\tilde{g}_1)\tilde{u}^*)_i, \qquad i \in R(g).$$

Since $\tilde{E}$ is closed under $P(f_*)$, it follows from (2.4.3) that

$$u_i(f_*^\infty) = u_i(\tilde{f}_*^\infty), \qquad i \in \tilde{E}.$$

Because $\tilde{f}_*^\infty$ is average optimal in model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, we can write, using (5.3.4),

$$(5.3.8) \qquad u_i \geq u_i(f_*^\infty) = u_i(\tilde{f}_*^\infty) = \tilde{u}_i^* - (P^*(\tilde{f}_*)\tilde{u}^*)_i \geq \tilde{u}_i^* - (P^*(\tilde{g}_1)\tilde{u}^*)_i, \quad i \in \tilde{E}.$$

Then, (5.3.7) and (5.3.8) imply that

$$u_i = u_i(f_*^\infty), \quad i \in R(g),$$

which completes the proof. □

<u>REMARK 5.3.1</u>. The policy f_*^∞ defined in the above theorem is bias optimal for the states that are recurrent under at least one bias optimal policy. Unfortunately, this set of states is unknown; we only know that it is a subset of $\tilde{E}$. Moreover, (5.3.8) implies that

$$u_i \geq \tilde{u}_i^* - (P^*(\tilde{f}_*)\tilde{u}^*)_i = u_i^* + \psi_i, \quad i \in \tilde{E}.$$

<u>DEFINITION 5.3.1</u>. A vector $z \in \mathbb{R}^N$ is said to be *bias superharmonic* if

$$(5.3.9) \qquad \begin{cases} \sum_j (\delta_{ij}-p_{iaj})z_j \geq r_{ia} - \phi_i & a \in \bar{A}(i),\ i \in E \\ z_i \geq u_i^* + \psi_i & i \in \tilde{E}, \end{cases}$$

where ϕ, u^*, ψ, $\tilde{E}$ and $\bar{A}(i)$ are as defined in the previous part of this section.

<u>THEOREM 5.3.2</u>. *The bias-value-vector* u *is the smallest bias superharmonic vector.*

<u>PROOF</u>. We first show that u is bias superharmonic. We have already seen in remark 5.3.1 that

$$u_i \geq u_i^* + \psi_i, \quad i \in \tilde{E}.$$

Next, we assume that

$$(5.3.10) \qquad \sum_j (\delta_{ij}-p_{iaj})u_j < r_{ia} - \phi_i \quad \text{for some } i \in E \text{ and } a \in \bar{A}(i).$$

Let g^∞ be a bias optimal policy. Then, using theorem 2.4.1, we can write

$$(5.3.11) \qquad (I-P(g))u = (I-P(g))D(g)r(g) = (I-P^*(g))r(g) = r(g)-\phi.$$

Define the policy g_1^∞ by

$$g_1(j) := \begin{cases} g(j) & j \neq i \\ a & j = i. \end{cases}$$

Since $g_1(j) \in \bar{A}(j)$, $j \in E$, we have $P^*(g_1)\phi = \phi$. The transition matrices $P(g)$ and $P(g_1)$ only differ in row i. Hence, (5.3.10) and (5.3.11) imply

$$\begin{cases} u_i - (P(g_1)u)_i < r_i(g_1) - \phi_i \\ u_j - (P(g_1)u)_j = r_j(g_1) - \phi_j & j \neq i. \end{cases}$$

Suppose that $i \in T(g_1)$. Then $R(g_1) \subset R(g)$ and, consequently

$$(5.3.12) \qquad u_j(g_1^\infty) = u_j(g^\infty) = u_j, \quad j \in R(g_1).$$

Hence

$$(5.3.13) \qquad (P^*(g_1)u)_i = [P^*(g_1)u(g_1^\infty)]_i = [P^*(g_1)D(g_1)r(g_1)]_i = 0.$$

Since $i \in T(g_1)$, it follows from (2.4.3) that $d_{ii}(g_1) > 0$. Then, we obtain

$$u_i(g_1^\infty) = [D(g_1)r(g_1)]_i > [D(g_1)\{(I-P(g_1))u + \phi\}]_i =$$

$$[u - P^*(g_1)u + D(g_1)P^*(g_1)\phi]_i = [u - P^*(g_1)u]_i = u_i,$$

implying a contradiction.
If $i \in R(g_1)$, then $p^*_{ii}(g_1) > 0$, and consequently

$$0 = [P^*(g_1)(I-P(g_1))u]_i < [P^*(g_1)(r(g_1)-\phi)]_i = \phi_i(g_1^\infty)-\phi_i \leq 0,$$

implying also a contradiction. Therefore, it has been shown that u is a bias superharmonic vector.
Let z also be a bias superharmonic vector. Assume that

$$[(I-P(g))z]_i > r_i(g)-\phi_i \quad \text{for some } i \in R(g).$$

Then,

$$0 = [P^*(g)(I-P(g))z]_i > [P^*(g)(r(g)-\phi)]_i = 0,$$

which yields a contradiction. Hence,

$$\begin{cases} [(I-P(g))z]_i = r_i(g)-\phi_i, & i \in R(g) \\ [(I-P(g))z]_i \geq r_i(g)-\phi_i, & i \in T(g). \end{cases}$$

Since $D(g))_{\cdot i} \geq 0$ for $i \in T(g)$, we obtain

$$u = u(g^\infty) = D(g)r(g) \leq D(g)\{(I-P(g))z+\phi\} = z-P^*(g)z.$$

If $i \in R(g)$, then (5.3.7) and (5.3.8) imply that $u_i = u_i^* + \psi_i$. Because z is bias superharmonic, we get $z_i \geq u_i$, $i \in R(g)$. Consequently,

$$u \leq z - P^*(g)z \leq z - P^*(g)u = z - P^*(g)D(g)r(g) = z.$$

Hence, we have shown that u is the smallest bias superharmonic vector. □

<u>REMARK 5.3.2</u>. The property of bias superharmonicity depends on the value of u^* which is found in the optimal solution of program (5.3.1). However, the property of being the smallest bias superharmonic vector is independent of which optimal solution is found.

<u>REMARK 5.3.3</u>. The result of theorem 5.3.2 is related to the functional equations of undiscounted Markov decision theory (cf. SCHWEITZER & FEDERGRUEN [1978]).

From theorem 5.3.2 it follows that the bias-value-vector u can be found as the optimal solution of the following linear programming problem

$$(5.3.14) \qquad \min\left\{\sum_j \beta_j z_j \,\middle|\, \begin{array}{ll} \sum_j(\delta_{ij}-p_{iaj})z_j \geq r_{ia} - \phi_i, & a \in \bar{A}(i),\ i \in E \\ z_i \geq u_i^* + \psi_i, & i \in \tilde{E} \end{array}\right\}.$$

The dual program is

$$(5.3.15) \qquad \textit{maximize } \sum_{i\in E}\sum_{a\in \bar{A}(i)} (r_{ia}-\phi_i)\tilde{x}_{ia} + \sum_{i\in \tilde{E}} (u_i^*+\psi_i)\tilde{y}_i$$

$$\textit{subject to } \sum_{i\in E}\sum_{a\in \bar{A}(i)} (\delta_{ij}-p_{iaj})\tilde{x}_{ia} + \sum_{i\in \tilde{E}} \delta_{ij}\tilde{y}_i = \beta_j, \quad j \in E$$

$$\tilde{x}_{ia} \geq 0,\ a \in \bar{A}(i),\ i \in E;\ \tilde{y}_i \geq 0,\ i \in \tilde{E}.$$

The next theorem shows that a pure and stationary bias optimal policy can be obtained from an optimal solution of the linear program (5.3.15). If we solve this linear program by the simplex method, then an extreme optimal solution is obtained and, furthermore, we obtain the bias-value-vector u as the optimal solution of program (5.3.14). The solution of this pair of dual linear programs will be the third part of the algorithm.

THEOREM 5.3.3. *Let* $(\tilde{x}^*,\tilde{y}^*)$ *be an extreme optimal solution of program* (5.3.15). *Suppose that* $\tilde{f}_*^\infty$ *is the policy defined in* (5.3.6). *Then, the pure and stationary policy* g_*^∞*, where*

$$g_*(i) := \begin{cases} \tilde{f}_*(i) & i \in E_* := \{j \in \tilde{E} \mid u_j = u_j^*+\psi_j\} \\ a_i \in \bar{A}(i) \text{ such that } \tilde{x}^*_{ia_i} > 0 & i \in E\backslash E_*, \end{cases}$$

is bias optimal.

PROOF. Suppose that $j \in E\backslash E_*$. Then, the complementary slackness property of linear programming (corollary 1.3.1) implies that $\tilde{y}_j^* = 0$. From the constraints of program (5.3.15) it follows that

$$\sum_{a\in \bar{A}(i)} \tilde{x}^*_{ja} = \beta_j + \sum_{i\in E}\sum_{a\in \bar{A}(i)} p_{iaj}\tilde{x}^*_{ia} \geq \beta_j > 0.$$

Hence, the policy g_*^∞ is well-defined.
The proof of this theorem has the same structure as the proof of theorem 4.2.4, i.e. we first prove three separate propositions and then we complete the proof of the theorem.

PROPOSITION 5.3.1. *Let* f_*^∞ *be any policy for the Markov decision problem* (E,A,p,r) *such that* $f_*(i) = \tilde{f}_*(i),\ i \in \tilde{E}$. *Then,*

$$u_i - (P(f_*)u)_i = r_i(f_*) - \phi_i, \quad i \in E_*.$$

PROOF. Notice that from the construction of $\tilde{E}$ it follows that $\tilde{E}$ is closed under $P(f_*)$. Hence, $(P(f_*)u)_i = (P(\tilde{f}_*)\tilde{u})_i$ and $(P^*(f_*)u)_i = (P^*(\tilde{f}_*)\tilde{u})_i$, $i \in \tilde{E}$, where $\tilde{u}$ is the restriction of u to the states of $\tilde{E}$. Furthermore, (5.3.3) implies that $[P^*(f_*)r(f_*)]_i = \phi_i$, $i \in \tilde{E}$. Suppose that $u_j-(P(f_*)u)_j \neq r_j(f_*)-\phi_j$ for some $j \in E_*$. Then, the constraints of program (5.3.14) imply that

$$u_i - (P(f_*)u)_i \geq r_i(f_*) - \phi_i \quad i \in E_*$$

and

$$u_j - (P(f_*)u)_j > r_j(f_*) - \phi_j, \quad \text{where } j \in E_*.$$

If $j \in R(f_*)$, then we get a contradiction, namely

$$0 = [P^*(f_*)\{u-P(f_*)u\}]_j > [P^*(f_*)(r(f_*)-\phi)]_j = \phi_j(f_*^\infty)-\phi_j = 0.$$

Consequently, we have

$$u_i - (P(f_*)u)_i = r_i(f_*)-\phi_i, \qquad i \in R(f_*)\cap E_*.$$

From formula (2.4.3), it follows that

$$d_{jk}(f_*) = 0, \quad k \notin \tilde{E}, \quad d_{jk}(f_*) \geq 0, \quad k \in T(f_*) \quad \text{and} \quad d_{jj}(f_*) > 0.$$

Hence, we can write, using the results of theorem 2.4.1,

$$(5.3.16) \qquad u_j(\tilde{f}_*^\infty) = \sum_k d_{jk}(f_*)r_k(f_*)$$
$$< \sum_k d_{jk}(f_*)\{u_k - (P(f_*)u)_k + \phi_k\} =$$
$$u_j - (P^*(f_*)u)_j \leq u_j - (P^*(f_*)u(f_*))_j = u_j.$$

Since $\tilde{f}_*^\infty$ is an average optimal policy in the AMD-model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, it follows from (5.3.4) that

$$(5.3.17) \qquad u_i(\tilde{f}_*^\infty) = u_i^* + \psi_i = u_i, \quad i \in E_*.$$

Because $j \in E_*$, (5.3.16) is contradictory to (5.3.17). This completes the proof of the proposition.

PROPOSITION 5.3.2. E_* *is closed under* $P(g_*)$.

PROOF. Let f_*^∞ be any policy for the Markov decision problem (E,A,p,r) such that $f_*(i) = \tilde{f}_*(i)$, $i \in \tilde{E}$. Since $g_*(i) = f_*(i)$, $i \in E_*$, it is sufficient to prove that E_* is closed under $P(f_*)$. By proposition 5.3.1, (5.3.17) and theorem 2.4.1, we have for any $i \in E_*$

$$\begin{aligned} 0 &= u_i - (P(f_*)u)_i - r_i(f_*) + \phi_i \\ &= u_i(f_*^\infty) + [P(f_*)(u(f_*^\infty)-u)]_i - (P(f_*)u(f_*^\infty))_i - r_i(f_*) + \phi_i \\ &= [P(f_*)(u(f_*^\infty)-u)]_i + [\{D(f_*)(I-P(f_*))-I+P^*(f_*)\}r(f_*)]_i \\ &= [P(f_*)(u(f_*^\infty)-u)]_i \\ &= \sum_{j\in\tilde{E}\backslash E_*} p_{ij}(f_*)(u_j(f_*^\infty)-u_j). \end{aligned}$$

Since $u_j(f_*^\infty) - u_j < 0$ for every $j \in \tilde{E}\backslash E_*$, it follows that $p_{ij}(f_*) = 0$, $i \in E_*$, $j \in \tilde{E}\backslash E_*$. Because $f_*(i) \in \tilde{A}(i)$, $i \in \tilde{E}$, it follows from the construction of $\tilde{E}$ that $\tilde{E}$ is closed under $P(f_*)$. Hence, E_* is closed under $P(f_*)$.

PROPOSITION 5.3.3. *The states of* $E\backslash E_*$ *are transient in the Markov chain induced by* $P(g_*)$.

PROOF. Suppose that there is a state $j \in E\backslash E_*$ which is recurrent under $P(g_*)$. Since E_* is closed under $P(g_*)$, there has to exist a nonempty ergodic set $J \subset E\backslash E_*$. Let $J = \{j_1,j_2,\ldots,j_m\}$. The constraints of program (5.3.15) imply that

$$\sum_{a\in\bar{A}(j)} \tilde{x}_{ja} + \tilde{y}_j = \beta_j + \sum_i \sum_{a\in\bar{A}(i)} p_{iaj}\tilde{x}_{ia} \ge \beta_j > 0, \quad j \in \tilde{E}$$

and

$$\sum_{a\in\bar{A}(j)} \tilde{x}_{ja} = \beta_j + \sum_i \sum_{a\in\bar{A}(i)} p_{iaj}\tilde{x}_{ia} \ge \beta_j > 0, \quad j \in E\backslash\tilde{E}.$$

Since $(\tilde{x}^*,\tilde{y}^*)$ is an extreme solution and since the linear program has N constraints, it follows that we have in each state either $\tilde{y}_j^* > 0$ and $\tilde{x}_{ja}^* = 0$ for all $a \in \bar{A}(j)$ or $\tilde{x}_{ja}^* > 0$ for exactly one $a \in \bar{A}(j)$, say for the action a_j. From the complementary slackness property of linear programming it follows that $\tilde{y}_i^* = 0$ for every $i \in \tilde{E}\backslash E_*$. Hence, in every state of J we have exactly one positive variable, namely $\tilde{x}^*_{j_i a_{j_i}}$, $i = 1,2,\ldots,m$. Consequently, by theorem 1.2.2, the vectors $\{q^i,\ i = 1,2,\ldots,m\}$ where

$$q_k^i = \delta_{j_i k} - p_{j_i a_{j_i} k}, \qquad k = 1,2,\dots,N$$

are linearly independent. The definition of g_*^∞ implies that $g_*(j_i) = a_{j_i}$, $i = 1,2,\dots,m$. Since J is closed under $P(g_*)$, we have $q_k^i = 0$, $k \notin J$, $i = 1,2,\dots,m$. Hence, the contracted (i.e. delete the components $k \in E\backslash J$ which are all zeroes) vectors $\{b^i, i = 1,2,\dots,m\}$, where

$$b_k^i = \delta_{j_i j_k} - p_{j_i a_{j_i} j_k}, \quad k = 1,2,\dots,m,$$

are also linearly independent. On the other hand, we have

$$\textstyle\sum_{k=1}^m b_k^i = \sum_{k=1}^m (\delta_{j_i j_k} - p_{j_i a_{j_i} j_k}) = 1 - \sum_{k=1}^m p_{j_i a_{j_i} j_k} = 0,$$

which contradicts the independency of the vectors $\{b^i, i = 1,2,\dots,m\}$. This completes the proof of the proposition.

We can complete the proof of the theorem as follows. Since $\tilde{x}^*_{ig_*(i)} > 0$, $i \in E\backslash E_*$, it follows from the complementary slackness property that

$$u_i - (P(g_*)u)_i = r_i(g_*) - \phi_i, \quad i \in E\backslash E_*. \tag{5.3.18}$$

$P(g_*)$ and $P(f_*)$, where f_*^∞ is the policy of proposition 5.3.1, have the same rows i for $i \in E_*$. Consequently, (5.3.18) and proposition 5.3.1 imply that

$$u - P(g_*)u = r(g_*) - \phi.$$

Since $g_*(i) \in \bar{A}(i)$, $i \in E$, we have $\phi = P^*(g_*)\phi$ and, consequently

$$D(g_*)\phi = D(g_*)P^*(g_*)\phi = 0.$$

Then,

$$u(g_*^\infty) = D(g_*)r(g_*) = D(g_*)(I-P(g_*))u = u - P^*(g_*)u.$$

From proposition 5.3.3 we get $R(g_*) \subset E_*$. Moreover, because E_* is closed under $P(g_*)$ and, by (5.3.17), $u_i = u_i(\tilde{f}_*^\infty) = u_i(g_*^\infty)$, $i \in E_*$, we obtain

$$u(g_*^\infty) = u - P^*(g_*)u = u - P^*(g_*)u(g_*) = u,$$

i.e. g_*^∞ is a bias optimal policy. □

Above, we have derived that a pure and stationary bias optimal policy can be determined by the following algorithm.

ALGORITHM XXIII for the construction of a pure and stationary bias optimal policy (general case).

step 1a: Take any choice for the numbers β_j such that $\beta_j > 0$, $j \in E$, and $\Sigma_j \beta_j = 1$.

step 1b: Compute an optimal solution (ϕ^*, u^*) of the linear programming problem

$$min\left\{\sum_{j\in E}\beta_j\hat{\phi}_j \;\middle|\; \begin{array}{ll} \sum_{j\in E}(\delta_{ij}-p_{iaj})\hat{\phi}_j \geq 0, & a\in A(i), i\in E \\ \hat{\phi}_i + \sum_{j\in E}(\delta_{ij}-p_{iaj})\hat{u}_j \geq r_{ia}, & a\in A(i), i\in E \end{array}\right\}.$$

step 1c: Determine the following sets:

$$\bar{A}(i) := \{a \in A(i) \mid \sum_{j\in E}(\delta_{ij}-p_{iaj})\phi_j^* = 0\}, \qquad i \in E.$$

$$\tilde{A}(i) := \{a \in \bar{A}(i) \mid \phi_i^* + \sum_{j\in E}(\delta_{ij}-p_{iaj})u_j^* = r_{ia}\}, \qquad i \in E.$$

$$\tilde{E} := \{i \in E \mid \tilde{A}(i) \neq \emptyset\}.$$

step 1d: If $p_{iaj} = 0$ for all $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$, then go to step 2a. Otherwise, go to step 1e.

step 1e: Take $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$ such that $p_{iaj} > 0$; $\tilde{A}(i) := \tilde{A}(i)\backslash\{a\}$; if $\tilde{A}(i) = \emptyset$, then $\tilde{E} := \tilde{E}\backslash\{i\}$; go to step 1d.

step 2a: Use the simplex method to compute optimal solutions (ψ^*, v^*) and (t^*, s^*) of the pair of dual linear programming problems

$$min\left\{\sum_{j\in\tilde{E}}\beta_j\tilde{\psi}_j \;\middle|\; \begin{array}{ll} \sum_{j\in\tilde{E}}(\delta_{ij}-p_{iaj})\tilde{\psi}_j \geq 0, & a\in\tilde{A}(i), i\in\tilde{E} \\ \tilde{\psi}_i + \sum_{j\in\tilde{E}}(\delta_{ij}-p_{iaj})\tilde{v}_j \geq -u_i^*, & a\in\tilde{A}(i), i\in\tilde{E} \end{array}\right\}$$

and

$$max\left\{\sum_{i\in\tilde{E}}(-u_i^*)\sum_{a\in\tilde{A}(i)}t_{ia} \;\middle|\; \begin{array}{ll} \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)}(\delta_{ij}-p_{iaj})t_{ia} = 0, & j\in\tilde{E} \\ \sum_{a\in\tilde{A}(i)}t_{ia} + \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)}(\delta_{ij}-p_{iaj})s_{ia} = \beta_j, & j\in\tilde{E} \\ t_{ia}, s_{ia} \geq 0, \; a \in \tilde{A}(i), & i\in\tilde{E} \end{array}\right\}$$

respectively.

step 2b: Take any policy $\tilde{f}_*^\infty$, where

$$\tilde{f}_*(i) := a_i \in \tilde{A}(i) \text{ such that} \begin{cases} t^*_{ia_i} > 0 & i \in E_{t^*} \\ s^*_{ia_i} > 0 & i \in E\backslash E_{t^*}. \end{cases}$$

step 3a: Use the simplex method to compute optimal solutions z^* and (x^*,y^*) of the pair of dual linear programming problems

$$\begin{aligned} &\textit{minimize} \quad \textstyle\sum_{j\in E}\beta_j z_j \\ &\textit{subject to} \ \textstyle\sum_{j\in E}(\delta_{ij}-p_{iaj})z_j \geq r_{ia} - \phi^*_i, \qquad a \in \bar{A}(i),\ i \in E \\ &\qquad\qquad\qquad z_i \geq u^*_i + \psi^*_i, \qquad\qquad i \in \tilde{E} \end{aligned}$$

and

$$\begin{aligned} &\textit{maximize} \quad \textstyle\sum_{i\in E}\sum_{a\in\bar{A}(i)}(r_{ia}-\phi^*_i)x_{ia} + \sum_{i\in\tilde{E}}(u^*_i+\psi^*_i)y_i \\ &\textit{subject to} \ \textstyle\sum_{i\in E}\sum_{a\in\bar{A}(i)}(\delta_{ij}-p_{iaj})x_{ia} + \sum_{i\in\tilde{E}}\delta_{ij}y_i = \beta_j,\ j \in E \\ &\qquad\qquad\qquad x_{ia} \geq 0,\ a \in \bar{A}(i),\ i \in E;\ y_i \geq 0\ ,\ i \in \tilde{E} \end{aligned}$$

respectively.

step 3b: Determine the set $E_* := \{i \in \tilde{E} \,|\, z^*_i = u^*_i + \psi^*_i\}$.

step 3c: Take g_*^∞ such that

$$g_*(i) := \begin{cases} \tilde{f}_*(i) & i \in E_* \\ a_i \text{ such that } x^*_{ia_i} > 0 & i \in E\backslash E_*. \end{cases}$$

The algorithm is displayed in the following simple example.

EXAMPLE 5.3.1. Consider the model of figure 5.3.1. The following calculations can easily be verified.

step 1a: We define $\beta_1:=\beta_2:=\beta_3:=\beta_4:= 1/4$.

step 1b: $\phi^* = (1,1,1,1)^T$; $u^* = (2,1,0,6)^T$.

step 1c: $\bar{A}(1)=\bar{A}(2)=\bar{A}(3)=\bar{A}(4)=\{1,2\}$;

$\tilde{A}(1)=\tilde{A}(2)=\tilde{A}(3)=\{1,2\}$; $\tilde{A}(4)=\emptyset$

$\tilde{E} = \{1,2,3\}$.

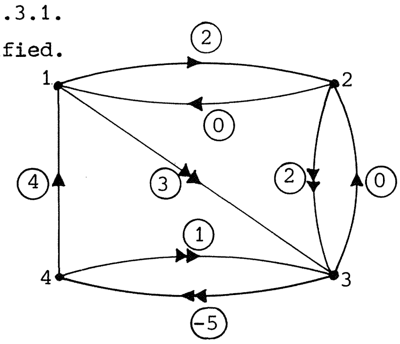


Figure 5.3.1

step 1e: $i=3,\ a=2,\ j=4$: $\tilde{A}(3)=\{1\}$.

step 1d: $p_{iaj}=0$ for all $i\in\tilde{E},\ a\in\tilde{A}(i),\ j\in E\backslash\tilde{E}$.

step 2a: $\psi^*=(-1/2,-1/2,-1/2)^T$; $v^*=(1/2,0,1/2)^T$;

$t^*_{11}=t^*_{12}=t^*_{21}=0,\ t^*_{22}=t^*_{31}=3/8;\ s^*_{11}=1/4,\ s^*_{12}=s^*_{21}=s^*_{31}=0,\ s^*_{22}=1/8.$

step 2b: $\tilde{f}_*(1)=1,\ \tilde{f}_*(2)=2,\ \tilde{f}_*(3)=1$.

step 3a: $z^*=(3/2,1/2,-1/2,9/2)^T$; $x^*_{11}=x^*_{12}=x^*_{21}=x^*_{22}=x^*_{31}=x^*_{32}=x^*_{42}=0,$

$x^*_{41}=1/4;\ y^*_1=1/2,\ y^*_2=1/4,\ y^*_3=1/4.$

step 3b: $E_*=\{1,2,3\}$.

step 3c: $g_*(1)=1,\ g_*(2)=2,\ g_*(3)=1,\ g_*(4)=1$.

5.4. LINEAR PROGRAMMING APPROACH (SPECIAL CASES)

In this section we present three special cases which were also considered for the average reward criterion (see the sections 4.5 and 4.6). In the *weak unichain case* (i.e. when assumption 4.5.1 is satisfied), the linear programming problems which occur in the steps 1b and 2a can be simplified. For the problem used in step 1b, we have presented a simpler program in section 4.5. For the problem studied in step 2a, we take actions from $\tilde{A}(i),\ i\in\tilde{E}$. Hence (cf. formula (5.3.3)), any pure and stationary policy is average optimal in the AMD-model $(\tilde{E},\tilde{A},p,r)$. Consequently, the assumption of weak unichainedness is also verified in the AMD-model $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$ of step 2a. Moreover, since the AMD-value-vector ϕ of model (E,A,p,r) has identical components, we have $\bar{A}(i)=A(i)$ for every $i\in E$. Therefore, the algorithm for the weak unichain case can be formulated as follows.

ALGORITHM XXIV for the construction of a pure and stationary bias optimal policy (weak unichain case).

step 1a: Compute an optimal solution (ϕ^*,u^*) of the linear programming problem

$$min\{\hat{\phi}\,|\,\hat{\phi}+\textstyle\sum_{j\in E}(\delta_{ij}-p_{iaj})\hat{u}_j\ \geq r_{ia},\quad a\in A(i),\ i\in E\}.$$

step 1b: Determine the following sets:

$$\tilde{A}(i)\ :=\{a\in A(i)\,|\,\phi^*+\textstyle\sum_{j\in E}(\delta_{ij}-p_{iaj})u^*_j=r_{ia}\},\qquad i\in E.$$

$\tilde{E} := \{i \in E | \tilde{A}(i) \neq \emptyset\}$.

<u>step 1c</u>: If $p_{iaj} = 0$ for all $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$, then go to step 2a. Otherwise, go to step 1d.

<u>step 1d</u>: Take $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$ such that $p_{iaj} > 0$; $\tilde{A}(i) := \tilde{A}(i)\backslash\{a\}$; if $\tilde{A}(i) = \emptyset$, then $\tilde{E} := \tilde{E}\backslash\{i\}$; go to step 1c.

<u>step 2a</u>: Use the simplex method to compute optimal solutions (ψ^*, v^*) and t^* of the pair of dual linear programming problems

$$min\{\tilde{\psi} | \tilde{\psi} + \sum_{j\in\tilde{E}}(\delta_{ij} - p_{iaj})\tilde{v}_j \geq -u_i^*, \quad a \in \tilde{A}(i),\ i \in \tilde{E}\}$$

and

$$max\left\{\sum_{i\in\tilde{E}}(-u_i^*)\sum_{a\in\tilde{A}(i)} t_{ia} \;\middle|\; \begin{array}{l} \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)}(\delta_{ij} - p_{iaj})t_{ia} = 0,\ j \in \tilde{E} \\ \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)} t_{ia} = 1 \\ t_{ia} \geq 0,\ a \in \tilde{A}(i),\ i \in \tilde{E} \end{array}\right\}$$

respectively.

<u>step 2b</u>: Take $\tilde{f}_*(i)$ such that $t^*_{i\tilde{f}_*(i)} > 0$, $i \in E_{t^*}$; Let $E_o := E_{t^*}$.

<u>step 2c</u>: If $E_o = \tilde{E}$, then go to step 3a. Otherwise, go to step 2d.

<u>step 2d</u>: Take $i \in \tilde{E}\backslash E_o$, $a \in \tilde{A}(i)$, $j \in E_o$ such that $p_{iaj} > 0$; $\tilde{f}(i) := a$; $E_o := E_o \cup \{i\}$; go to step 2c.

<u>step 3a</u>: Use the simplex method to compute optimal solutions z^* and (x^*, y^*) of the pair of dual linear programming problems

$$min\left\{\sum_{j\in E} z_j \;\middle|\; \begin{array}{ll} \sum_{j\in E}(\delta_{ij} - p_{iaj})z_j \geq r_{ia} - \phi^* & a \in A(i),\ i \in E \\ z_i \geq u_i^* + \psi^* & i \in \tilde{E} \end{array}\right\}$$

and

$$\begin{aligned} &maximize \quad \sum_{i\in E}\sum_{a\in A(i)}(r_{ia} - \phi^*)x_{ia} + \sum_{i\in\tilde{E}}(u_i^* + \psi^*)y_i \\ &subject\ to\ \sum_{i\in E}\sum_{a\in A(i)}(\delta_{ij} - p_{iaj})x_{ia} + \sum_{i\in\tilde{E}}\delta_{ij}y_i = 1, \quad j \in E \\ &\qquad x_{ia} \geq 0,\ a \in A(i),\ i \in E;\ y_i \geq 0, \quad i \in \tilde{E} \end{aligned}$$

respectively.

step 3b: Determine the set $E_* := \{i \in \tilde{E} \mid z_i^* = u_i^* + \psi^*\}$.
step 3c: Take g_*^∞ such that

$$g_*(i) := \begin{cases} \tilde{f}_*(i) & , \quad i \in E_* \\ a_i \text{ such that } x_{ia_i}^* > 0, & i \in E \setminus E_*. \end{cases}$$

In the *completely ergodic case* (i.e. when assumption 4.6.1 is satisfied) the algorithm becomes rather simple. Since all states are recurrent under every pure and stationary policy, lemma 5.3.1 and theorem 5.3.1 imply that $\tilde{E} = E$ and that $\tilde{f}_*^\infty$, defined in step 2, is a bias optimal policy. Hence, step 3 can be deleted and we obtain the following algorithm.

ALGORITHM XXV for the construction of a pure and stationary bias optimal policy (completely ergodic case).

step 1a: Compute an optimal solution (ϕ^*, u^*) of the linear programming problem

$$\min\{\hat{\phi} \mid \hat{\phi} + \textstyle\sum_{j \in E} (\delta_{ij} - p_{iaj}) \hat{u}_j \geq r_{ia}, \quad a \in A(i),\ i \in E\}.$$

step 1b: Determine

$$\tilde{A}(i) := \{a \in A(i) \mid \phi^* + \textstyle\sum_{j \in E} (\delta_{ij} - p_{iaj}) u_j^* = r_{ia}\}, \quad i \in E.$$

step 2a: Use the simplex method to compute an optimal solution t^* of the linear programming problem

$$\max\left\{ \sum_{i \in E} (-u_i^*) \sum_{a \in \tilde{A}(i)} t_{ia} \;\middle|\; \begin{array}{ll} \sum_{i \in E} \sum_{a \in \tilde{A}(i)} (\delta_{ij} - p_{iaj}) t_{ia} = 0, & j \in E \\ \sum_{i \in E} \sum_{a \in \tilde{A}(i)} t_{ia} = 1 & \\ t_{ia} \geq 0, \ a \in \tilde{A}(i),\ i \in E & \end{array} \right\}.$$

step 2b: Take $\tilde{f}_*^\infty$ such that $t_{i\tilde{f}_*(i)}^* > 0$, $i \in E$.

We close this chapter with the presentation of the algorithm for the *unichain case*, i.e. when assumption 4.6.2 is satisfied. From the results of the sections 4.6 and 5.3 it is straightforward that in this case a bias optimal policy can be determined by the following algorithm.

ALGORITHM XXVI for the construction of a pure and stationary bias optimal policy (unichain case).

step 1a: Compute an optimal solution (ϕ^*, u^*) of the linear programming problem

$$min\{\hat{\phi} \mid \hat{\phi} + \sum_{j\in E}(\delta_{ij}-p_{iaj})\hat{u}_j \geq r_{ia}, \quad a \in A(i),\ i \in E\}.$$

step 1b: Determine the following sets:

$$\tilde{A}(i) := \{a \in A(i) \mid \phi^* + \sum_{j\in E}(\delta_{ij}-p_{iaj})u_j^* = r_{ia}\}, \quad i \in E.$$

$$\tilde{E} := \{i \in E \mid \tilde{A}(i) \neq \emptyset\}.$$

step 1c: If $p_{iaj} = 0$ for all $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$, then go to step 2a. Otherwise, go to step 1d.

step 1d: Take $i \in \tilde{E}$, $a \in \tilde{A}(i)$, $j \in E\backslash\tilde{E}$ such that $p_{iaj} > 0$; $\tilde{A}(i) := \tilde{A}(i)\backslash\{a\}$; if $\tilde{A}(i) = \emptyset$, then $\tilde{E} := \tilde{E}\backslash\{i\}$; go to step 1c.

step 2a: Use the simplex method to compute optimal solutions (ψ^*, v^*) and t^* of the pair of dual linear programming problems

$$min\{\tilde{\psi} \mid \tilde{\psi} + \sum_{j\in\tilde{E}}(\delta_{ij}-p_{iaj})\tilde{v}_j \geq -u_i^*, \quad a \in \tilde{A}(i),\ i \in \tilde{E}\}$$

and

$$max\left\{\sum_{i\in\tilde{E}}(-u_i^*)\sum_{a\in\tilde{A}(i)} t_{ia} \;\middle|\; \begin{array}{lr} \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)}(\delta_{ij}-p_{iaj})t_{ia} = 0, & j\in\tilde{E} \\ \sum_{i\in\tilde{E}}\sum_{a\in\tilde{A}(i)} t_{ia} = 1 & \\ t_{ia} \geq 0, & a\in\tilde{A}(i), i\in\tilde{E} \end{array}\right\}$$

respectively.

step 2b: Take $\tilde{f}_*^\infty$ such that

$$\tilde{f}_*(i) := \begin{cases} a_i \quad \text{where } t^*_{ia_i} > 0, & i \in E_{t^*} \\ \text{arbitrarily} & i \in \tilde{E}\backslash E_{t^*}. \end{cases}$$

step 3a: Use the simplex method to compute optimal solutions z^* and (x^*, y^*) of the pair of dual linear programming problems

$$\min\left\{\sum_{j\in E} z_j \;\middle|\; \begin{array}{ll} \sum_{j\in E}(\delta_{ij}-p_{iaj})z_j \geq r_{ia}-\phi^* , & a \in A(i),\ i \in E \\ z_i \geq u_i^* + \psi^*, & i \in \tilde{E} \end{array}\right\}$$

and

$$\begin{array}{ll} \textit{maximize} & \sum_{i\in E}\sum_{a\in A(i)}(r_{ia}-\phi^*)x_{ia} + \sum_{i\in\tilde{E}}(u_i^*+\psi^*)y_i \\ \textit{subject to} & \sum_{i\in E}\sum_{a\in A(i)}(\delta_{ij}-p_{iaj})x_{ia} + \sum_{i\in\tilde{E}}\delta_{ij}y_i = 1, \quad j \in E \\ & x_{ia} \geq 0,\ a \in A(i),\ i \in E;\ y_i \geq 0,\ i \in \tilde{E} \end{array}$$

respectively.

<u>step 3b</u>: Determine the set $E_* := \{i \in \tilde{E} \mid z_i^* = u_i^* + \psi^*\}$.

<u>step 3c</u>: Take g_*^∞ such that

$$g_*(i) := \begin{cases} \tilde{f}_*(i) & i \in E_* \\ a_i \ \text{ such that } \ x_{ia_i}^* > 0, & i \in E \backslash E_*. \end{cases}$$

CHAPTER 6

TWO-PERSON ZERO-SUM STOCHASTIC GAMES IN WICH ONE PLAYER CONTROLS THE TRANSITION PROBABILITIES

6.1. INTRODUCTION AND SUMMARY

In this chapter we investigate a two-person zero-sum stochastic game. This game can be described as follows. Consider a system with a finite state space $E = \{1,2,\dots,N\}$ that is observed at discrete time points $t = 1,2,\dots$. If the system is in state i (at some time point t), then both players choose simultaneously an action from their own finite action sets A(i) and B(i) for player I and player II respectively. If in state i player I chooses action $a \in A(i)$ and player II action $b \in B(i)$, then the following occurs:

1. Player I receives an immediate reward r_{iab} from player II.
2. The next state of the system is chosen according to the transition probabilities p_{iabj}, where $p_{iabj} \geq 0$ and $\Sigma_j\, p_{iabj} \leq 1$ for every $a \in A(i)$, $b \in B(i)$, $i \in E$.

A *two-person zero-sum stochastic game* is denoted by a five-tuple (E,A,B,p,r), where

- E is the *state space*
- $A = \cup_{i\in E}\, A(i)$ is the *action space for player I*
- $B = \cup_{i\in E}\, B(i)$ is the *action space for player II*
- p is a *transition probability* from $E \times A \times B$ to E
- r is a real-valued *reward function* on $E \times A \times B$

($E\times A\times B$ has to be interpreted as $\{(i,a,b)\,|\,i \in E,\ a \in A(i),\ b \in B(i)\}$). Stochastic games are also called *Markov games*. If the action space for one of the two players consists of one element, then the game becomes a Markov decision problem.

Let H_t denote the set of possible *histories* of the system up to time t, i.e.

$$H_t := \left\{ (i_1,a_1,b_1,\dots,i_{t-1},a_{t-1},b_{t-1},i_t) \;\middle|\; \begin{array}{l} i_k\in E,\ a_k\in A(i_k),\ b_k\in B(i_k) \\ k = 1,2,\dots,t-1;\ i_t\in E \end{array} \right\}.$$

A *decision rule* π^t for player I at time t is a nonnegative function on $H_t \times A$ such that for every $(i_1, a_1, b_1, \ldots, b_{t-1}, i_t) \in H_t$

$$\pi^t_{i_1 a_1 b_1 \ldots b_{t-1} i_t a_t} = 0 \quad \text{if } a_t \notin A(i_t)$$

and

$$\sum_{a_t} \pi^t_{i_1 a_1 b_1 \ldots b_{t-1} i_t a_t} = 1.$$

A *policy* R_1 for player I is a sequence of decision rules: $R_1 = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$. A decision rule ρ^t for player II at time t is a nonnegative function on $H_t \times B$ such that for every $(i_1, a_1, b_1, \ldots, b_{t-1}, i_t) \in H_t$

$$\rho^t_{i_1 a_1 b_1 \ldots b_{t-1} i_t b_t} = 0 \quad \text{if } b_t \notin B(i_t)$$

and

$$\sum_{b_t} \rho^t_{i_1 a_1 b_1 \ldots b_{t-1} i_t b_t} = 1.$$

A policy R_2 for player II is a sequence of decision rules: $R_2 = (\rho^1, \rho^2, \ldots, \rho^t, \ldots)$. If the decision rules of a policy are independent of the histories and the time points, then the policy is said to be *stationary*; furthermore, if the decision rules are nonrandomized, then the policy is said to be *pure*.

For any pair (R_1, R_2) of policies for player I and player II, we denote by $p^t_{ijab}(R_1, R_2)$ the probability that - given that the system starts in state i - the system is at time t in state j and then the actions a and b are chosen by player I and player II respectively. Let $\{X_t, t = 1,2,\ldots\}$, $\{Y_t, t = 1,2,\ldots\}$ and $\{Z_t, t = 1,2,\ldots\}$ be the sequences of random variables denoting the observed states, the actions chosen by player I and the actions chosen by player II respectively. Then, we can also write

$$p^t_{ijab}(R_1, R_2) = \mathbb{P}_{R_1, R_2}(X_t = j, Y_t = a, Z_t = b \mid X_1 = i).$$

The expected reward in the t-th period, when the policies R_1 and R_2 are used and i is the initial state, is denoted by $v^t_i(R_1, R_2)$, i.e.

$$v^t_i(R_1, R_2) := \sum_j \sum_a \sum_b \mathbb{P}_{R_1, R_2}(X_t = j,\ Y_t = a,\ Z_t = b \mid X_1 = i) \cdot r_{jab}.$$

The *expected total reward* over an infinite horizon, when the policies R_1

and R_2 are used and i is the initial state, is denoted by $v_i(R_1,R_2)$, i.e.

$$v_i(R_1,R_2) := \sum_{t=1}^{\infty}\sum_j\sum_a\sum_b \mathbb{P}_{R_1,R_2}(X_t=j,\ Y_t=a,\ Z_t=b \mid X_1=i)\cdot r_{jab}.$$

Using the above notation we assume that $\lim_{T\to\infty} \Sigma_{t=1}^{T} v_i^t(R_1,R_2)$ exists (possibly $+\infty$ or $-\infty$). For a Markov game with as utility function the total reward criterion we will use the name *TMG-model*. Player I wants to maximize his rewards and player II wants to minimize his payments. Hence, the aim is to find policies R_1^* and R_2^* such that

$$(6.1.1) \qquad v(R_1,R_2^*) \le v(R_1^*,R_2^*) \le v(R_1^*,R_2) \qquad \text{for all policies } R_1,R_2.$$

If R_1^* and R_2^* satisfy (6.1.1), then R_1^* and R_2^* are called *optimal policies* for player I and player II respectively. We are also interested in the value of $v(R_1^*,R_2^*)$ which will be denoted by *val(TMG)* and is called the value of the TMG-model or the *value of the game*.

In section 6.2 we consider the total reward criterion under the contraction assumption as introduced in section 3.4. It is well-known that in this model the value of the game exists. We will see that, in general, the value of the game does not lie in the same field as the field generated by the data r_{iab}, p_{iabj}, $i,j \in E$, $a \in A(i)$, $b \in B(i)$. In the simplex method only the operations addition, subtraction, multiplication and division are used. Hence, in general, the value of the game cannot be computed by linear programming. If we assume that the transition probabilities only depend on one player, say player I, then it can be shown that the value as well as stationary optimal policies for both players can be computed by linear programming. For this reason we investigate the model in which one player controls the transition probabilities. We shall show that the value of the game is the smallest vector which satisfies a superharmonic property. Then, we can formulate a pair of dual linear programs. Stationary optimal policies as well as val(TMG) can be obtained from optimal solutions of these linear programs. Hence, we can present an algorithm to compute these quantities by linear programming.

Section 6.3 deals with the average reward criterion. The *expected average reward* over an infinite horizon, when the policies R_1 and R_2 are used and state i is the initial state, is denoted by $\phi_i(R_1,R_2)$ and defined by

$$\phi_i(R_1,R_2) := \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T}\sum_j\sum_a\sum_b \mathbb{P}_{R_1,R_2}(X_t=j,Y_t=a,Z_t=b \mid X_1=i)\cdot r_{jab}.$$

This model is called the *AMG-model*. The policies R_1^* and R_2^* are said to be optimal for player I and player II respectively if

(6.1.2) $\qquad \phi(R_1,R_2^*) \le \phi(R_1^*,R_2^*) \le \phi(R_1^*,R_2) \quad$ for all policies R_1,R_2.

If R_1^* and R_2^* are optimal policies, then $\phi(R_1^*,R_2^*)$ is the value of the game, denoted by *val(AMG)*.

Also for the AMG-model, we shall assume that only one player controls the transition probabilities. We will present a pair of dual linear programming problems, and we will prove that stationary optimal policies as well as the value of the game can be obtained from optimal solutions of these linear programs. Hence, we can formulate a finite algorithm to construct stationary optimal policies. Furthermore, the linear programming approach provides a new proof for the existence of the value of a stochastic game in which one player controls the transition probabilities. We close section 6.3 by a description in which way the algorithm may be simplified in the unichain case.

LEMMA 6.1.1. *Let* f *be a real-valued function defined on* $X\times Y$, *where* X *and* Y *are given sets. Suppose that* $x^* \in X$ *and* $y^* \in Y$ *satisfy*

$$f(x,y^*) \le f(x^*,y^*) \le f(x^*,y) \quad \textit{for every } x \in X \textit{ and } y \in Y.$$

Then,

$$f(x^*,y^*) = sup_{x\in X} inf_{y\in Y} f(x,y) = inf_{y\in Y} sup_{x\in X} f(x,y).$$

PROOF. Since

$$sup_{x\in X} f(x,y) \ge f(x,y) \qquad \text{for every } x \in X \text{ and } y \in Y,$$

we have

$$inf_{y\in Y} sup_{x\in X} f(x,y) \ge inf_{y\in Y} f(x,y) \quad \text{for every } x \in X.$$

Consequently,

(6.1.3) $\qquad inf_{y\in Y} sup_{x\in X} f(x,y) \ge sup_{x\in X} inf_{y\in Y} f(x,y).$

Since $f(x,y^*) \le f(x^*,y^*)$ for every $x \in X$, it follows that $f(x^*,y^*) = sup_{x\in X} f(x,y^*)$. Hence, we can write

(6.1.4) $\qquad inf_{y\in Y} sup_{x\in X} f(x,y) \le sup_{x\in X} f(x,y^*) = f(x^*,y^*) \le f(x^*,y), \quad y \in Y.$

Similarly, we can derive that

$$(6.1.5) \qquad \sup_{x\in X}\inf_{y\in Y} f(x,y) \geq \inf_{y\in Y} f(x^*,y) = f(x^*,y^*) \geq f(x,y^*), \quad x \in X.$$

Combining (6.1.3), (6.1.4) and (6.1.5) yields

$$\begin{aligned} f(x,y^*) &\leq \sup_{x\in X}\inf_{y\in Y} f(x,y) \\ &\leq \inf_{y\in Y}\sup_{x\in X} f(x,y) \\ &\leq f(x^*,y) \qquad \text{for every } x \in X \text{ and } y \in Y. \end{aligned}$$

Hence,

$$f(x^*,y^*) = \sup_{x\in X}\inf_{y\in Y} f(x,y) = \inf_{y\in Y}\sup_{x\in X} f(x,y),$$

completing the proof of the theorem. □

COROLLARY 6.1.1.

(i) *If* (R_1^*,R_2^*) *is a pair of optimal policies for the TMG-model, then*

$$v(R_1^*,R_2^*) = \sup_{R_1}\inf_{R_2} v(R_1,R_2) = \inf_{R_2}\sup_{R_1} v(R_1,R_2).$$

(ii) *If* (R_1^*,R_2^*) *is a pair of optimal policies for the AMG-model, then*

$$\phi(R_1^*,R_2^*) = \sup_{R_1}\inf_{R_2} \phi(R_1,R_2) = \inf_{R_2}\sup_{R_1} \phi(R_1,R_2).$$

Let π^∞ and ρ^∞ be stationary policies for player I and player II respectively. We introduce the following notations:

$$\begin{aligned}
r_{ia}(\rho) &:= \textstyle\sum_b r_{iab}\rho_{ib} && a \in A(i),\ i \in E, \\
r_{ib}(\pi) &:= \textstyle\sum_a r_{iab}\pi_{ia} && b \in B(i),\ i \in E, \\
r_i(\pi,\rho) &:= \textstyle\sum_a\sum_b r_{iab}\pi_{ia}\rho_{ib} && i \in E, \\
p_{iaj}(\rho) &:= \textstyle\sum_b p_{iabj}\rho_{ib} && a \in A(i),\ i,j \in E, \\
p_{ibj}(\pi) &:= \textstyle\sum_a p_{iabj}\pi_{ia} && b \in B(i),\ i,j \in E, \\
p_{ij}(\pi,\rho) &:= \textstyle\sum_a\sum_b p_{iabj}\pi_{ia}\rho_{ib} && i,j \in E.
\end{aligned}$$

REMARK 6.1.1. Let ρ^∞ be any stationary policy for player II. Consider the Markov decision problem $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, where

$$\tilde{E} := E,$$

$$\tilde{A}(i) := A(i), \qquad i \in \tilde{E},$$

$$\tilde{p}_{iaj} := p_{iaj}(\rho), \qquad a \in \tilde{A}(i),\ i,j \in \tilde{E},$$

$$\tilde{r}_{ia} := r_{ia}(\rho), \qquad a \in \tilde{A}(i), \quad i \in \tilde{E}.$$

Let $R_1 = (\pi^1,\pi^2,\ldots)$ be any policy for player I. Then R_1 induces a policy $\tilde{R}_1 = (\tilde{\pi}^1,\tilde{\pi}^2,\ldots)$ for the Markov decision problem $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, where

$$\tilde{\pi}^t_{i_1a_1\ldots a_{t-1}i_ta_t} := \mathbb{P}_{R_1,\rho^\infty}(Y_t=a_t \mid X_1=i_1,Y_1=a_1,\ldots,Y_{t-1}=a_{t-1},X_t=i_t)$$

for every $t \in \mathbb{N}$ and every history $(i_1,a_1,\ldots,a_{t-1},i_t)$. Then, by induction on t, it can easily be verified that

$$(6.1.6) \qquad \mathbb{P}_{\tilde{R}_1}(X_1 = i_1, Y_1 = a_1,\ldots,Y_{t-1} = a_{t-1}, X_t = i_t, Y_t = a_t) =$$

$$\mathbb{P}_{R_1,\rho^\infty}(X_1 = i_1, Y_1 = a_1,\ldots,Y_{t-1} = a_{t-1}, X_t = i_t, Y_t = a_t)$$

for every $t \in \mathbb{N}$, every history $(i_1,a_1,\ldots,a_{t-1},i_t)$ and every $a_t \in A(i_t)$. (6.1.6) implies that

$$\mathbb{P}_{\tilde{R}_1,\rho^\infty}(X_1 = i_1, Y_1 = a_1, Z_1 = b_1,\ldots,X_t = i_t, Y_t = a_t, Z_t = b_t) =$$

$$\mathbb{P}_{R_1,\rho^\infty}(X_1 = i_1, Y_1 = a_1, Z_t = b_1,\ldots,X_t = i_t, Y_t = a_t, Z_t = b_t)$$

for every $(i_1,a_1,\ldots,i_t,a_t,b_t)$, $t \in \mathbb{N}$. Therefore, the policies $(\tilde{R}_1,\rho^\infty)$ and (R_1,ρ^∞) are equivalent for any utility function. However, the policy $\tilde{R}_1$ is a feasible policy for the Markov decision problem $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$. If $\tilde{v}(\tilde{R}_1)$ and $\tilde{\phi}(\tilde{R}_1)$ denote the expected total reward and the expected average reward respectively in the Markov decision problem $(\tilde{E},\tilde{A},\tilde{p},\tilde{r})$, then we have

1. $\tilde{v}(\tilde{R}_1) = v(R_1,\rho^\infty)$
2. $sup_{R_1} v(R_1,\rho^\infty) = sup_\pi v(\pi^\infty,\rho^\infty)$

3. $\tilde{\phi}(\tilde{R}_1) = \phi(R_1,\rho^\infty)$

4. $sup_{R_1}\,\phi(R_1,\rho^\infty) = sup_\pi\,\phi(\pi^\infty,\rho^\infty)$.

Furthermore, changing the roles of the players I and II, we obtain

5. $inf_{R_2}\,v(\pi^\infty,R_2) = inf_\rho\,v(\pi^\infty,\rho^\infty)$

6. $inf_{R_2}\,\phi(\pi^\infty,R_2) = inf_\rho\,\phi(\pi^\infty,\rho^\infty)$.

6.2. TOTAL REWARD CRITERION

In this section we consider the TMG-model under the following contraction assumption (cf. assumption 3.4.1).

<u>ASSUMPTION 6.2.1</u>. *There exists a vector* $\mu \gg 0$ *and a scalar* $\alpha \in [0,1)$ *such that*

$$\textstyle\sum_j p_{iabj}\mu_j \le \alpha\,\mu_i, \quad a \in A(i),\ b \in B(i),\ i \in E.$$

Assumption 6.2.1 guarantees that the expected total reward is well-defined for any pair (R_1,R_2) of policies. The following theorem has been proved already in 1953 by SHAPLEY [1953] for the discounted Markov game, i.e. the TMG-model under assumption 6.2.1 with $\mu = e$. The extension of the theorem to general positive μ-vectors is straightforward (cf. VAN DER WAL & WESSELS [1977]).

<u>THEOREM 6.2.1</u>. *There exist stationary optimal policies for both players.*

The above theorem implies that val(TMG) exists. The next example will show that, in general, val(TMG) is not an element of the field generated by the data r_{iab}, p_{iabj}, $a \in A(i)$, $b \in B(i)$, $i,j \in E$. Hence, this val(TMG) cannot be computed as solution of a linear program which has all coefficients in this field. Since we study in this monograph linear programming methods, we shall not discuss the general TMG-model, but a model with an additional assumption. Under this assumption, we can compute val(TMG) as well as stationary optimal policies by linear programming. The TMG-model under this additional assumption was first studied by PARTHASARATHY & RAGHAVAN [1977]. The following example is also due to them.

<u>EXAMPLE 6.2.1</u>. Consider the discounted TMG-model of figure 6.2.1 with $\alpha = 0.5$. The interpretation of the figures for TMG-models is similar to

the interpretation of the figures for TMD-models except that a positive $p_{ik\ell j}$ is indicated by an arc from state i to state j with k times ● and ℓ times ▶.

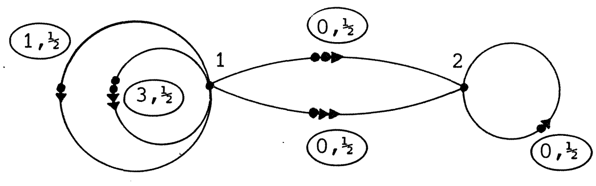


Figure 6.2.1

Let y := val(TMG). Since $v_2(R_1,R_2) = 0$ for all R_1,R_2, we have $y_2 = 0$. It can be shown that y_1 is the value of the matrix game with pay-off matrix

$$\begin{pmatrix} 1+\frac{1}{2}y_1 & 0 \\ 0 & 3+\frac{1}{2}y_1 \end{pmatrix}.$$

Then, using results from the theory of matrix games (e.g. KARLIN [1959] p.50), one can find that

$$y_1 = \frac{(1+\frac{1}{2}y_1)\cdot(3+\frac{1}{2}y_1)}{(1+\frac{1}{2}y_1)+(3+\frac{1}{2}y_1)},$$

implying that $y_1 = \frac{1}{3}(-4+\sqrt{13})$. Hence, $[\text{val(TMG)}]_1$ is not an element of the field of the rational numbers, i.e. the field generated by the data of the above problem.

DEFINITION 6.2.1. A vector $y \in \mathbb{R}^N$ is said to be *TMG-superharmonic* if there exists a stationary policy ρ^∞ for player II such that

$$y_i \geq r_{ia}(\rho) + \sum_j p_{iaj}(\rho)y_j, \quad a \in A(i),\ i \in E.$$

THEOREM 6.2.2. *val(TMG) is the smallest TMG-superharmonic vector.*

PROOF. Let $(\pi^*)^\infty$ and $(\rho^*)^\infty$ be stationary optimal policies for player I and player II respectively (theorem 6.2.1 implies the existence of such a pair of policies). If player II uses policy $(\rho^*)^\infty$, then the stochastic game may be interpreted as a Markov decision problem (see remark 6.1.1). Furthermore, since $(\rho^*)^\infty$ is optimal for player II, we have $sup_{R_1} v(R_1,(\rho^*)^\infty) = \text{val(TMG)}$. Hence, the TMD-model has TMD-value-vector val(TMG). Consequently, theorem 3.4.1 implies that val(TMG) is TMD-superharmonic, i.e.

$$[\text{val(TMG)}]_i \geq r_{ia}(\rho^*) + \sum_j p_{iaj}(\rho^*)[\text{val(TMG)}]_j, \quad a \in A(i),\ i \in E.$$

Therefore, val(TMG) is also TMG-superharmonic.

Suppose that y is another TMG-superharmonic vector with corresponding stationary policy ρ^∞. Then, it follows from definition 6.2.1 that $y \geq r(\pi^*,\rho) + P(\pi^*,\rho)y$. Assumption 6.2.1 and theorem 2.3.1 imply that $(I-P(\pi^*,\rho))^{-1} = \Sigma_{t=1}^{\infty} P^{t-1}(\pi^*,\rho)$. Hence,

$$y \geq \sum_{t=1}^{\infty} P^{t-1}(\pi^*,\rho)\, r(\pi^*,\rho) = v((\pi^*)^\infty,\rho^\infty).$$

Since $(\pi^*)^\infty$ is optimal for player I, we have

$$y \geq v((\pi^*)^\infty,\rho^\infty) \geq v((\pi^*)^\infty,(\rho^*)^\infty) = \text{val(TMG)}.$$

This completes the proof. □

From theorem 6.2.2 it follows that val(TMG) is the optimal solution of the following nonlinear programming problem in which the objective function is linear and there are linear as well as quadratic constraints (cf. ROTHBLUM [1979]):

$$\begin{aligned} &\textit{minimize} \quad \sum_j \beta_j y_j \\ &\textit{subject to } y_i \geq \sum_b r_{iab}\rho_{ib} + \sum_j \sum_b p_{iabj}\rho_{ib} y_j, \quad a \in A(i),\ i \in E, \\ &\qquad \sum_b \rho_{ib} = 1 \qquad\qquad\qquad\qquad\qquad\qquad i \in E, \\ &\qquad\quad \rho_{ib} \geq 0 \qquad\qquad\qquad\qquad\qquad , \quad b \in B(i),\ i \in E, \end{aligned}$$

where $\beta_j > 0$, $j \in E$, are given numbers.

To obtain a linear program we assume that we have in the remaining part of this section the following assumption.

ASSUMPTION 6.2.2. *The transition probabilities* p_{iabj}, $j \in E$, *do not depend on* b *for all* $i \in E$, $a \in A(i)$.

Because of assumption 6.2.2, we will denote the transition probabilities p_{iabj} by p_{iaj} and the transition matrix $P(\pi,\rho)$ by $P(\pi)$. Under this assumption we obtain the following linear programming problem

$$(6.2.1)\qquad min\left\{\sum_j\beta_j y_j \left|\begin{array}{llr}\sum_j(\delta_{ij}-p_{iaj})y_j - \sum_b r_{iab}\rho_{ib} & \geq 0, & a\in A(i),\ i\in E\\ \sum_b\rho_{ib} & = 1 & ,\ i\in E\\ \rho_{ib} & \geq 0, & b\in B(i),\ i\in E\end{array}\right.\right\}.$$

The dual linear programming problem is

$$(6.2.2)\qquad max\left\{\sum_i z_i \left|\begin{array}{llr}\sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} & = \beta_j, & j\in E\\ -\sum_a r_{iab}x_{ia} + z_i & \leq 0\ , & b\in B(i),\ i\in E\\ x_{ia} & \geq 0\ , & a\in A(i),\ i\in E\end{array}\right.\right\}.$$

<u>THEOREM 6.2.3</u>. *Let* (y^*,ρ^*) *and* (x^*,z^*) *be optimal solutions of the linear programming problems* (6.2.1) *and* (6.2.2) *respectively. Define the stationary policy* $(\pi^*)^\infty$ *by*

$$\pi^*_{ia} := x^*_{ia}\Big/\sum_a x^*_{ia}, \qquad a\in A(i),\ i\in E.$$

Then, $(\pi^*)^\infty$ *and* $(\rho^*)^\infty$ *are stationary optimal policies for player I and player II respectively, and* y^* *is the value of the game.*

<u>PROOF</u>. Theorem 6.2.2 implies that y^* is the value of the game. Since

$$\sum_a x^*_{ja} = \beta_j + \sum_i\sum_a p_{iaj}x^*_{ia} \geq \beta_j > 0, \qquad j\in E,$$

the stationary policy $(\pi^*)^\infty$ is well-defined. From the constraints of program (6.2.1) it follows that

$$(I-P(\pi))y^* \geq r(\pi,\rho^*) \qquad \text{for every stationary policy } \pi^\infty.$$

Since $(I-P(\pi))^{-1} = \Sigma_{t=1}^\infty P^{t-1}(\pi)$, we get

$$(6.2.3)\qquad y^* \geq \sum_{t=1}^\infty P^{t-1}(\pi)r(\pi,\rho^*) = v(\pi^\infty,(\rho^*)^\infty) \qquad \text{for every stationary policy } \pi^\infty.$$

$\pi^*_{ia} > 0$ if and only if $x^*_{ia} > 0$ and, consequently, the complementary slackness property of linear programming (cf. corollary 1.3.1) implies that

$$\sum_a\pi^*_{ia}\cdot\{\sum_j(\delta_{ij}-p_{iaj})y^*_j\} = \sum_a\pi^*_{ia}\cdot\sum_b r_{iab}\rho^*_{ib}, \qquad i\in E.$$

Hence,

$$(I-P(\pi^*))y^* = r(\pi^*,\rho^*),$$

implying that

$$y^* = (I-P(\pi^*))^{-1}r(\pi^*,\rho^*) = v((\pi^*)^\infty,(\rho^*)^\infty).$$

Analogously to theorem 3.4.2, we can obtain

$$x^*_{ia} = [\beta^T(I-P(\pi^*))^{-1}]_i \cdot \pi^*_{ia}, \quad a \in A(i),\ i \in E.$$

Since the optima of (6.2.1) and (6.2.2) are equal, we get

$$\sum_j \beta_j v_j((\pi^*)^\infty,(\rho^*)^\infty) = \sum_j \beta_j y^*_j = \sum_i z^*_i \le$$

$$\sum_i \sum_a \sum_b r_{iab}\rho_{ib} x^*_{ia} = \sum_j \beta_j v_j((\pi^*)^\infty,\rho^\infty) \quad \text{for every } \rho^\infty.$$

Hence, $(\rho^*)^\infty$ is a stationary optimal policy in the Markov decision problem corresponding to policy $(\pi^*)^\infty$ for player I. Consequently,

$$(6.2.4) \quad y^* \le v((\pi^*)^\infty,\rho^\infty) \quad \text{for every stationary policy } \rho^\infty.$$

Since $sup_{R_1} v(R_1,(\rho^*)^\infty) = sup_\pi v(\pi^\infty,(\rho^*)^\infty)$ and $inf_{R_2} v((\pi^*)^\infty,R_2) = inf_\rho v((\pi^*)^\infty,\rho^\infty)$ (see remark 6.1.1), it follows from (6.2.3) and (6.2.4) that

$$v(R_1,(\rho^*)^\infty) \le v((\pi^*)^\infty,(\rho^*)^\infty) \le v((\pi^*)^\infty,R_2) \quad \text{for all } R_1,R_2,$$

i.e. $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for player I and player II respectively. □

<u>REMARK 6.2.1</u>. Since the optimal policies and the value of the game are obtained as optimal solutions of the linear programs (6.2.1) and (6.2.2), the components of the value of the game as well as the components of the optimal decision rules belong to the algebraic field generated by the rewards and the transition probabilities. This result is also shown by PARTHASARATHY & RAGHAVAN [1978].

REMARK 6.2.2. In this remark we will show that the optimality of the policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$, which were defined in theorem 6.2.3, can also be established without the use of theorem 6.2.1. Then, we have a constructive proof for the existence of the value of the game and the existence of stationary optimal policies for the two players. This proof only needs results from the theory of linear programming and the theory of Markov decision processes. Consider the linear programming problem (6.2.2). By theorem 3.4.8,

$$P := \left\{ x \left| \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j, & j \in E \\ x_{ia} \geq 0, & a \in A(i),\ i \in E \end{array} \right. \right\}$$

is feasible and bounded, and it follows from the constraints of problem (6.2.2) that this linear program has a finite optimal solution. Again, let (y^*, ρ^*) and (x^*, z^*) be optimal solutions of the linear programs (6.2.1) and (6.2.2) respectively. Similarly as in the proof of theorem 6.2.3 it can be shown that

$$v(R_1, (\rho^*)^\infty) \leq y^* = v((\pi^*)^\infty, (\rho^*)^\infty) \leq v((\pi^*)^\infty, R_2)$$

for all policies R_1, R_2, i.e. $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies and y^* is the value of the game.

ALGORITHM XXVII for the construction of the value of the game and of stationary optimal policies for the two players in a contracting TMG-model in which one player controls the transition probabilities.

step 1: Choose the numbers β_j such that $\beta_j > 0$, $j \in E$.
step 2: Compute optimal solutions (y^*, ρ^*) and (x^*, z^*) of the pair of dual linear programming problems

$$\min \left\{ \sum_j \beta_j y_j \left| \begin{array}{lll} \sum_j (\delta_{ij} - p_{iaj}) y_j - \sum_b r_{iab} \rho_{ib} & \geq 0, & a \in A(i),\ i \in E \\ \sum_b \rho_{ib} & = 1, & i \in E \\ \rho_{ib} & \geq 0, & b \in B(i),\ i \in E \end{array} \right. \right\}$$

and

$$\max\left\{\sum_i z_i \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = \beta_j, & & j \in E \\ -\sum_a r_{iab}x_{ia} + z_i \le 0, & & b \in B(i),\ i \in E \\ x_{ia} \ge 0, & & a \in A(i),\ i \in E \end{array}\right\}$$

respectively.

<u>step 3</u>: val(TMG) := y^*;
$(\rho^*)^\infty$ is an optimal policy for player II;
$(\pi^*)^\infty$, where $\pi^*_{ia} := x^*_{ia}/\Sigma_a x^*_{ia}$, $a \in A(i)$, $i \in E$, is an optimal policy for player I.

For any stationary policy π^∞ for player I, we define

$$x_{ia}(\pi) := [\beta^T(I-P(\pi))^{-1}]_i \cdot \pi_{ia} \quad , \qquad a \in A(i),\ i \in E,$$

$$z_i(\pi) := \min_{b\in B(i)} r_{ib}(\pi) \cdot \sum_a x_{ia}(\pi), \qquad i \in E.$$

The relation between the stationary policies and the feasible solutions of program (6.2.2) is given in the following theorem.

<u>THEOREM 6.2.4</u>.

(i) $(x(\pi),z(\pi))$ *is a feasible solution of the linear programming problem* (6.2.2) *with*

$$\sum_i z_i(\pi) := \min_\rho \sum_j \beta_j v_j(\pi^\infty,\rho^\infty).$$

(ii) *If* (x,z) *is a feasible solution of problem* (6.2.2), *then* $x = x(\pi)$ *and* $z \le z(\pi)$, *where*

$$\pi_{ia} := x_{ia}\Big/\sum_a x_{ia}\ , \qquad a \in A(i),\ i \in E.$$

<u>PROOF</u>.

(i) Theorem 3.4.2 implies that $\Sigma_i\Sigma_a(\delta_{ij}-p_{iaj})x_{ia}(\pi) = \beta_j$, $j \in E$, and $x_{ia}(\pi) \ge 0$, $a \in A(i)$, $i \in E$. Furthermore, we have

$$z_i(\pi) \le r_{ib}(\pi)\cdot\sum_a x_{ia}(\pi) = \sum_a r_{iab}x_{ia}(\pi), \quad b \in B(i),\ i \in E.$$

Hence, $(x(\pi),z(\pi))$ is a feasible solution of program (6.2.2). Let ρ^∞ be any stationary policy for player II. Then, we can write

$$(6.2.5) \qquad \sum_i z_i(\pi) = \sum_i(\sum_b \rho_{ib})z_i(\pi)$$
$$\le \sum_i\sum_b\sum_a r_{iab}\rho_{ib}x_{ia}(\pi) = \sum_j \beta_j v_j(\pi^\infty,\rho^\infty).$$

Define the stationary policy $\tilde{\rho}^{\infty}$ by

$$\tilde{\rho}_{ib} := \begin{cases} 1 & b = b_i \\ 0 & b \neq b_i, \end{cases}$$

where b_i satisfies

$$z_i(\pi) = r_{ib_i}(\pi)\cdot\sum_a x_{ia}(\pi), \qquad i \in E.$$

Thus,

$$(6.2.6) \qquad \sum_i z_i(\pi) = \sum_i\sum_b\sum_a r_{ib}(\pi)\cdot\tilde{\rho}_{ib}\cdot x_{ia}(\pi) = \sum_j \beta_j v_j(\pi^{\infty},\tilde{\rho}^{\infty}).$$

From (6.2.5) and (6.2.6), it follows that

$$\sum_i z_i(\pi) = \min_\rho \sum_j \beta_j v_j(\pi^{\infty},\rho^{\infty}).$$

(ii) Let (x,z) be any feasible solution of problem (6.2.2). Theorem 3.4.2 implies that $x = x(\pi)$. Hence, z satisfies

$$z_i \le \sum_a r_{iab} x_{ia}(\pi) = \sum_a r_{iab}\pi_{ia}\cdot\sum_a x_{ia}(\pi) = r_{ib}(\pi)\cdot\sum_a x_{ia}(\pi)$$

for every $b \in B(i)$ and $i \in E$. Consequently,

$$z_i \le \min_{b\in B(i)} r_{ib}(\pi)\cdot\sum_a x_{ia}(\pi) = z_i(\pi), \qquad i \in E,$$

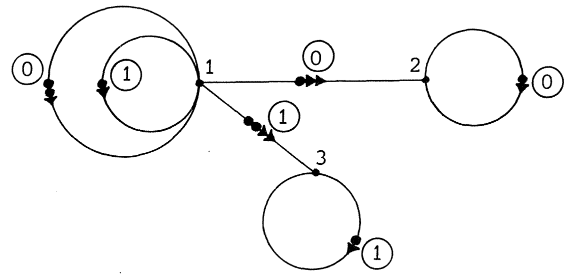
which completes the proof of the theorem. □

<u>REMARK 6.2.3</u>. The linear programming approach is also applicable to solve the *constrained Markov game*, where the constraints are imposed on the expected state-action frequencies for the player who controls the transitions. The analysis is analogous to the treatment of the constrained Markovian decision problem of section 3.3 (cf. HORDIJK & KALLENBERG [1981e]).

6.3. AVERAGE REWARD CRITERION

In this section we deal with the AMG-model. As in chapter 4, we assume that the summation of the transition probabilities equals 1, i.e. $\Sigma_j p_{iabj}=1$ for every $i \in E$, $a \in A(i)$, $b \in B(i)$. In the AMG-model, in general, there do not exist stationary optimal policies as shown in the following example due to GILLETTE [1957].

EXAMPLE 6.3.1. Suppose that the AMG-model corresponding to figure 6.3.1 has stationary optimal policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$ for player I and player II respectively. Then,



$$\phi(\pi^\infty,(\rho^*)^\infty) \le \phi((\pi^*)^\infty,(\rho^*)^\infty) \le \phi((\pi^*)^\infty,\rho^\infty)$$

Figure 6.3.1

for all stationary policies π^∞ and ρ^∞. Hence (cf. corollary 6.1.1),

$$sup_\pi inf_\rho \phi(\pi^\infty,\rho^\infty) = inf_\rho sup_\pi \phi(\pi^\infty,\rho^\infty).$$

However, it can be verified that the model of figure 6.3.1 satisfies

$$\tfrac{1}{2} = sup_\pi inf_\rho \phi_1(\pi^\infty,\rho^\infty) < inf_\rho sup_\pi \phi_1(\pi^\infty,\rho^\infty) = 1.$$

REMARK 6.3.1. BLACKWELL & FERGUSON [1968] have shown that for the model of figure 6.3.1

$$sup_{R_1} inf_{R_2} \phi_1(R_1,R_2) = inf_{R_2} sup_{R_1} \phi_1(R_1,R_2) = \tfrac{1}{2}.$$

Moreover, they have proved that there do not exist optimal policies for the two players; only player II has an optimal policy. Recently, MONASH [1979] and MERTENS & NEYMAN [1980] have shown that any AMG-model satisfies the minimax theorem, i.e.

$$sup_{R_1} inf_{R_2} \phi(R_1,R_2) = inf_{R_2} sup_{R_1} \phi(R_1,R_2).$$

DEFINITION 6.3.1. A vector $\psi \in \mathbb{R}^N$ is said to be *AMG-superharmonic* if there exist a vector $t \in \mathbb{R}^N$ and a stationary policy ρ^∞ for player II such that

$$\psi_i \ge \textstyle\sum_j p_{iaj}(\rho)\psi_j \quad , \quad a \in A(i),\ i \in E,$$

$$\psi_i + t_i \ge r_{ia}(\rho) + \textstyle\sum_j p_{iaj}(\rho)t_j, \quad a \in A(i),\ i \in E.$$

THEOREM 6.3.1. *If there exist stationary optimal policies for both players, then val(AMG) is the smallest AMG-superharmonic vector.*

PROOF. Suppose that $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for the two players. Since player II uses a stationary policy, the AMG-model may be interpreted as an AMD-model with rewards $r_{ia}(\rho^*)$ and transition probabilities $p_{iaj}(\rho^*)$ (see remark 6.1.1). Because $(\rho^*)^\infty$ is an optimal policy for player II, we have furthermore, $sup_{R_1}\phi(R_1,(\rho^*)^\infty) = \text{val(AMG)}$. Consequently, val(AMG) is the AMD-value-vector in the corresponding AMD-model. Theorem 4.2.2 implies that val(AMG) is AMD-superharmonic. Hence, val(AMG) is also AMG-superharmonic with corresponding stationary policy $(\rho^*)^\infty$ for player II.

Suppose that ψ is also AMG-superharmonic with corresponding vector t and policy ρ^∞. Then, definition 6.3.1 implies that

$$\psi \geq P^*(\pi^*,\rho)\psi \quad \text{and} \quad \psi \geq r(\pi^*,\rho) + (I-P(\pi^*,\rho))t.$$

Hence, we get

$$\psi \geq P^*(\pi^*,\rho)\{r(\pi^*,\rho) + (I-P(\pi^*,\rho))t\} = P^*(\pi^*,\rho)r(\pi^*,\rho) = \phi((\pi^*)^\infty,\rho^\infty).$$

Since the policy $(\pi^*)^\infty$ is optimal for player I, it follows that

$$\psi \geq \phi((\pi^*)^\infty,\rho^\infty) \geq \phi((\pi^*)^\infty,(\rho^*)^\infty) = \text{val(AMG)},$$

i.e. val(AMG) is the smallest AMG-superharmonic vector. □

From theorem 6.3.1 it follows that, if there are stationary optimal policies for both players, then the value of the game can be computed as the optimal solution of the following mathematical programming problem

$$min\left\{\sum_j \beta_j\psi_j \;\middle|\; \begin{array}{lll} \sum_j(\delta_{ij}-\sum_b p_{iabj}\rho_{ib})\psi_j & \geq 0, & a\in A(i),\ i\in E \\ \psi_i+\sum_j(\delta_{ij}-\sum_b p_{iabj}\rho_{ib})t_j-\sum_b r_{iab}\rho_{ib} & \geq 0, & a\in A(i),\ i\in E \\ \sum_b\rho_{ib} & = 1, & i\in E \\ \rho_{ib} & \geq 0, & b\in B(i),\ i\in E \end{array}\right\},$$

where $\beta_j > 0$, $j \in E$, are given numbers.

REMARK 6.3.2. In BEWLEY & KOHLBERG [1978] sufficient conditions can be found for the existence of stationary optimal policies in an AMG-model.

An example of such a condition is the case that assumption 6.3.1, which is stated below, is satisfied.

Since we are interested in the computation of stationary optimal policies by linear programming, we have to impose an assumption to our model. Similarly as in the previous section for the TMG-model, we will assume that the transition probabilities depend only on the maximizing player. The following assumption holds for the remaining part of this section.

ASSUMPTION 6.3.1. *The transition probabilities* p_{iabj}, $j \in E$, *do not depend on* b *for all* $i \in E$, $a \in A(i)$.

We will denote the transition probabilities p_{iabj} by p_{iaj} and the transition matrix $P(\pi,\rho)$ by $P(\pi)$. Theorem 6.3.1, remark 6.3.2 and assumption 6.3.1 imply that val(AMG) can be found as the optimal solution of the following linear programming problem:

$$(6.3.1)\qquad \min\left\{\sum_j \beta_j \psi_j \;\middle|\; \begin{array}{lll} \sum_j (\delta_{ij}-p_{iaj})\psi_j & \geq 0, & a\in A(i), i\in E \\ \psi_i + \sum_j (\delta_{ij}-p_{iaj}) t_j - \sum_b r_{iab}\rho_{ib} & \geq 0, & a\in A(i), i\in E \\ \sum_b \rho_{ib} & = 1, & i\in E \\ \rho_{ib} & \geq 0, & b\in B(i), i\in E \end{array}\right\}.$$

The dual linear programming problem is

$$(6.3.2)\qquad \max\left\{\sum_i z_i \;\middle|\; \begin{array}{lll} \sum_i\sum_a (\delta_{ij}-p_{iaj}) x_{ia} & = 0, & j\in E \\ \sum_a x_{ja} + \sum_i\sum_a (\delta_{ij}-p_{iaj}) y_{ia} & = \beta_j, & j\in E \\ -\sum_a r_{iab} x_{ia} + z_i & \leq 0, & b\in B(i), i\in E \\ x_{ia}, y_{ia} & \geq 0, & a\in A(i), i\in E \end{array}\right\}.$$

THEOREM 6.3.2. *Let* (ψ^*, t^*, ρ^*) *and* (x^*, y^*, z^*) *be optimal solutions of the linear programming problems* (6.3.1) *and* (6.3.2) *respectively. Define the stationary policy* $(\pi^*)^\infty$ *by*

$$\pi^*_{ia} := \begin{cases} x^*_{ia} \Big/ \sum_a x^*_{ia}, & a \in A(i),\ i \in E_{x^*} \\ y^*_{ia} \Big/ \sum_a y^*_{ia}, & a \in A(i),\ i \in E \backslash E_{x^*}. \end{cases}$$

Then, $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for player I and player II respectively, and ψ^* is the value of the game.

PROOF. From theorem 6.3.1 and BEWLEY & KOHLBERG [1978] it follows that ψ^* is the value of the game. The constraints of program (6.3.2) imply that

$$\sum_a x^*_{ja} + \sum_a y^*_{ja} = \beta_j + \sum_i \sum_a p_{iaj} y^*_{ia} \geq \beta_j > 0, \quad j \in E.$$

Hence, the policy $(\pi^*)^\infty$ is well-defined. The constraints of program (6.3.1) imply for any policy π^∞

$$\psi^* \geq P(\pi)\psi^* \quad \text{and} \quad \psi^* \geq r(\pi,\rho^*) - (I-P(\pi))t^*.$$

Therefore, we obtain

$$(6.3.3) \quad \psi^* \geq P^*(\pi)\psi^* \geq P^*(\pi)r(\pi,\rho^*) - P^*(\pi)(I-P(\pi))t^* = \phi(\pi^\infty,(\rho^*)^\infty)$$

for any policy π^∞. Since $\pi^*_{ia} > 0$ if and only if

$$\begin{cases} x^*_{ia} > 0 & \text{for } a \in A(i) \text{ and } i \in E_{x^*} \\ y^*_{ia} > 0 & \text{for } a \in A(i) \text{ and } i \in E\backslash E_{x^*}, \end{cases}$$

if follows from the complementary slackness property of linear programming (cf. corollary 1.3.1), that

$$\sum_a \pi^*_{ia} \cdot \{\sum_j (\delta_{ij} - p_{iaj})\psi^*_j\} = 0, \quad i \in E\backslash E_{x^*},$$

$$\sum_a \pi^*_{ia} \cdot \{\psi^*_i + \sum_j (\delta_{ij} - p_{iaj})t^*_j - \sum_b r_{iab}\rho^*_{ib}\} = 0, \quad i \in E_{x^*}.$$

Suppose that

$$\pi^*_{ka_k} \cdot \{\sum_j (\delta_{kj} - p_{ka_kj})\psi^*_j\} \neq 0$$

for some $k \in E_{x^*}$ and $a_k \in A(k)$. Then, the definition of π^* and the constraints of program (6.3.1) imply that

$$x^*_{ka_k} \cdot \{\sum_j (\delta_{kj} - p_{ka_kj})\psi^*_j\} > 0.$$

Hence, we get

$$\sum_i \sum_a x^*_{ia} \cdot \{\sum_j (\delta_{ij} - p_{iaj}) \psi^*_j\} > 0,$$

which is contradictory to

$$\sum_i \sum_a x^*_{ia} \cdot \{\sum_j (\delta_{ij} - p_{iaj}) \psi^*_j\} =$$

$$= \sum_j \{\sum_i \sum_a (\delta_{ij} - p_{iaj}) x^*_{ia}\} \psi^*_j = 0.$$

Therefore, we have

$$\begin{cases} \sum_a \pi^*_{ia} \cdot \{\sum_j (\delta_{ij} - p_{iaj}) \psi^*_j\} = 0, & i \in E, \\ \sum_a \pi^*_{ia} \cdot \{\psi^*_i + \sum_j (\delta_{ij} - p_{iaj}) t^*_j - \sum_b r_{iab} \rho^*_{ib}\} = 0, & i \in E_{x^*}, \end{cases}$$

i.e.

$$\begin{cases} [(I - P(\pi^*)) \psi^*]_i = 0, & i \in E, \\ \psi^*_i + [(I - P(\pi^*)) t^*]_i = r_i(\pi^*, \rho^*), & i \in E_{x^*}. \end{cases}$$

Since E_{x^*} is the set of recurrent states in the Markov chain induced by $P(\pi^*)$ (see proposition 4.3.3), we obtain

$$(6.3.4) \qquad \psi^* = P^*(\pi^*) \psi^* = P^*(\pi^*) r(\pi^*, \rho^*) = \phi((\pi^*)^\infty, (\rho^*)^\infty).$$

Let $x^*_i := \Sigma_a x^*_{ia}$, $i \in E$. Suppose that $E_1, E_2, \ldots, E_m$ are the ergodic sets and that F is the set of transient states in the Markov chain induced by $P(\pi^*)$. Let $n_k := |E_k|$, $k = 1,2,\ldots,m$. Then, we shall show that

$$(x^*)^T = \gamma^T P^*(\pi^*),$$

for certain vector $\gamma >> 0$, where

$$(6.3.5) \qquad \gamma_\ell := \begin{cases} \frac{1}{n}, & \ell \in F \\ \frac{1}{n_k} \sum_{j \in E_k} \{x^*_j - \frac{1}{n} \sum_{i \in F} p^*_{ij}(\pi^*)\}, & \ell \in E_k,\ k = 1,2,\ldots,m \end{cases}$$

(choose n sufficiently large such that $\gamma >> 0$).
Then, definition (6.3.5) implies that

(6.3.6) $$\sum_\ell \sum_{j\in E_k} \gamma_\ell p^*_{\ell j}(\pi^*) =$$

$$= \sum_{\ell\in F}\sum_{j\in E_k} \gamma_\ell p^*_{\ell j}(\pi^*) + \sum_{i\in E_k}\sum_{j\in E_k} \gamma_i p^*_{ij}(\pi^*)$$

$$= \frac{1}{n}\sum_{\ell\in F}\sum_{j\in E_k} p^*_{\ell j}(\pi^*) + \sum_{i\in E_k}\gamma_i$$

$$= \frac{1}{n}\sum_{\ell\in F}\sum_{j\in E_k} p^*_{\ell j}(\pi^*) + \sum_{j\in E_k}\{x^*_j - \frac{1}{n}\sum_{\ell\in F} p^*_{\ell j}(\pi^*)\}$$

$$= \sum_{j\in E_k} x^*_j, \quad k = 1,2,\dots,m.$$

From program (6.3.2) and the definition of π^* it follows that $(x^*)^T = (x^*)^T P(\pi^*)$ and, consequently, $(x^*)^T = (x^*)^T P^*(\pi^*)$. Because, by proposition 4.3.3, $x^*_i = 0$ for all $i \in F$, and, by theorem 2.3.2, $p^*_{\cdot i}(\pi^*) = 0$ for all $i \in F$, we have

(6.3.7) $$x^*_i = (\gamma^T P^*(\pi^*))_i = 0, \quad i \in F.$$

For any $i \in E_k$, we obtain using (6.3.6)

(6.3.8) $$x^*_i = \sum_j x^*_j p^*_{ji}(\pi^*) = \sum_{j\in E_k} x^*_j p^*_{ji}(\pi^*) = p^*_{ii}(\pi^*)\cdot\sum_{j\in E_k} x^*_j =$$

$$= p^*_{ii}(\pi^*)\cdot\sum_\ell\sum_{j\in E_k}\gamma_\ell p^*_{\ell j}(\pi^*) = \sum_\ell \gamma_\ell\cdot\sum_{j\in E_k} p^*_{\ell j}(\pi^*)\cdot p^*_{ji}(\pi^*) =$$

$$= \sum_\ell \gamma_\ell \cdot p^*_{\ell i}(\pi^*) = (\gamma^T P^*(\pi^*))_i.$$

Hence, (6.3.7) and (6.3.8) imply that $(x^*)^T = \gamma^T P^*(\pi^*)$. Again using the complementary slackness property yields

$$\sum_i\sum_b \rho^*_{ib}(z^*_i - \sum_a r_{iab} x^*_{ia}) = 0.$$

Therefore,

(6.3.9) $$\sum_i z^*_i = \sum_i\sum_b \rho^*_{ib} z^*_i = \sum_i\sum_b\sum_a r_{iab}\rho^*_{ib}\pi^*_{ia}\cdot\sum_a x^*_{ia}$$

$$= \sum_i (\gamma^T P^*(\pi^*))_i r_i(\pi^*,\rho^*) = \gamma^T\phi((\pi^*)^\infty,(\rho^*)^\infty).$$

For any stationary policy ρ^∞ for player II, we have in view of the constraints of the linear program (6.3.2)

$$(6.3.10) \qquad \sum_i z_i^* = \sum_i \sum_b \rho_{ib} z_i^* \le \sum_i \sum_b \sum_a r_{iab} \rho_{ib} \pi_{ia}^* \cdot \sum_a x_{ia}^* = \gamma^T \phi((\pi^*)^\infty, \rho^\infty).$$

If the policy $(\pi^*)^\infty$ is used by player I, then the AMG-model may be viewed as an AMD-model (cf. remark 6.1.1). Since $\gamma >> 0$, it follows from (6.3.9), (6.3.10) and the property that an optimal policy maximizes the rewards simultaneously for all initial states, that

$$(6.3.11) \qquad \phi((\pi^*)^\infty, (\rho^*)^\infty) \le \phi((\pi^*)^\infty, \rho^\infty) \quad \text{for every stationary policy } \rho^\infty.$$

Since $sup_{R_1} \phi(R_1, (\rho^*)^\infty) = sup_\pi \phi(\pi^\infty, (\rho^*)^\infty)$ and $inf_{R_2} \phi((\pi^*)^\infty, R_2) = inf_\rho \phi((\pi^*)^\infty, \rho^\infty)$ (see remark 6.1.1), it follows from (6.3.3), (6.3.4) and (6.3.11) that

$$\phi(R_1, (\rho^*)^\infty) \le \phi((\pi^*)^\infty, (\rho^*)^\infty) \le \phi((\pi^*)^\infty, R_2)$$

for all R_1, R_2, i.e. $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for player I and player II respectively. □

<u>REMARK 6.3.3</u>. Recently, we learned that another proof of the above theorem was developed by VRIEZE [1980] at the same time.

<u>REMARK 6.3.4</u>. We can show the optimality of the stationary policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$, defined in theorem 6.3.2, without using the results of BEWLEY & KOHLBERG [1978]. This provides a constructive proof for the existence of the value of the game and of stationary optimal policies.

Consider the linear programming problem (6.3.2). Since any feasible solution (x,y,z) satisfies

$$\sum_i z_i \le \sum_i \sum_a r_{iab} x_{ia} \le M \cdot \sum_i \sum_a x_{ia} = M \cdot \sum_j \beta_j,$$

where $M := max_{i,a,b} r_{iab}$, the linear program (6.3.2) has a finite optimum. Using the results of chapter 4, it is obvious that this linear program is also feasible. Hence, the pair of dual linear programming problems (6.3.1) and (6.3.2) has finite optimal solutions, say (ψ^*, t^*, ρ^*) and (x^*, y^*, z^*) respectively. In the proof of theorem 6.3.2 we have shown that

$$\phi(R_1, (\rho^*)^\infty) \le \psi^* = \phi((\pi^*)^\infty, (\rho^*)^\infty) \le \phi((\pi^*)^\infty), R_2)$$

for all policies R_1 and R_2, i.e. $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for player I and player II respectively, and ψ^* is the value of the game.

ALGORITHM XXVIII for the construction of val(AMG) and of stationary optimal policies for the two players in an AMG-model in which one player controls the transition probabilities (multichain case).

step 1: Take the numbers β_j such that $\beta_j > 0$, $j \in E$, and $\Sigma_j \beta_j = 1$.

step 2: Compute optimal solutions (ψ^*, t^*, ρ^*) and (x^*, y^*, z^*) of the pair of dual linear programming problems

$$\min\left\{\sum_j \beta_j \psi_j \;\middle|\; \begin{array}{lll} \sum_j (\delta_{ij} - p_{iaj})\psi_j & \geq 0, & a \in A(i), i \in E \\ \psi_i + \sum_j (\delta_{ij} - p_{iaj}) t_j - \sum_b r_{iab} \rho_{ib} & \geq 0, & a \in A(i), i \in E \\ \sum_b \rho_{ib} & = 1, & i \in E \\ \rho_{ib} & \geq 0, & b \in B(i), i \in E \end{array}\right\}$$

and

$$\max\left\{\sum_i z_i \;\middle|\; \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} & = 0, & j \in E \\ \sum_a x_{ja} + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} & = \beta_j, & j \in E \\ -\sum_a r_{iab} x_{ia} + z_i & \leq 0, & b \in B(i), i \in E \\ x_{ia}, y_{ia} & \geq 0, & a \in A(i), i \in E \end{array}\right\}$$

respectively.

step 3: val(AMG) := ψ^*;

$(\rho^*)^\infty$ is an optimal policy for player II;

$$(\pi^*)^\infty, \text{ where } \pi^*_{ia} := \begin{cases} x^*_{ia} \big/ \sum_a x^*_{ia}, & a \in A(i),\ i \in E_{x^*} \\ y^*_{ia} \big/ \sum_a y^*_{ia}, & a \in A(i),\ i \in E \backslash E_{x^*}, \end{cases}$$

is an optimal policy for player I.

REMARK 6.3.5. From the linear programming approach it also follows that the value and the optimal stationary policies lie in the same ordered field as the data. This property was already shown by PARTHASARATY & RAGHAVAN [1978]. The first finite algorithm to compute stationary optimal policies was developed by FILAR & RAGHAVAN [1979]. Their algorithm seems to have a prohibitive amount of computations.

REMARK 6.3.6. Analogously to the results of section 4.7 the constrained Markov game can be solved by linear programming. An extensive treatment of this subject can be found in HORDIJK & KALLENBERG [1981f].

We close this section with the *unichain case*, i.e. when assumption 4.6.2 is satisfied. For this case we propose algorithm XXIX. In theorem 6.3.3 we will prove that this algorithm finds stationary optimal policies for both players as well as the value of the game.

ALGORITHM XXIX for the construction of val(AMG) and of stationary optimal policies for the two players in an AMG-model in which one player controls the transition probabilities (unichain case).

step 1: Compute optimal solutions (ψ^*, t^*, ρ^*) and (x^*, z^*) of the pair of dual linear programming problems

$$(6.3.12)\quad \min\left\{\psi \;\middle|\; \begin{array}{lll} \psi + \sum_j (\delta_{ij} - p_{iaj}) t_j - \sum_b r_{iab}\rho_{ib} \geq 0, & a \in A(i), & i \in E \\ \sum_b \rho_{ib} = 1, & & i \in E \\ \rho_{ib} \geq 0, & b \in B(i), & i \in E \end{array}\right\}$$

and

$$(6.3.13)\quad \max\left\{\sum_i z_i \;\middle|\; \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = 0, & & j \in E \\ \sum_i \sum_a x_{ia} = 1 & & \\ -\sum_a r_{iab} x_{ia} + z_i \leq 0, & b \in B(i), & i \in E \\ x_{ia} \geq 0, & a \in A(i), & i \in E \end{array}\right\}$$

respectively.

step 2: $\mathrm{val(AMG)} := \psi^* \cdot e$;

$(\rho^*)^\infty$ is an optimal policy for player II;

$$(\pi^*)^\infty, \text{ where } \pi^*_{ia} := \begin{cases} x^*_{ia} \Big/ \sum_a x^*_{ia}, & a \in A(i),\ i \in E_{x^*} \\ \text{arbitrarily}, & a \in A(i),\ i \in E \backslash E_{x^*}, \end{cases}$$

is an optimal policy for player I.

THEOREM 6.3.3. *Suppose that we have a unichained AMG-model. Then, algorithm XXIX gives the value of the game as well as stationary optimal policies for the two players.*

PROOF. Lemma 4.6.1 together with theorem 2.3.2 imply that for any stationary policy π^∞ the stationary matrix $P^*(\pi)$ has identical rows. Hence, val(AMG) has identical components. Then, by theorem 6.3.1, val(AMG) is the optimal solution of the linear programming problem (6.3.12). Moreover, we have

$$(x^*)^T(I-P(\pi^*)) = 0 \quad \text{and} \quad (x^*)^T e = 1.$$

Consequently,

$$(6.3.14) \qquad x^* = p^*(\pi^*),$$

where $p^*(\pi^*)$ is the vector corresponding to the identical rows of $P^*(\pi^*)$. From the constraints of program (6.3.12) it follows that

$$(6.3.15) \qquad \psi^*\cdot e \geq \phi(\pi^\infty,(\rho^*)^\infty) \quad \text{for every stationary policy } \pi^\infty.$$

By the complementary slackness property it holds that

$$(6.3.16) \qquad \textstyle\sum_i z_i^* = \sum_i\sum_a\sum_b r_{iab}\rho_{ib}^*\pi_{ia}^*\cdot x_i^* = (p^*(\pi^*))^T r(\pi^*,\rho^*) =$$

$$= \phi_j((\pi^*)^\infty,(\rho^*)^\infty) \qquad \text{for every } j \in E.$$

Then, by theorem 1.3.4, we obtain

$$(6.3.17) \qquad \psi^* = \textstyle\sum_i z_i^* = \phi_j((\pi^*)^\infty,(\rho^*)^\infty) \quad \text{for every } j \in E.$$

The constraints of program (6.3.13) imply that

$$(6.3.18) \qquad \textstyle\sum_i z_i^* \leq \sum_i\sum_a\sum_b r_{iab}\rho_{ib}\pi_{ia}^*\cdot x_i^* = (p^*(\pi^*))^T r(\pi^*,\rho) =$$

$$\phi_j((\pi^*)^\infty,\rho^\infty) \quad \text{for every stationary policy } \rho^\infty \text{ and } j \in E.$$

Combining (6.3.15), (6.3.17) and (6.3.18) yields

$$\phi(\pi^\infty,(\rho^*)^\infty) \leq \psi^*\cdot e = \phi((\pi^*)^\infty,(\rho^*)^\infty) \leq \phi((\pi^*)^\infty,\rho^\infty)$$

for all stationary policies π^∞ and ρ^∞. Then, using remark 6.1.1, it follows that $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies and that $\psi^*\cdot e$ is the value of the game. □

CHAPTER 7

SEMI-MARKOV DECISION PROCESSES

7.1. INTRODUCTION AND SUMMARY

In this chapter we shall investigate the *semi-Markov decision process* which was introduced by DE CANI [1964], HOWARD [1963], JEWELL [1963a], [1963b] and SCHWEITZER [1965]. In the discrete Markov decision model that was studied in the preceding chapters, the decision time points were equidistant. In the semi-Markov model, the times between the decision time points will be random variables. We can describe the semi-Markov decision model in the following way.

Consider a dynamic system that is observed at decision time points t, starting at t = 0. At each decision time point the system is in one of a finite number of states and an action has to be chosen. Let $E = \{1,2,\dots,N\}$ be the *state space* and $A(i)$ the finite set of possible *actions* in state i, $i \in E$. If the system is in state i and action $a \in A(i)$ is chosen, then the following occurs independently of the history of the process:

1. The next state of the process is chosen according to the *transition probabilities* p_{iaj}, where $p_{iaj} \geq 0$ and $\Sigma_j p_{iaj} = 1$ for every $a \in A(i)$ and $i,j \in E$.
2. Conditional on the event that the next state is j, the *sojourn time* t_{iaj} until the next decision time point is a random variable with probability distribution $F_{iaj}(t)$, i.e. $F_{iaj}(t) = \mathbb{P}(t_{iaj} \leq t)$.
3. A reward r_{ia} is earned immediately and, in addition, a reward rate s_{ia} is imposed until the next transition occurs, i.e. if the next decision time point falls after t_{ia} units of times, then the *reward* in this epoch is given by $r_{ia} + t_{ia} \cdot s_{ia}$.

A semi-Markov decision process is also called a *Markov renewal program.*

A *policy* R is a sequence of decision rules: $R = (\pi^1,\pi^2,\dots)$, where π^n denotes the decision rule for the n-th decision time point. This decision rule may depend on the whole history of the process, i.e. on the observed states $\{X_1,X_2,\ldots,X_n\}$ and the chosen actions $\{Y_1,Y_2,\ldots,Y_{n-1}\}$. A policy is called *stationary* if the chosen action only depends on the state of the process; if this choice is nonrandomized, then the policy is said to be *pure and stationary*. Similarly as for the Markov decision model, we denote by C, C_S and C_D the set of all policies, stationary policies and pure and stationary policies, respectively.

In section 7.2 we discuss the expected *discounted* reward criterion. We introduce for this model the concept of superharmonicity and we prove that the reward vector of an optimal policy is the smallest superharmonic vector. We can compute this vector as optimal solution of a linear program. Furthermore, we will show that the complementary slackness property of linear programming provides an optimal policy from the optimal solution of the dual program. Moreover, this dual program will give the equivalence between the semi-Markov model and a contracting TMD-model. Hence, also for the semi-Markov model we may apply the results shown in section 3.4 as

- one-to-one correspondence between stationary policies and feasible solutions of the dual program
- policy improvement
- elimination of suboptimal actions.

Some of the above observations were already presented in WESSELS & VAN NUNEN [1975]. However, their analysis was based on the correspondence between stationary policies and feasible solutions of the dual program. In our treatment the results are consequences of the concept of superharmonicity.

Section 7.3 deals with the *undiscounted* rewards. Also for this model we can present the property of superharmonicity. Using DENARDO [1971], we shall show that the reward vector of an optimal policy is the smallest superharmonic vector. Similarly as in chapter 4, we can formulate a linear program such that a pure and stationary optimal policy can be obtained directly from the optimal solution of the linear program. This linear program was also used by DENARDO & FOX [1968], but they did not show how an optimal policy can be found. The linear programming problem can be transformed into the linear program which was derived for the AMD-model. The transformations are the same as proposed by SCHWEITZER [1971]. By these transformations we show that the semi-Markov model with the average reward criterion is equivalent to an AMD-model.

We close the chapter by the presentation of simplified algorithms for the weak unichain case, the unichain case and the completely ergodic case.

7.2. DISCOUNTED REWARDS

Let $\alpha \in (0,1)$ be any discount factor. Then, $\alpha^t = e^{-\lambda t}$, where $t \in \mathbb{R}^1$ and $\lambda := -\ell n\alpha$. Throughout this section we have the following assumption.

ASSUMPTION 7.2.1. $\int_0^\infty e^{-\lambda t} dF_{iaj}(t) < 1$ *for every* $i,j \in E$ *and* $a \in A(i)$.

REMARK 7.2.1. Assumption 7.2.1 guarantees that the probability distributions $F_{iaj}(t)$ are not degenerated in $t = 0$. Consequently, the expected number of transitions in a finite interval is finite. Furthermore, DENARDO [1967] has shown that the discounted Markov renewal program with assumption 7.2.1 possesses the contraction property. We shall call this model a *DRD-model*.

For any policy R and any initial state i, we define the *expected discounted reward* $v_i^\lambda(R)$ by

$$(7.2.1) \qquad v_i^\lambda(R) := \mathbb{E}_R[\textstyle\sum_{n=1}^\infty e^{-\lambda(T_1+T_2+\dots+T_{n-1})} \cdot \{r_{X_nY_n} + s_{X_nY_n} \displaystyle\int_0^{T_n} e^{-\lambda t} dt\} \mid X_1 = i],$$

where $T_1+T_2+\dots+T_{n-1} := 0$ for $n = 1$.

LEMMA 7.2.1.

$$v_i^\lambda(R) = \textstyle\sum_{n=1}^\infty \sum_j \sum_a \displaystyle\int_0^\infty r_{ja}^* e^{-\lambda t} d\pi_{iaj}(n,t,R), \qquad i \in E,\ R \in C,$$

where

$$r_{ja}^* := r_{ja} + s_{ja} \textstyle\sum_k p_{jak} \displaystyle\int_0^\infty \int_0^\tau e^{-\lambda t} dt dF_{jak}(\tau)$$

and

$$\pi_{iaj}(n,t,R) := \mathbb{P}_R(X_n = j,\ Y_n = a,\ T_1+T_2+\dots+T_{n-1} \le t \mid X_1 = i).$$

PROOF. First, we remark that

$$\mathbb{E}_R[r_{X_nY_n} + s_{X_nY_n} \int_0^{T_n} e^{-\lambda t} dt \mid X_n = j,\ Y_n = a] =$$

$$= \textstyle\sum_k \mathbb{E}_R[r_{ja} + s_{ja} \displaystyle\int_0^{T_n} e^{-\lambda t} dt \mid X_n = j,\ Y_n = a,\ X_{n+1} = k] \cdot \mathbb{P}_R[X_{n+1} = k \mid X_n = j,\ Y_n = a]$$

$$= \textstyle\sum_k p_{jak}\{r_{ja} + s_{ja} \displaystyle\int_0^\infty \int_0^\tau e^{-\lambda t} dt dF_{jak}(\tau)\} = r_{ja}^*.$$

Since the random variables $T_1+T_2+\ldots+T_{n-1}$ and T_n are conditional independent, given X_n and Y_n, we obtain

$$\mathbb{E}_R[e^{-\lambda(T_1+T_2+\ldots+T_{n-1})}\cdot\{r_{X_nY_n}+s_{X_nY_n}\int_0^{T_n} e^{-\lambda t}dt\}|X_1 = i] =$$

$$= \sum_j\sum_a \int_0^\infty e^{-\lambda t}\cdot r^*_{ja}\cdot d\mathbb{P}_R(X_n = j, Y_n = a, T_1+T_2+\ldots+T_{n-1} \le t|X_1 = i)$$

$$= \sum_j\sum_a r^*_{ja} \int_0^\infty e^{-\lambda t}d\pi_{iaj}(n,t,R).$$

$\int_0^\infty e^{-\lambda t}d\pi_{iaj}(n,t,R)$ may be interpreted as the expected discounted probability that $X_n = j$, $Y_n = a$, given $X_1 = i$. We have the recursion

$$(7.2.2)\qquad \sum_a \int_0^\infty e^{-\lambda t}d\pi_{iaj}(n,t,R) = \sum_\ell\sum_b \int_0^\infty e^{-\lambda t}d\pi_{ib\ell}(n-1,t,R)\cdot p_{\ell bj}\int_0^\infty e^{-\lambda s}dF_{\ell bj}(s).$$

We define

$$(7.2.3)\qquad w_n := \sum_j\sum_a \int_0^\infty e^{-\lambda t}d\pi_{iaj}(n,t,R), \quad n \in \mathbb{N},$$

$$(7.2.4)\qquad M := max_{i,a}\{r_{ia} + \lambda^{-1}\cdot s_{ia}\},$$

$$(7.2.5)\qquad \rho := max_{i,a,j}\int_0^\infty e^{-\lambda t}dF_{iaj}(t).$$

Then, (7.2.2), (7.2.3) and (7.2.5) imply that

$$w_n = \sum_\ell\sum_b \int_0^\infty e^{-\lambda t}d\pi_{ib\ell}(n-1,t,R)\cdot\sum_j p_{\ell bj}\int_0^\infty e^{-\lambda s}dF_{\ell bj}(s)$$

$$\le \sum_\ell\sum_b \int_0^\infty e^{-\lambda t}d\pi_{ib\ell}(n-1,t,R)\cdot\rho =$$

$$= \rho\, w_{n-1} \le \ldots \le \rho^{n-1}\cdot w_1 = \rho^{n-1}.$$

Furthermore, we have

$$r^*_{ja} \le |r_{ja}| + |s_{ja}| \cdot \sum_k p_{jak} \int_0^\infty \lambda^{-1}(1-e^{-\lambda\tau})\,dF_{jak}(\tau)$$

$$\le |r_{ja}| + \lambda^{-1}|s_{ja}| \cdot \sum_k p_{jak} \int_0^\infty dF_{jak}(\tau) \le M.$$

Consequently,

$$\sum_{n=1}^\infty \sum_j \sum_a \int_0^\infty |r^*_{ja}| e^{-\lambda t} d\pi_{iaj}(n,t,R) \le \sum_{n=1}^\infty M w_n \le \frac{M}{1-\rho} < \infty.$$

Hence,

$$v_i^\lambda(R) = \mathbb{E}_R[\sum_{n=1}^\infty e^{-\lambda(T_1+T_2+\ldots+T_{n-1})} \cdot \{r_{X_nY_n} + s_{X_nY_n} \int_0^{T_n} e^{-\lambda t} dt\} | X_1 = i]$$

$$= \sum_{n=1}^\infty \mathbb{E}_R[e^{-\lambda(T_1+T_2+\ldots+T_{n-1})} \cdot \{r_{X_nY_n} + s_{X_nY_n} \int_0^{T_n} e^{-\lambda t} dt\} | X_1 = i]$$

$$= \sum_{n=1}^\infty \sum_j \sum_a \int_0^\infty r^*_{ja} e^{-\lambda t} d\pi_{iaj}(n,t,R), \quad i \in E,\ R \in C. \qquad \square$$

NOTATION 7.2.1. We will denote the DRD-model by the five-tuple (E,A,p,r^*,F).

DEFINITION 7.2.1. The *DRD-value-vector* v^λ is defined by $v_i^\lambda := sup_R\, v_i^\lambda(R)$, $i \in E$.

From the proof of lemma 7.2.1 it follows that $|v_i^\lambda| \le \frac{M}{1-\rho}$, $i \in E$.

DEFINITION 7.2.2. A vector $\tilde{v} \in \mathbb{R}^N$ is *DRD-superharmonic* if

$$\tilde{v}_i \ge r^*_{ia} + \sum_j p_{iaj} \int_0^\infty e^{-\lambda t} dF_{iaj}(t) \cdot \tilde{v}_j, \qquad a \in A(i),\ i \in E,$$

where r^*_{ia} is defined as in lemma 7.2.1.

THEOREM 7.2.1. *The DRD-value-vector* v^λ *is the smallest DRD-superharmonic vector.*

PROOF. Choose $\varepsilon > 0$ arbitrarily. Take policy R_j such that $v_j^\lambda(R_j) \ge v_j^\lambda - \varepsilon$, $j \in E$. Let $a_i \in A(i)$ be such that

$$r^*_{ia_i} + \sum_j p^*_{ia_ij} v_j^\lambda = max_a \{r^*_{ia} + \sum_j p^*_{iaj} v_j^\lambda\}, \quad i \in E, \tag{7.2.6}$$

where

$$p^*_{iaj} := p_{iaj} \int_0^\infty e^{-\lambda t} dF_{iaj}(t).$$

We denote by $\hat{R}$ the policy that chooses at $t = 0$ action a_i, for initial state i, and then follows policy R_j, if the next state is j, while the process is considered as starting in state j. Then we obtain

$$\begin{aligned} v^\lambda_i \geq v^\lambda_i(\hat{R}) &= r^*_{ia_i} + \sum_j p^*_{ia_i j} v^\lambda_j(R_j) \\ &\geq r^*_{ia_i} + \sum_j p^*_{ia_i j} v^\lambda_j - \varepsilon \sum_j p^*_{ia_i j} \\ &\geq \max_a \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\} - \varepsilon \cdot \rho, \quad i \in E, \end{aligned}$$

where ρ is defined by (7.2.5). Since ε is arbitrarily chosen, it follows that

$$\text{(7.2.7)} \qquad v^\lambda_i \geq \max_a \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\}, \qquad i \in E,$$

i.e. v^λ is a DRD-superharmonic vector.

Next, we will show that (7.2.7) holds with equalities instead of inequalities. Let $R = (\pi^1, \pi^2, \ldots)$ be any policy. Then, we can write

$$v^\lambda_i(R) = \sum_a \pi^1_{ia} \cdot \{r^*_{ia} + \sum_j p^*_{iaj} u^\lambda_j(R)\}, \ i \in E,$$

where $u^\lambda_j(R)$ represents the expected discounted reward earned from the second decision time point, given that the state at the second decision time point is j. Therefore, $u^\lambda_j(R) \leq v^\lambda_j$, $j \in E$. Hence,

$$\begin{aligned} v^\lambda_i(R) &\leq \sum_a \pi^1_{ia} \cdot \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\} \\ &\leq \sum_a \pi^1_{ia} \cdot \max_a \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\} = \\ &\max_a \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\}, \quad i \in E. \end{aligned}$$

Since R is arbitrarily chosen, we obtain

$$\text{(7.2.8)} \qquad v^\lambda_i \leq \max_a \{r^*_{ia} + \sum_j p^*_{iaj} v^\lambda_j\}, \qquad i \in E.$$

Combining (7.2.7) and (7.2.8) yields

$$(7.2.9) \qquad v_i^\lambda = max_a \{r_{ia}^* + \sum_j p_{iaj}^* v_j^\lambda\}, \qquad i \in E.$$

Suppose that $\tilde{v}$ is also a DRD-superharmonic vector. Let a_i, $i \in E$, again satisfy (7.2.6). Then, we have

$$\tilde{v}_i - v_i^\lambda \geq r_{ia_i}^* + \sum_j p_{ia_ij}^* \tilde{v}_j - r_{ia_i}^* - \sum_j p_{ia_ij}^* v_j^\lambda = \sum_j p_{ij}(\tilde{v}_j - v_j^\lambda), \quad i \in E,$$

where $p_{ij} := p_{ia_ij}^*$ for all $i,j \in E$. Thus, we may write in vector notation

$$\tilde{v} - v^\lambda \geq P(\tilde{v}-v^\lambda) \geq \ldots \geq P^n(\tilde{v}-v^\lambda) \qquad \text{for all } n \in \mathbb{N}.$$

Using assumption 7.2.1 and (7.2.5), we obtain

$$\|P\| = max_i \sum_j p_{ia_ij}^* \leq max_i \sum_j p_{ia_ij} \cdot \rho = \rho < 1.$$

Consequently, $\lim_{n\to\infty} P^n = 0$. Hence, it follows that

$$\tilde{v} - v^\lambda \geq \lim_{n\to\infty} P^n(\tilde{v}-v^\lambda) = 0,$$

i.e. $\tilde{v} \geq v^\lambda$. This completes the proof that v^λ is the smallest DRD-superharmonic vector. □

<u>DEFINITION 7.2.3</u>. A policy R^* is said to be an *optimal policy* for the DRD-model if $v^\lambda(R^*) = v^\lambda$.

<u>THEOREM 7.2.2</u>. *Let* $a_i \in A(i)$ *satisfy*

$$r_{ia_i}^* + \sum_j p_{ia_ij}^* v_j^\lambda = v_i^\lambda, \qquad i \in E.$$

Then, the pure and stationary policy f^∞, *where* $f(i) := a_i$, $i \in E$, *is an optimal policy for the DRD-model.*

<u>PROOF</u>.

$$v_i^\lambda(f^\infty) - v_i^\lambda = r_{ia_i}^* + \sum_j p_{ia_ij}^* v_j^\lambda(f^\infty) - r_{ia_i}^* - \sum_j p_{ia_ij}^* v_j^\lambda$$

$$= \sum_j p_{ia_ij}^* (v_j^\lambda(f^\infty) - v_j^\lambda), \qquad i \in E.$$

Let $P := (p^*_{ia_ij})$. Then, similarly as in the proof of theorem 7.2.1, we obtain

$$v^\lambda(f^\infty)-v^\lambda = P(v^\lambda(f^\infty)-v^\lambda) = P^n(v^\lambda(f^\infty)-v^\lambda) \to 0 \quad \text{for } n \to \infty.$$

Consequently, $v^\lambda(f^\infty) = v^\lambda$, i.e. f^∞ is an optimal policy. □

Theorem 7.2.1 implies that the DRD-value-vector v^λ can be found as optimal solution of the linear programming problem

$$(7.2.10) \qquad min\{\sum_j \beta_j \tilde{v}_j \mid \sum_j (\delta_{ij}-p^*_{iaj})\tilde{v}_j \geq r^*_{ia}, \quad a \in A(i),\ i \in E\},$$

where $\beta_j > 0$, $j \in E$, are given numbers. The dual program of (7.2.10) is

$$(7.2.11) \qquad max\left\{\sum_i\sum_a r^*_{ia}x_{ia} \;\middle|\; \begin{array}{ll} \sum_i\sum_a(\delta_{ij}-p^*_{iaj})x_{ia} = \beta_j, & j \in E \\ x_{ia} \geq 0\ , & a \in A(i),\ i \in E \end{array}\right\}.$$

THEOREM 7.2.3. *Let x^* be an optimal solution of the linear programming problem* (7.2.11). *Then, any pure and stationary policy f^∞_* such that* $x^*_{if_*(i)} > 0$, $i \in E$, *is an optimal policy.*

PROOF. Since v^λ is the finite optimal solution of program (7.2.10), the dual program (7.2.11) has also a finite optimal solution. Let x^* be any optimal solution of program (7.2.11). Then,

$$\sum_a x^*_{ja} = \beta_j + \sum_i\sum_a p^*_{iaj}x^*_{ia} \geq \beta_j > 0, \quad j \in E.$$

The complementary slackness property of linear programming (cf. corollary 1.3.1) implies that

$$\sum_j(\delta_{ij}-p^*_{if_*(i)j})v^\lambda_j = r^*_{if_*(i)}, \qquad i \in E.$$

It follows from theorem 7.2.2 that f^∞_* is an optimal policy. □

A pure and stationary optimal policy for the DRD-model can be determined by the following algorithm.

ALGORITHM XXX for the construction of a pure and stationary optimal policy *in a discounted semi-Markov model.*

step 1: Compute

$$r^*_{ia} := r_{ia} + \lambda^{-1} s_{ia} \sum_j p_{iaj}\{1 - \int_0^\infty e^{-\lambda t} dF_{iaj}(t)\}, \quad a \in A(i),\ i \in E,$$

$$p^*_{iaj} := p_{iaj} \int_0^\infty e^{-\lambda t} dF_{iaj}(t) \quad , \quad a \in A(i),\ i,j \in E.$$

step 2: Choose the numbers β_j such that $\beta_j > 0$, $j \in E$.

step 3: Compute an optimal solution x^* of the linear program

$$\max\left\{ \sum_i \sum_a r^*_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_i \sum_a (\delta_{ij} - p^*_{iaj}) x_{ia} = \beta_j, & j \in E \\ x_{ia} \ge 0\ , & a \in A(i),\ i \in E \end{array} \right\}.$$

step 4: Take f_*^∞ such that $x^*_{if_*(i)} > 0$, $i \in E$.

REMARK 7.2.2. Consider the TMD-model (E,A,p^*,r^*). It can easily be verified that this model satisfies the contraction assumption of section 3.4 for $\mu := e$ and $\alpha := \rho$. Furthermore, algorithm IX applied on this TMD-model is identical to algorithm XXX. It can also easily be verified that $v^\lambda(\pi^\infty) = v^*(\pi^\infty)$ for every stationary policy π^∞, where $v^*(\pi^\infty)$ is the expected total reward in the TMD-model. Therefore, the TMD-model (E,A,p^*,r^*) may be considered as *equivalent* to the DRD-model (E,A,p,r^*,F), and we may apply the results of section 3.4 to the DRD-model.

REMARK 7.2.3. The above analysis is also applicable on the *two-person zero-sum semi-Markov game* in which one player controls the transition probabilities and the sojourn times. This DRG-model can be described as follows. If in state i player I chooses action $a \in A(i)$ and player II action $b \in B(i)$, then the following occurs:

1. The next state of the process is chosen according to the transition probabilities p_{iaj}.
2. Conditional on the event that the next state is j, the time t_{iaj} until the next decision time point is a random variable with probability distribution $F_{iaj}(t)$.
3. Player I receives an immediate reward r_{iab} from player II, and, in addition, player II is indebted to player I an amount $s_{iab} \cdot t_{ia}$ if the next decision time point falls after t_{ia} units of time.

If we define

$$r^*_{iab} := r_{iab}+s_{iab}\sum_j p_{iaj}\int_0^\infty\int_0^t e^{-\lambda s}dsdF_{iaj}(t), \quad a\in A(i),\ b\in B(i),\ i\in E,$$

$$p^*_{iaj} := p_{iaj}\int_0^\infty e^{-\lambda t}dF_{iaj}(t) \quad , \quad a\in A(i),\ i,j\in E,$$

then similarly to theorem 6.2.1 we can prove that there exist stationary optimal policies for both players. Moreover, it can straightforward be shown that the DRG-model (E,A,B,p,r^*,F) and the TMG-model (E,A,B,p^*,r^*) are equivalent and that algorithm XXVII applied on the TMG-model provides stationary optimal policies for both players in the DRG-model.

7.3. UNDISCOUNTED REWARDS

For any policy R and any initial state i, the *average reward per unit time* is denoted by $\chi_i(R)$ and defined by

$$\chi_i(R) := \liminf_{T\to\infty} \frac{1}{T} v_i^T(R),$$

where $v_i^T(R)$ denotes the expected undiscounted reward earned in the interval $[0,T)$. For a Markov renewal program with as utility function the average reward per unit time, we will use the name *ARD-model*. The *ARD-value-vector* χ is defined by

$$\chi_i := sup_R \chi_i(R), \quad i \in E,$$

and policy R^* is said to be *optimal for the* ARD-model if $\chi(R^*) = \chi$. A policy R_o is called a *Blackwell optimal* policy if there is a $\lambda_o > 0$ such that $v^\lambda(R_o) = v^\lambda$ for every $\lambda \in (0,\lambda_o]$.

Throughout this section we have the following assumption.

<u>ASSUMPTION 7.3.1.</u> $0 < \int_0^\infty t^2 dF_{iaj}(t) < \infty$ *for all* $a \in A(i)$, $i,j \in E$.

The above assumption implies the following results due to DENARDO [1971]:

1. Let π^∞ be any stationary policy. Then,

(7.3.1) $\quad v^\lambda(\pi^\infty) = \lambda^{-1}\chi(\pi^\infty) + w(\pi^\infty) + \varepsilon(\lambda),$

where $\lim_{\lambda\downarrow 0} \varepsilon(\lambda) = 0$.
Moreover, $\chi(\pi^\infty)$ is the unique solution of the equations

$$(7.3.2) \qquad \begin{cases} (I-P(\pi))x = 0 \\ P^*(\pi)T(\pi)x = P^*(\pi)\hat{r}(\pi), \end{cases}$$

where

$$\hat{r}_i(\pi) := \sum_a \pi_{ia} \cdot \{r_{ia} + s_{ia} \sum_j p_{iaj} \int_0^\infty t dF_{iaj}(t)\}, \quad i \in E,$$

and $T(\pi)$ is the diagonal matrix with $t_{ij}(\pi) := \delta_{ij}\tau_i(\pi)$ and

$$\tau_i(\pi) := \sum_a \pi_{ia} \cdot \sum_j p_{iaj} \int_0^\infty t dF_{iaj}(t), \quad i,j \in E.$$

Furthermore, $w(\pi^\infty)$ is a solution of the linear system

$$(7.3.3) \qquad (I-P(\pi))y = \hat{r}(\pi) - T(\pi)\chi(\pi^\infty).$$

2. There exists a Blackwell optimal pure and stationary policy.

LEMMA 7.3.1. $\liminf_{\lambda\downarrow 0} \lambda v_i^\lambda(R) \geq \chi_i(R), \quad i \in E, R \in C.$

PROOF. Since $v_i^\lambda(R) = \int_0^\infty e^{-\lambda t} dv_i^t(R)$, $i \in E$, $R \in C$, $\lambda > 0$, the proof follows from an Abelian theorem (cf. WIDDER [1946], chapter V). □

THEOREM 7.3.1. *Any pure and stationary Blackwell optimal policy is also optimal for the ARD-model.*

PROOF. Let f_o^∞ be a Blackwell optimal policy. Take an arbitrary $R \in C$. Then, (7.3.1) and lemma 7.3.1 imply that

$$\chi_i(R) \leq \liminf_{\lambda\downarrow 0} \lambda v_i^\lambda(R) \leq \liminf_{\lambda\downarrow 0} \lambda v_i^\lambda(f_o^\infty) = \chi_i(f_o^\infty), \quad i \in E.$$

Consequently, $\chi(f_o^\infty) = \chi$, i.e. f_o^∞ is an optimal policy for the ARD-model. □

From theorem 7.3.1 it follows that for the determination of an optimal policy in the ARD-model, we may restrict ourselves to the pure and stationary policies. Consider a pure and stationary policy f^∞. Then, (7.3.2) and lemma 2.4.2 imply that $\chi(f^\infty)$ depends on the rewards and the transition times only through $\hat{r}(f)$ and $\tau(f)$ respectively. Hence, for the computation of $\chi(f^\infty)$ it is sufficient to know the values τ_{ia}, $a \in A(i)$, $i \in E$, where

$$(7.3.4) \qquad \tau_{ia} := \sum_j p_{iaj} \int_0^\infty t dF_{iaj}(t),$$

instead of explicit knowledge about the probability distributions $F_{iaj}(t)$. We may assume that

$$(7.3.5) \qquad F_{iaj}(t) = \begin{cases} 0 & t < \tau_{ia} \\ 1 & t = \tau_{ia} \end{cases}, \quad a \in A(i),\ i,j \in E.$$

Therefore, we shall denote an ARD-model by $(E,A,p,\hat{r},\tau)$, where $\hat{r}_{ia} := r_{ia} + s_{ia} \cdot \tau_{ia}$, $a \in A(i)$, $i \in E$.

<u>DEFINITION 7.3.1</u>. A vector $\tilde{\chi} \in \mathbb{R}^N$ is *ARD-superharmonic* if there exists a vector $\tilde{w}$ such that

$$\tilde{\chi}_i \geq \sum_j p_{iaj} \tilde{\chi}_j \qquad , \quad a \in A(i),\ i \in E,$$

$$\tau_{ia} \tilde{\chi}_i + \tilde{w}_i \geq \hat{r}_{ia} + \sum_j p_{iaj} \tilde{w}_j, \quad a \in A(i),\ i \in E.$$

<u>THEOREM 7.3.2</u>. *The ARD-value-vector* χ *is the smallest ARD-superharmonic vector.*

<u>PROOF</u>. (cf. theorem 4.2.1). Let f_o^∞ be any pure and stationary Blackwell optimal policy. Since there exists a $\lambda_o > 0$ such that

$$v^\lambda(f_o^\infty) = v^\lambda \quad \text{for every } \lambda \in (0,\lambda_o],$$

theorem 7.2.1 implies that

$$v_i^\lambda(f_o^\infty) \geq r_{ia}^* + \sum_j p_{iaj} \int_0^\infty e^{-\lambda t} dF_{iaj}(t) \cdot v_j^\lambda(f_o^\infty),$$

$$a \in A(i),\ i \in E,\ \lambda \in (0,\lambda_o],$$

where

$$r^*_{ia} := r_{ia} + s_{ia} \sum_j p_{iaj} \int_0^\infty \int_0^t e^{-\lambda s} ds dF_{iaj}(t).$$

Then, it follows from (7.3.5) that

$$v^\lambda_i(f^\infty_o) \geq r_{ia}+s_{ia} \sum_j p_{iaj} \lambda^{-1}(1-e^{-\lambda\tau_{ia}})+ \sum_j p_{iaj} e^{-\lambda\tau_{ia}} \cdot v^\lambda_j(f^\infty_o)$$

for all $a \in A(i)$, $i \in E$, $\lambda \in (0,\lambda_o]$. Using (7.3.1) and the expansion $e^{-\lambda\tau_{ia}} = 1-\lambda\tau_{ia} + o(\lambda)$, we obtain

$$\lambda^{-1}\chi_i(f^\infty_o) + w_i(f^\infty_o) + o(1) \geq$$

$$\lambda^{-1}\sum_j p_{iaj}\chi_j(f^\infty_o) + r_{ia} + s_{ia}\tau_{ia} + \sum_j p_{iaj} \cdot w_j(f^\infty_o) - \tau_{ia}\sum_j p_{iaj}\chi_j(f^\infty_o)$$

for all $a \in A(i)$, $i \in E$ and $\lambda \in (0,\lambda_o]$. Since $\chi(f^\infty_o) = \chi$, it follows that

$$\text{(7.3.6)} \qquad \chi_i \geq \sum_j p_{iaj}\chi_j \qquad\qquad a \in A(i),\ i \in E,$$

and

$$w_i(f^\infty_o) \geq \hat{r}_{ia} + \sum_j p_{iaj} w_j(f^\infty_o) - \tau_{ia}\sum_j p_{iaj}\chi_j =$$

$$\hat{r}_{ia} + \sum_j p_{iaj} w_j(f^\infty_o) - \tau_{ia}\chi_i, \qquad\qquad a \in \bar{A}(i),\ i \in E,$$

where

$$\bar{A}(i) := \{a \in A(i) \mid \chi_i = \sum_j p_{iaj}\chi_j\}, \qquad\qquad i \in E.$$

Then, similarly as in theorem 4.2.2 we can prove that

$$\text{(7.3.7)} \qquad \tilde{w}_i \geq \hat{r}_{ia} + \sum_j p_{iaj}\tilde{w}_j - \tau_{ia}\chi_i, \qquad\qquad a \in A(i),\ i \in E,$$

where

$$\tilde{w} := w(f^\infty_o) - M \cdot \chi$$

and

$$M := \min\left\{\frac{\tau_{ia}\chi_i + w_i(f^\infty_o) - \hat{r}_{ia} - \sum_j p_{iaj} w_j(f^\infty_o)}{\chi_i - \sum_j p_{iaj}\chi_j} \,\middle|\, a \in A^*(i),\ i \in E\right\},$$

with

$$A^*(i) := \{a \in A(i) \mid \tau_{ia}\chi_i + w_i(f^\infty_o) < \hat{r}_{ia} + \sum_j p_{iaj} w_j(f^\infty_o)\}, \quad i \in E.$$

(if $A^*(i) = \emptyset$ for all $i \in E$, then we define $M := 0$). Consequently, (7.3.6) and (7.3.7) imply that the ARD-value-vector χ is ARD-superharmonic.

Suppose that $\tilde{\chi}$ is also ARD-superharmonic with corresponding vector $\tilde{w}$. Then,

$$(I-P(f_o))\tilde{\chi} \geq 0 \quad \text{and} \quad T(f_o)\tilde{\chi} + \tilde{w} \geq \hat{r}(f_o) + P(f_o)\tilde{w}.$$

Consequently,

$$(I-P(f_o))\tilde{\chi} \geq 0 \quad \text{and} \quad P^*(f_o)T(f_o)\tilde{\chi} \geq P^*(f_o)\hat{r}(f_o).$$

Then, (7.3.2) and lemma 2.4.3 imply that $\tilde{\chi} \geq \chi$, completing the proof that χ is the smallest ARD-superharmonic vector. □

Since χ is the smallest ARD-superharmonic vector, we consider the following linear programming problem:

$$\text{(7.3.8)} \qquad \min\left\{ \sum_j \beta_j \tilde{\chi}_j \;\middle|\; \begin{array}{lr} \sum_j (\delta_{ij} - p_{iaj})\tilde{\chi}_j & \geq 0 \;,\; a \in A(i), i \in E \\ \tau_{ia}\tilde{\chi}_i + \sum_j (\delta_{ij} - p_{iaj})\tilde{w}_j & \geq \hat{r}_{ia},\; a \in A(i), i \in E \end{array} \right\},$$

where $\beta_j > 0$, $j \in E$, are given numbers with $\Sigma_j \beta_j = 1$. The dual linear programming problem is:

$$\text{(7.3.9)} \qquad \max\left\{ \sum_i \sum_a \hat{r}_{ia} x_{ia} \;\middle|\; \begin{array}{llr} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} & = 0 \;, & j \in E \\ \sum_a \tau_{ja} x_{ja} + \sum_i \sum_a (\delta_{ij} - p_{iaj}) y_{ia} & = \beta_j, & j \in E \\ & x_{ia}, y_{ia} \geq 0 \;, & a \in A(i), i \in E \end{array} \right\}.$$

<u>THEOREM 7.3.3</u>. *If* (x^*, y^*) *is an optimal extreme solution of the linear program* (7.3.9), *then the policy* f_*^∞, *where*

$$f_*(i) := a_i \text{ such that} \begin{cases} x^*_{ia_i} > 0, & i \in E_{x^*} \\ y^*_{ia_i} > 0, & i \in E \backslash E_{x^*}, \end{cases}$$

is an optimal policy for the ARD-model.

<u>PROOF</u>. (cf. theorem 4.2.4). Let (χ^*, w^*) be an optimal solution of the linear

programming problem (7.3.8). Then, $\chi^* = \chi$, and analogously to the proof of theorem 4.2.4 we can show that

1. f_*^∞ is well-defined.
2. $\Sigma_j(\delta_{ij}-p_{if_*(i)j})\chi_j = 0, \qquad i \in E,$
3. $\tau_{if_*(i)}\chi_i + \Sigma_j(\delta_{ij}-p_{if_*(i)j})w_j^* = \hat{r}_{if_*(i)}, \quad i \in E_{x^*}.$
4. The states of $E\backslash E_{x^*}$ are transient in the Markov chain induced by $P(f_*)$.

From the above properties it follows that

$$\begin{cases} (I-P(f_*))\chi = 0 \\ P^*(f_*)T(f_*)\chi = P^*(f_*)\hat{r}(f_*). \end{cases}$$

Hence, (7.3.2) implies that $\chi(f_*^\infty) = \chi$, i.e. f_*^∞ is an optimal policy. □

<u>REMARK 7.3.1</u>. The linear programming problems (7.3.4) and (7.3.5) were already proposed by DENARDO & FOX [1968]. However, they only proved that the program (7.3.8) determines the vector χ, but they did not prove the optimality of the policy f_*^∞.

ALGORITHM XXXI for the construction of a pure and stationary optimal policy in an undiscounted semi-Markov model (multichain case).

<u>step 1</u>: Take any choice of the numbers β_j such that $\beta_j > 0$, $j \in E$, and $\Sigma_j\beta_j = 1$.

<u>step 2</u>: Use the simplex method to compute an optimal solution (x^*,y^*) of the linear programming problem

$$max\left\{\sum_i\sum_a \hat{r}_{ia}x_{ia} \;\middle|\; \begin{array}{lr} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & j\in E \\ \sum_a \tau_{ja}x_{ja} + \sum_i\sum_a(\delta_{ij}-p_{iaj})y_{ia} = \beta_j, & j\in E \\ x_{ia}, y_{ia} \geq 0, & a\in A(i), i\in E \end{array}\right\}.$$

<u>step 3</u>: For each $i \in E$, take an arbitrary action a_i from the set $A^*(i)$, where

$$A^*(i) := \begin{cases} \{a \mid x_{ia}^* > 0\} & \text{if } i \in E_{x^*} \\ \{a \mid y_{ia}^* > 0\} & \text{if } i \in E\backslash E_{x^*}. \end{cases}$$

<u>step 4</u>: f_*^∞, where $f_*(i) := a_i$, $i \in E$, is a pure and stationary optimal policy.

Consider the linear programming problem (7.3.8) and substitute

$$(7.3.10)\quad \begin{cases} \bar{x}_i := \tilde{x}_i & , i \in E \\ \bar{w}_i := \tau^{-1}\cdot\tilde{w}_i & , i \in E \\ \bar{r}_{ia} := \tau_{ia}^{-1}\cdot\hat{r}_{ia} & , a \in A(i),\ i \in E \\ \bar{p}_{iaj} := \delta_{ij} - (\delta_{ij}-p_{iaj})\cdot(\tau/\tau_{ia}), & a \in A(i),\ i,j \in E, \end{cases}$$

where τ satisfies

$$0 < \tau \le min_{i,a}\left\{\frac{\tau_{ia}}{1-p_{iai}}\ \middle|\ p_{iai} \ne 1\right\}.$$

Then, $\bar{p}_{iaj} \ge 0$ and $\Sigma_j\bar{p}_{iaj} = 1$ for every $a \in A(i)$, $i,j \in E$. Furthermore, we obtain

$$\sum_j(\delta_{ij}-p_{iaj})\tilde{x}_j \ge 0 \iff \sum_j(\delta_{ij}-\bar{p}_{iaj})\bar{x}_j\cdot(\tau_{ia}/\tau) \ge 0$$

$$\iff \sum_j(\delta_{ij}-\bar{p}_{iaj})\bar{x}_j \ge 0 \quad \text{for all } a \in A(i),\ i \in E,$$

and

$$\tau_{ia}\tilde{x}_i + \sum_j(\delta_{ij}-p_{iaj})\tilde{w}_j \ge \hat{r}_{ia} \iff \tau_{ia}\bar{x}_i + \sum_j(\delta_{ij}-\bar{p}_{iaj})\cdot(\tau_{ia}/\tau)\cdot\tau\bar{w}_i \ge \hat{r}_{ia}$$

$$\iff \bar{x}_i + \sum_j(\delta_{ij}-\bar{p}_{iaj})\bar{w}_i \ge \bar{r}_{ia}$$

$$\text{for all } a \in A(i),\ i \in E.$$

Hence, the linear program (7.3.8) can also be written as

$$(7.3.11)\quad min\left\{\sum_j\beta_j\bar{x}_j\ \middle|\ \begin{array}{ll} \sum_j(\delta_{ij}-\bar{p}_{iaj})\bar{x}_j \ge 0 & , a \in A(i),\ i \in E \\ \bar{x}_i + \sum_j(\delta_{ij}-\bar{p}_{iaj})\bar{w}_j \ge \bar{r}_{ia}, & a \in A(i),\ i \in E \end{array}\right\}$$

and $(\tilde{x},\tilde{w})$ is a feasible solution of program (7.3.8) if and only if $(\tilde{x},\tau^{-1}\cdot\tilde{w})$ is a feasible solution of program (7.3.11). The transformations (7.3.10) were proposed by SCHWEITZER [1971].

REMARK 7.3.2. Notice that the linear programming problem (7.3.11) is similar to program (4.2.10), with $\bar{p}_{iaj}$ and $\bar{r}_{ia}$ instead of p_{iaj} and r_{ia}. Hence, algorithm XIV for the AMD-model $(E,A,\bar{p},\bar{r})$ is identical to algorithm XXXI for the ARD-model $(E,A,p,\hat{r},\tau)$. Furthermore, it can easily be verified that $\chi(\pi^\infty) = \bar{\phi}(\pi^\infty)$ for every stationary policy π^∞, where $\bar{\phi}(\pi^\infty)$ is the expected average reward in the AMD-model $(E,A,\bar{p},\bar{r})$. Hence, the ARD-model and the corresponding AMD-model may be viewed as *equivalent*.

REMARK 7.3.3. Linear programming can also be used for the *two-person zero-sum semi-Markov game* in which one player controls the transition probabilities and the transition times. If we define

$$\hat{r}_{iab} := r_{iab} + s_{iab} \cdot \tau_{ia}, \qquad a \in A(i),\ b \in B(i),\ i \in E,$$

then similarly as in section 6.3 it can be shown that there exist stationary optimal policies for both players. Moreover, it can be proved that algorithm XXVIII applied on the transformed AMG-model $(E,A,B,\bar{p},\bar{r})$, where

$$\bar{p}_{iaj} := \delta_{ij} - (\delta_{ij} - p_{iaj}) \cdot (\tau/\tau_{ia}), \qquad a \in A(i),\ i,j \in E,$$

$$\bar{r}_{iab} := \tau_{ia}^{-1} \cdot \hat{r}_{iab} \qquad , \qquad a \in A(i),\ b \in B(i),\ i \in E,$$

yields stationary optimal policies for the two players.

We close this section with the presentation of algorithms for the *weak unichain case*, the *unichain case* and the *completely ergodic case*. We say that an ARD-model is weakly unichained, unichained or completely ergodic if the equivalent AMD-model satisfies assumption 4.5.1, assumption 4.6.2 or assumption 4.6.1 respectively. Then, the results of the sections 4.5 and 4.6 imply that we can use the following algorithms.

ALGORITHM XXXII for the construction of a pure and stationary optimal policy in an undiscounted semi-Markov model (weak unichain case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$max\left\{ \sum_i \sum_a \hat{r}_{ia} x_{ia} \;\middle|\; \begin{array}{lll} \sum_i \sum_a (\delta_{ij} - p_{iaj}) x_{ia} = 0, & & j \in E \\ \sum_i \sum_a \tau_{ia} x_{ia} \quad = 1 & & \\ \qquad\qquad x_{ia} \geq 0, & a \in A(i),\ i \in E \end{array} \right\}.$$

step 2: Take $f_*(i)$ such that $x^*_{if_*(i)} > 0$, $i \in E_{x^*}$.

step 3: Let $E_o := E_{x^*}$.

step 4: If $E_o = E$, then f_*^∞ is an optimal policy (STOP). Otherwise, go to step 5a.

step 5a: Choose a triple (i,a_i,j) that satisfies $i \in E \backslash E_o$, $a_i \in A(i)$, $j \in E_o$ and $p_{ia_ij} > 0$.

step 5b: Define $f_*(i) := a_i$ and $E_o := E_o \cup \{i\}$; go to step 4.

ALGORITHM XXXIII for the construction of a pure and stationary optimal policy in an undiscounted semi-Markov model (unichain case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$max\left\{\sum_i\sum_a \hat{r}_{ia}x_{ia} \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & & j \in E \\ \sum_i\sum_a \tau_{ia}x_{ia} = 1 & & \\ x_{ia} \geq 0, & a \in A(i), & i \in E \end{array}\right\}.$$

step 2: Take f_*^∞ such that

$$f_*(i) := \begin{cases} a_i \text{ where } x^*_{ia_i} > 0, & i \in E_{x^*} \\ \text{arbitrarily}, & i \in E \backslash E_{x^*}. \end{cases}$$

ALGORITHM XXXIV for the construction of a pure and stationary optimal policy in an undiscounted semi-Markov model (completely ergodic case).

step 1: Use the simplex method to compute an optimal solution x^* of the linear programming problem

$$max\left\{\sum_i\sum_a \hat{r}_{ia}x_{ia} \;\middle|\; \begin{array}{lll} \sum_i\sum_a(\delta_{ij}-p_{iaj})x_{ia} = 0, & & j \in E \\ \sum_i\sum_a \tau_{ia}x_{ia} = 1 & & \\ x_{ia} \geq 0, & a \in A(i), & i \in E \end{array}\right\}.$$

step 2: Take f_*^∞ such that $x^*_{if_*(i)} > 0$, $i \in E$.

REFERENCES

BATHER, J. [1973]. *Optimal decision procedures for finite Markov chains. Part II: Communicating systems*. Advances in Applied Probability, 5, 521-540.

BELLMAN, R. [1957]. *Dynamic programming*. Princeton University Press, Princeton.

BEWLEY, T. & E. KOHLBERG [1978]. *On stochastic games with stationary optimal strategies*. Mathematics of Operations Research, 3, 104-125.

BLACKWELL, D. [1962]. *Discrete dynamic programming*. Annals of Mathematical Statistics, 33, 719-726.

BLACKWELL, D. & T.S. FERGUSON [1968]. *The big match*. Annals of Mathematical Statistics, 39, 159-163.

BLAND, R.G. [1977]. *New finite pivoting rules for the simplex method*. Mathematics of Operations Research, 2, 103-107.

COLLATZ, L. & W. WETTERLING [1966]. *Optimierungsaufgaben*. Springer, Berlin.

DANTZIG, G.B. [1963]. *Linear programming and extensions*. Princeton University Press, Princeton.

DE CANI, J.S. [1964]. *A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity*. Management Science, 10, 716-733.

DE GHELLINCK, G.T. [1960]. *Les Problèmes de décisions sequentielles*. Cahiers du Centre d'Etudes de Recherche Opérationelle, 2, 161-179.

DE GHELLINCK, G.T. & G.D. EPPEN [1967]. *Linear programming solutions for separable Markovian decision problems*. Management Science, 13, 371-394.

DENARDO, E.V. [1967]. *Contraction mappings in the theory underlying dynamic programming*. SIAM Review, 9, 165-177.

DENARDO, E.V. [1970a]. *Computing a bias-optimal policy in a discrete-time Markov decision problem*. Operations Research, 18, 279-289.

DENARDO, E.V. [1970b]. *On linear programming in a Markov decision problem*. Management Science, 16, 281-288.

DENARDO, E.V. [1971]. *Markov renewal programs with small interest rates.* Annals of Mathematical Statistics, 42, 477-496.

DENARDO, E.V. [1973]. *A Markov decision problem,* pp. 33-68 in T.C. Hu & S.M. Robinson (eds.), *Mathematical programming,* Academic Press, New York.

DENARDO, E.V. & B.L. FOX [1968]. *Multichain Markov renewal programs.* SIAM Journal on Applied Mathematics, 16, 468-487.

DENARDO, E.V. & B.L. MILLER [1968]. *An optimality condition for discrete dynamic programming with no discounting.* Annals of Mathematical Statistics, 39, 1220-1227.

DENARDO, E.V. & U.G. ROTHBLUM [1979]. *Optimal stopping, exponential utility, and linear programming.* Mathematical Programming, 16, 228-244.

D'EPENOUX, F. [1960]. *Sur un problème de production et de stockage dans l'aléatoire.* Revue Française de Recherche Operationelle, 14, 3-16.

DERMAN, C. [1970]. *Finite state Markovian decision processes.* Academic Press, New York.

DERMAN, C. & R. STRAUCH [1966]. *A note on memoryless rules for controlling sequential control problems.* Annals of Mathematical Statistics, 37, 276-278.

DERMAN, C. & A.F. VEINOTT Jr. [1972]. *Constrained Markov decision chains.* Management Science, 19, 389-390.

DIRICKX, Y.M.J. & M.R. RAO [1979]. *Linear programming methods for computing gain-optimal policies in Markov decision models.* Cahiers du Centre d'Etudes de Recherche Opérationelle, 21, 133-142.

DOOB, J.L. [1953]. *Stochastic processes.* Wiley, New York.

FELLER, W. [1968]. *An introduction to probability theory and its applications.* Volume I, third edition, Wiley, New York.

FILAR, J.A. & T.E.S. RAGHAVAN [1979]. *An algorithm for solving an undiscounted stochastic game in which one player controls transitions.* Report.

FOX, B.L. [1966]. *Markov renewal programming by linear fractional programming.* SIAM Journal on Applied Mathematics, 14, 1418-1432.

FOX, B.L. & D.M. LANDI [1968]. *An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix.* Communications of the ACM, 11, 619-621.

GILLETTE, D. [1957]. *Stochastic games with zero stop probabilities*, pp. 179-187 in M. Dresher, A.W. Tucker and P. Wolfe (eds.), *Contributions to the theory of games*, volume 3, Annals of Mathematical Studies no. 39, Princeton University Press, Princeton.

HADLEY, G. [1962]. *Linear programming*. Addison-Wesley, Reading.

HEE, K.M. VAN, A. HORDIJK & J. VAN DER WAL [1977]. *Successive approximations for convergent dynamic programming*, pp. 183-211 in H.C. Tijms and J. Wessels (eds.) *Markov decision theory*, Mathematical Centre Tract no. 93, Mathematical Centre, Amsterdam.

HEILMANN, W.-R. [1977]. *Lineare Programmierung stochastischer dynamischer Entscheidungsmodelle*. Ph.D. dissertation, Hamburg.

HINDERER, K. [1970]. *Foundations of non-stationary dynamic programming with discrete time parameter*. Springer, Berlin.

HORDIJK, A. [1971]. *A sufficient condition for the existence of an optimal policy with respect to the average cost criterion in Markovian decision processes*. Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Prague, 263-274.

HORDIJK, A. [1974]. *Dynamic programming and Markov potential theory*. Mathematical Centre Tract no. 51, Mathematical Centre, Amsterdam.

HORDIJK, A. [1976]. *Stochastic dynamic programming*. Course notes, University of Leiden (in Dutch).

HORDIJK, A., R. DEKKER & L.C.M. KALLENBERG [1981]. *A simplex-like algorithm to compute a Blackwell-optimal policy*. Report no. 81-37, Institute of Applied Mathematics and Computer Science, University of Leiden.

HORDIJK, A. & L.C.M. KALLENBERG [1978a]. *Linear programming and Markov decision chains I*. Report no. 78-1, Institute of Applied Mathematics and Computer Science, University of Leiden.

HORDIJK, A. & L.C.M. KALLENBERG [1978b]. *Linear programming and Markov decision chains II*. Report no. 78-5, Institute of Applied Mathematics and Computer Science, University of Leiden.

HORDIJK, A. & L.C.M. KALLENBERG [1979a]. *Linear programming and Markov decision chains*. Management Science, 25, 352-362.

HORDIJK, A. & L.C.M. KALLENBERG [1979b]. *On solving Markov decision problems by linear programming*. Proceedings of the Manchester Conference on Markov decision processes (to appear).

HORDIJK, A. & L.C.M. KALLENBERG [1981a]. *Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints*. Report no. 81-30, Institute of Applied Mathematics and Computer Science, University of Leiden.

HORDIJK, A. & L.C.M. KALLENBERG [1981b]. *Constrained undiscounted stochastic dynamic programming*. Report no. 81-53, Institute of Applied Mathematics and Computer Science, University of Leiden.

HORDIJK, A. & L.C.M. KALLENBERG [1981c]. *Algorithms for the computation of an optimal transient policy in unconstrained and constrained dynamic programming*. Methods of Operations Research, 41, 37-40.

HORDIJK, A. & L.C.M. KALLENBERG [1981d]. *Linear programming methods for solving finite Markovian decision problems*, pp. 468-482 in G. Fandel et. al. (eds.) *Operations Research Proceedings*, Springer, Berlin.

HORDIJK, A. & L.C.M. KALLENBERG [1981e]. *Linear programming and Markov games I*, pp. 291-305 in O. Moeschlin and D. Pallaschke (eds.) *Game theory and mathematical economics*, North-Holland, Amsterdam.

HORDIJK, A. & L.C.M. KALLENBERG [1981f]. *Linear programming and Markov games II*, pp. 307-320 in O. Moeschlin and D. Pallaschke (eds.) *Game theory and mathematical economics*, North-Holland, Amsterdam.

HORDIJK, A. & K. SLADKY [1977]. *Sensitive optimality criteria in countable state dynamic programming*. Mathematics of Operation Research, 2, 1-14.

HORDIJK, A. & H.C. TIJMS [1970]. *Colloquium Markov-programming*. Mathematical Centre Report BC 1/70, Mathematical Centre, Amsterdam (in Dutch).

HOWARD, R.A. [1960]. *Dynamic programming and Markov processes*. M.I.T. Press, Cambridge, Massachusetts.

HOWARD, R.A. [1963]. *Semi-Markovian decision processes*. Proceedings International Statistical Institute, Ottawa.

JEWELL, W.S. [1963a]. *Markov renewal programming. I: Formulation, finite return models*. Operations Research, 11, 938-948.

JEWELL, W.S. [1963b]. *Markov renewal programming. II: Infinite return models, example*. Operations Research, 11, 949-971.

KALLENBERG, L.C.M. [1981a]. *Unconstrained and constrained dynamic programming over a finite horizon*. Report no. 81-46, Institute of Applied Mathematics and Computer Science, University of Leiden.

KALLENBERG, L.C.M. [1981b]. *Linear programming to compute a bias-optimal policy*, pp. 433-440 in B. Fleischman et.al. (eds.) *Operations Research Proceedings* 1981, Springer, Berlin.

KALLENBERG, L.C.M. [1983]. *Finite algorithms for positive and negative dynamic programming*. Report, Institute of Applied Mathematics and Computer Science, University of Leiden.

KARLIN, S. [1959]. *Mathematical methods and theory in games, programming and economics*. Volume I, Addison-Wesley, Reading, Massachusetts.

KEMENY, J.G. & J.L. SNELL [1960]. *Finite Markov chains*. Van Nostrand, Princeton, New Jersey.

MACQUEEN, J. [1967]. *A test for suboptimal actions in Markovian decision problems*. Operations Research, 15, 559-561.

MANNE, A.S. [1960]. *Linear programming and sequential decisions*. Management Science, 6, 259-267.

MERTENS, J.F. & A. NEYMAN [1980]. *Stochastic games*, Report, Université Catholique de Louvain.

MILLER, B.L. & A.F. VEINOTT Jr. [1969]. *Discrete dynamic programming with a small interest rate*. Annals of Mathematical Statistics, 40, 366-370.

MINE, H. & S. OSAKI [1970]. *Markovian decision processes*. Elsevier, New York.

MONASH, C.A. [1979]. *Stochastic games: the minimax theorem*. Ph.D. dissertation, Harvard University, Cambridge, Massachusetts.

NUNEN, J.A.E.E. Van [1976]. *Contracting Markov decision processes*. Mathematical Centre Tract no. 71, Mathematical Centre, Amsterdam.

NUNEN, J.A.E.E. Van & J. WESSELS [1977]. *Markov decision processes with unbounded rewards*, pp.1-24 in H.C. Tijms and J. Wessels (eds.) *Markov decision theory*, Mathematical Centre Tract no. 93, Mathematical Centre, Amsterdam.

PARTHASARATHY, T. & T.E.S. RAGHAVAN [1977]. *Finite algorithms for stochastic games*. Paper presented at the International Conference on Dynamic Programming, Vancouver.

PARTHASARATHY, T. & T.E.S. RAGHAVAN [1978]. *An order field property for stochastic games when one player controls the transition probabilities*. Paper presented at the Game Theory Conference, Cornell, Ithaca.

ROCKAFELLER, R.T. [1970]. *Convex analysis*. Princeton University Press, Princeton, New Jersey.

ROSS, S.M. [1970]. *Applied probability models with optimization applications*. Holden-Day, San Francisco, California.

ROTHBLUM, U.G. [1979]. *Solving stopping stochastic games by maximizing a linear function subject to quadratic constraints,* pp. 103-105 in O. Moeschlin and D. Pallaschke (eds.) *Game theory and related topics,* North Holland, Amsterdam.

SCHWEITZER, P.J. [1965]. *Perturbation theory and Markovian decision processes*. Ph.D. dissertation, M.I.T. Operations Research Center Report 15.

SCHWEITZER, P.J. [1971]. *Iterative solution of the functional equations of undiscounted Markov renewal programming*. Journal of Mathematical Analysis and Applications, 34, 495-501.

SCHWEITZER, P.J. & A. FEDERGRUEN [1978]. *The functional equations of undiscounted Markov renewal programming*. Mathematics of Operations Research, 3, 308-321.

SHAPLEY, L.S. [1953]. *Stochastic games*. Proceedings National Academy of Sciences U.S.A., 39, 1095-1100.

STRAUCH, R. & A.F. VEINOTT, Jr. [1966]. *A property of sequential control processes*. Rand McNally, Chicago, Illinois.

VEINOTT, A.F. Jr. [1966]. *On finding optimal policies in discrete dynamic programming with no discounting*. Annals of Mathematical Stattistics, 37, 1284-1294.

VEINOTT, A.F. Jr. [1969]. *Discrete dynamic programming with sensitive discount optimality criteria*. Annals of Mathematical Statistics, 40, 1635-1660.

VEINOTT, A.F. Jr. [1973]. *State-action frequencies in Markov decision chains*. Unpublished manuscript.

VEINOTT, A.F. Jr. [1974]. *Markov decision chains*. pp. 124-159 in G.B. Dantzig, & B.C. Eaves (eds.) Studies in Mathematics, Volume 10: *Studies in optimization*. The Mathematical Association of America.

VRIEZE, O. [1980]. *Linear programming and undiscounted stochastic games in which one player controls the transitions*. Mathematical Centre Report BW 122/80, Mathematical Centre, Amsterdam.

WESSELS, J. & J.A.E.E. Van NUNEN [1975]. *Discounted semi-Markov decision processes: linear programming and policy iteration*. Statistica Neerlandica, 29, 1-7.

WIDDER, D.V. [1946]. *The Laplace transform*. Princeton University Press, Princeton, New Jersey.

ZOUTENDIJK, G. [1960]. *Methods of feasible directions, a study in linear and nonlinear programming*. Elsevier, New York.

ZOUTENDIJK, G. [1976]. *Mathematical programming methods*. North-Holland, Amsterdam.

LIST OF ALGORITHMS

AUTHOR INDEX

SUBJECT INDEX

SYMBOL INDEX

(Symbols with only local significance are not included)

GREEK

LATIN

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.

CWI TRACTS

1 D.H.J. Epema. *Surfaces with canonical hyperplane sections.* 1984.

2 J.J. Dijkstra. *Fake topological Hilbert spaces and characterizations of dimension in terms of negligibility.* 1984.

3 A.J. van der Schaft. *System theoretic descriptions of physical systems.* 1984.

4 J. Koene. *Minimal cost flow in processing networks, a primal approach.* 1984.

5 B. Hoogenboom. *Intertwining functions on compact Lie groups.* 1984.

6 A.P.W. Böhm. *Dataflow computation.* 1984.

7 A. Blokhuis. *Few-distance sets.* 1984.

8 M.H. van Hoorn. *Algorithms and approximations for queueing systems.* 1984.

9 C.P.J. Koymans. *Models of the lambda calculus.* 1984.