ZOR – Mathematical Methods of Operations Research (1994) 40:1–42

Survey of Linear Programming for Standard and Nonstandard Markovian Control Problems. Part I: Theory

L. C. M. KALLENBERG

Leiden University, Department of Mathematics and Computer Science, P.O. Box 9512, 2300 RA Leiden. The Netherlands

Abstract: This paper gives an overview of linear programming methods for solving standard and nonstandard Markovian control problems. Standard problems are problems with the usual criteria such as expected total (discounted) rewards and average expected rewards; we also discuss a particular class of stochastic games. In nonstandard problems there are additional considerations as side constraints, multiple criteria or mean-variance tradeoffs. In a second companion paper efficient linear programing algorithms are discussed for some applications.

1 Introduction

The pioneering work in solving Markovian control problems by linear programming was made in 1960, at a very early stage in the development of Markov decision theory. The initial papers were written by D'Epenoux (1960) (with the translated version D'Epenoux (1963)), De Ghellinck (1960) and Manne (1960). D'Epenoux (1960) considered the discounted rewards; linear programming for the average reward criterion (in the unichain case without transient states) is due to De Ghellinck (1960) and independently Manne (1960). The first analysis of linear programming in the general multichain case has been presented by Denardo/Fox (1968) and Denardo (1970a). Hordijk/Kallenberg (1979) have shown that also in the multichain case only one linear program suffices. Many results about linear programming for standard Markovian control problems can be found in the thesis Kallenberg (1983) written under the supervision of Hordijk.

We will include in this survey the single-controller stochastic game. The linear programming results for this model are in the undiscounted case due to Hordijk/Kallenberg (1981a, 1981b) and Vrieze (1981). A detailed survey of algorithms for stochastic games is presented by Raghavan/Filar (1991).

For solving standard Markovian control problems there are three main methods, namely *successive approximation*, *policy improvement* and *linear programming*. For surveys on successive approximation we refer to Porteus (1980)

0340–9422/94/40:1/1–42 $2.50

and Federgruen/Schweitzer (1980); an overview of policy improvement can be found in Puterman (1988); for previous surveys on linear programming and Markov decision processes, we refer to Heilmann (1978) and Hordijk/Kallenberg (1980).

The linear programming approach is particularly useful for nonstandard problems. In this area we mention the following contributions. Derman/Klein (1965) have shown that finite horizon problems with side constraints can be handled by linear programming. The treatment of infinite horizon problems with side constraints or with multiple objectives is based on the state-action frequencies as developed by Derman (1970, Chapter 7) and Hordijk/Kallenberg (1984a, 1984b). For the mean-variance tradeoffs problem we mention the papers by Sobel (1985), White (1988) and Huang/Kallenberg (1994). The last paper unifies and extends formulations and results discussed by many other authors, and it is shown that the nonlinear mean-variance problem can be solved by parametric linear programming. This last observation is based on work by Kawai (1987).

In this overview we restrict ourselves to problems with a *finite* state space and *finite* action sets. For linear programming in more general Markovian decision models we refer to the thesis of Heilmann (1977).

2 Preliminaries

A *Markovian decision process* is defined as follows. At the beginning of each period $t = 1, 2, \ldots$ a process is observed by a decision maker to be in one of the states of a *finite state space* $E = \{1, 2, \ldots, N\}$. When the process is observed in state i the decision maker chooses an action a from a *finite action set* $A(i)$, and the following happens: an *immediate reward* r_{ia} is earned and the process moves to state j with probability p_{iaj}, where $p_{iaj} \geq 0$ and $\sum_j p_{iaj} = 1$ for every $(i, a) \in E \times A := \{(i, a) | a \in A(i), i \in E\}$.

A *policy* is a sequence of decision rules, one decision rule for each period. These rules may randomize over the actions and may depend on the whole history of the process. A policy is said to be a *Markov policy* if the decision rule π^t in period t only depends on the state at time t. It is well-known that Markov policies are sufficient for the usual utility functions (cf. Derman/Strauch (1966)). A *stationary policy* π is a policy which uses the same (randomized) rule in each period: π_{ia} is the probability that action $a \in A(i)$ is chosen in state $i \in E$. A policy is said to be *deterministic* if there is no randomization, i.e. it selects an action with certainty. Hence, a deterministic and stationary policy selects the same action in every period. The chosen action only depends on the state: such a policy may be considered as a function f that assigns to each state i an action $f(i) \in A(i)$.

Let the random variables X_t and Y_t denote the state and action, respectively, at time t ($t = 1, 2, \ldots$). $\mathbb{P}_R[X_t = j, Y_t = a | X_1 = i]$ is the notation for the probability that at time t the state is j and the action is a, given that policy R is used and that state i is the starting state.

A Standard Markovian Control Problems

In this part we discuss linear programming methods for several optimality criteria: (expected) discounted rewards, total rewards, average rewards, bias optimality and Blackwell optimality. We will present the results in such a way that the similarity between the different models becomes apparent. We always start with the notions of the *value-vector* and a related concept, namely a *superharmonic vector*. It can be shown that the value-vector is the smallest superharmonic vector. From this property a *primal linear program* is obtained which determines the value-vector. An optimal policy can be obtained from the corresponding solution of the *dual program*. Futhermore, the variables of the dual program can be interpreted as the *state-action frequencies*.

3 Discounted Reward Criterion

The *total expected discounted reward* for policy R, initial state i and discount factor $\alpha \in [0, 1)$ is denoted by $v_i^\alpha(R)$ and defined by

$$v_i^\alpha(R) := \sum_{t=1}^{\infty} \mathbb{E}_R[\alpha^{t-1} r_{X_t Y_t} | X_1 = i] \ . \tag{1}$$

Hence, $v_i^\alpha(R)$ can be writen as

$$v_i^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j,a} \mathbb{P}_R[X_t = j, Y_t = a | X_1 = i] \cdot r_{ja} \ . \tag{2}$$

$v_i^\alpha := \sup_R v_i^\alpha(R)$, $i \in E$, is called the *discounted-value-vector*. R^* is said to be a *discounted-optimal* policy if $v_i^\alpha(R^*) = v_i^\alpha$, $i \in E$. The existence of a discounted-optimal policy is not trivial: the supremum must be attained for all starting states simultaneously. Indeed, there exists a discounted-optimal policy, even a deterministic and stationary one (cf. Derman (1970)).

A vector $v \in \mathbb{R}^N$ is said to be *discounted-superharmonic* if

$$v_i \geq r_{ia} + \alpha \cdot \sum_j p_{iaj} v_j \quad \text{for every } (i, a) \in E \times A \ . \tag{3}$$

The following results are in essence due to D'Epenoux (1960). The presentation in terms of superharmonicity can be found in Kallenberg (1983). The next theorem is based on the well-known property (cf. Denardo (1967)) that the value-vector is the unique fixed-point of the contraction mapping $U: \mathbb{R}^N \to \mathbb{R}^N$, where $(Ux)_i := \max_{a \in A(i)} \{r_{ia} + \alpha \cdot \sum_j p_{iaj} x_j\}$.

Theorem 1: The value-vector v^α is the (componentwise) smallest discounted-superharmonic vector.

Corollary: v^α is the unique optimal solution of the linear primal program

$$\min \left\{ \sum_j \beta_j v_j \middle| \sum_j (\delta_{ij} - \alpha \cdot p_{iaj}) v_j \geq r_{ia}, (i, a) \in E \times A \right\} \tag{4}$$

where $\beta_j > 0, j \in E$, is arbitrarily chosen.
The dual linear program of (4) is

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \middle| \begin{array}{l} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj}) x_{ia} = \beta_j \ , \quad j \in E \\ \\ x_{ia} \geq 0 \ , \quad (i, a) \in E \times A \end{array} \right\} . \tag{5}$$

Theorem 2: Let x^* be an extreme optimal solution of (5). Then, for every $i \in E$ $x^*_{ia} > 0$ for exactly one action, say a_i, and the deterministic and stationary policy f^* with $f^*(i) = a_i$, $i \in E$, is an α-discounted-optimal policy.

Theorem 3: The mapping $x_{ia} \to \pi(x)$ with $\pi_{ia}(x) := x_{ia} / \sum_a x_{ia}$ is a one-to-one mapping of the feasible solutions of (5) onto the stationary policies with $\pi \to x_{ia}(\pi)$, where

$$x_{ia}(\pi) := \sum_j \beta_j \cdot \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{P}_\pi[X_t = i, Y_t = a | X_1 = j] \ , \tag{6}$$

as inverse mapping. The set of deterministic and stationary policies corresponds to the set of extreme solutions of (5).

Remarks:

1) Theorem 3 is due to De Ghellinck/Eppen (1967). It follows from the definition of $x_{ia}(\pi)$ in Theorem 3 that it is correct to give the variables of the dual program the interpretation of the expected discounted state-action frequencies, i.e. $x_{ia}(\pi)$ is the total expected discounted number of times that state i is entered and action a is chosen, given policy π and initial distribution β. Here, we assume that the β_j's are normalized, i.e. $\sum_j \beta_j = 1$.
2) The policy improvement method (cf. Howard (1960) and Blackwell (1962)) is equivalent to a special implementation to solve (5), namely a block pivoting simplex method (see Mine/Osaki (1970) and Dantzig (1963, p. 201)). Furthermore, this method can be considered as Newton's method to solve the fixed-point equation $Ux = x$ (cf. Puterman/Brumelle (1979)).
3) During the solution of (5) by the simplex method, it is possible to apply a *suboptimality test*. Consider the simplex tableau corresponding to the deterministic and stationary policy f with variables $x_{ia}(f)$. Then, it can be shown that the corresponding solution v of (4) satisfies $v = v^{\alpha}(f)$. By $d_{ia}(f)$ we denote the slack variable of equation (i, a) in (4), i.e. $d_{ia}(f) := v_i^{\alpha}(f) - \alpha \sum_j p_{ij}(\pi) v_j^{\alpha}(f) - r_{ia}(f)$, where $p_{ij}(\pi) := \sum_a p_{iaj}\pi_{ia}$. Let $d_i(f) := \min_a d_{ia}(f)$, $i \in E$. Then, the values of $d_i(f)$, $i \in E$, are known from the tableau. Let for $x \in \mathbb{R}^N$, $\operatorname{span}[x] := \max_i \{x_i\} - \min_i \{x_i\}$. If

$$d_{ia_i}(f) > d_i(f) + \alpha(1 - \alpha)^{-1} \operatorname{span}[d(f)] \ , \tag{7}$$

then action $a_i \in A(i)$ is suboptimal, as shown in Kallenberg (1983).
4) Stein (1988) has reported on computational results. It turns out that linear programming is efficient (in comparison with other methods) when the discounting factor α is close to 1 or when a high accuracy is desired.
5) The results given in this section also hold for semi-Markov decision processes (cf. Wessels/Van Numen (1975)).

4 Total Reward Criterion

For the total reward criterion we suppose that the transitions are *not necessarily probabilities*. We only assume that for every $(i, a) \in E \times A$ the numbers $p_{iaj}, j \in E$, are nonnegative. There are many applications of this general model, e.g. problems with an option to stop. The results of this section can be found in Kallenberg (1983), chapter 3.
For any Markov policy $R = (\pi^1, \pi^2, \ldots)$, we define

$$\begin{cases} P^t(R) := P(\pi^1)P(\pi^2)\cdots P(\pi^t) \ , \quad \text{where } [P(\pi^t)]_{ij} := \sum_a p_{iaj}\pi^t_{ia} \ , \quad t \in \mathbb{N} \\ P^0(R) := I \ , \quad \text{i.e. the identity matrix} \end{cases} \tag{8}$$

A Markov policy is called *transient* if $\lim_{T\to\infty} \sum_{t=1}^{T} [P^t(R)]_{ij} < \infty$ for all i, j. Then, for the transient Markov policy R and the starting state $i \in E$, the total expected reward $v_i(R)$ is well defined by

$$v_i(R) := \sum_{t=1}^{\infty} \mathbb{E}_R[r_{X_t Y_t} | X_1 = i] \ . \tag{9}$$

Hence, we can write

$$v_i(R) = \sum_{t=1}^{\infty} [P^{t-1}(R) r(\pi^t)]_i \ , \quad \text{where } [r(\pi^t)]_i := \sum_a r_{ia}\pi^t_{ia} \ , \quad t \in \mathbb{N} \ . \tag{10}$$

We will consider the problem to find a transient policy R^*, which is optimal with respect to the total reward criterion, i.e.

$$v_i(R^*) = \sup \{v_i(R) | R \text{ transient}\} \ , \quad i \in E \ . \tag{11}$$

We include the following assumptions:

Assumption 1: There exists a transient policy.

Assumption 2: $w_i := \sup\{v_i(R) | R \text{ transient}\}$ is finite for every $i \in E$.

Remarks:

1) The problem to find an optimal transient policy is related to an optimal stopping problem with an exponential utility function as studied by Denardo/ Rothblum (1979).
2) In Kallenberg (1983) is also shown that assumption 1 can be verified by linear programming, namely it is equivalent to the feasibility of the following linear system

$$\begin{cases} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j \ , & j \in E \\ x_{ia} \geq 0 \ , & (i, a) \in E \times A \end{cases} \tag{12}$$

where $\beta_j, j \in E$, is strictly positive.

3) w, defined in Assumption 2, is the *total-value-vector* for the total reward criterion.

For the present criterion the concept of superharmonicity becomes the following: a vector $v \in \mathbb{R}^N$ is said to be *total-superharmonic* if

$$v_i \geq r_{ia} + \sum_j p_{iaj} v_j \quad \text{for every } (i, a) \in E \times A \ . \tag{13}$$

Theorem 4: The total-value-vector w is the (componentwise) smallest total-superharmonic vector.

Corollary: w is the unique optimal solution of the linear primal program

$$\min \left\{ \sum_j \beta_j v_j \middle| \sum_j (\delta_{ij} - p_{iaj}) v_j \geq r_{ia}, (i, a) \in E \times A \right\} \tag{14}$$

where $\beta_j > 0, j \in E$, is arbitrarily chosen.
The dual linear program is

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \middle| \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = \beta_j \ , & j \in E \\ x_{ia} \geq 0 \ , & (i, a) \in E \times A \end{array} \right\} . \tag{15}$$

Theorem 5: Let x^* be an extreme optimal solution of (15). Then, for every $i \in E$, $x^*_{ia} > 0$ for exactly one action, say a_i, and the determinisitic and stationary policy f^* with $f^*(i) = a_i$, $i \in E$, is an optimal transient policy.

Theorem 6: The mapping $x_{ia} \to \pi(x)$ with $\pi_{ia}(x) := x_{ia} / \sum_a x_{ia}$ is a one-to-one mapping of the feasible solutions of (15) onto the stationary transient policies with $\pi \to x_{ia}(\pi)$, where $x_{ia}(\pi) := \sum_j \beta_j \sum_{t=1}^{\infty} \mathbb{P}_\pi[X_t = i, Y_t = a | X_1 = j]$, as inverse mapping. The set of deterministic and stationary transient policies corresponds to the set of extreme solutions of (15).

Remarks:

1) It follows from the definition of $x_{ia}(\pi)$ that the variables of the dual program have the interpretation of the total expected state-action frequencies, i.e. $x_{ia}(\pi)$ is the total expected number of times that state i is entered and action a is chosen, given policy π and initial distribution β. Here, we assume that the β_j's are normalized: $\sum_j \beta_j = 1$.
2) An interesting model is the case in which every policy is transient. In this case the model is equivalent to the discounted model and is called *contracting dynamic programming* (cf. Veinott (1969) and Van Nunen (1976)). It turns out (cf. Kallenberg (1983)) that this condition is equivalent to the boundedness of the following linear system $\sum_{i,a}(\delta_{ij} - p_{iaj})x_{ia} \leq \beta_j$, $j \in E$, with nonnegative variables x_{ia}, $(i, a) \in E \times A$, and where $\beta_j, j \in E$, is strictly positive. Hence, this means that the following linear program needs a finite solution:

$$\max \left\{ \sum_{i,a} x_{ia} \;\middle|\; \begin{array}{l} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} \leq \beta_j \;, \quad j \in E \\ x_{ia} \geq 0 \;, \quad (i, a) \in E \times A \end{array} \right\} . \tag{16}$$

   A related characterization can be found in Sutherland (1980).
3) During the solution of the dual program (15) it is possible to apply a sub-optimality test (cf. Hordijk/Kallenberg (1984a)).

5 Average Reward Criterion (General Case)

The *average expected reward* for policy R and initial state i is denoted by $g_i(R)$ and defined by

$$g_i(R) := \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_R[r_{X_t Y_t} | X_1 = i] \;. \tag{17}$$

Therefore, the average expected reward can be written as

$$g_i(R) = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j,a} \mathbb{P}_R[X_t = j, Y_t = a | X_1 = i] \cdot r_{ja} \;. \tag{18}$$

$g_i := \sup_R g_i(R)$, $i \in E$, is called the *average-value-vector*. R^* is said to be an *average-optimal policy* if $g_i(R^*) = g_i$, $i \in E$. For this criterion, a deterministic and stationary optimal policy f^* also exists (cf. Blackwell (1962)). Furthermore, it can be established (cf. Kallenberg (1983)) that f^* is also an optimal policy for the stronger criterion based on the definition

$$\hat{g}_i(R) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_R[r_{X_t Y_t} | X_1 = i] \ . \tag{19}$$

The approach of the average reward case by the concept of superharmonicity is due to Hordijk/Kallenberg (1979), where the proof of the results of this section can be found.

A vector $v \in \mathbb{R}^N$ is said to be *average-superharmonic* if there exists a vector $u \in \mathbb{R}^N$ such that the pair (v, u) satisfies

$$\begin{cases} v_i \geq \sum_j p_{iaj} v_j & \text{for every } (i, a) \in E \times A \\ v_i + u_i \geq r_{ia} + \sum_j p_{iaj} u_j & \text{for every } (i, a) \in E \times A \end{cases} \tag{20}$$

Theorem 7: The average-value-vector g is the (componentwise) smallest average-superharmonic vector.

Corollary: g is the unique v-part of an optimal solution (v, u) of the linear primal program

$$\min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij} - p_{iaj}) v_j \geq 0 & \text{for every } (i, a) \in E \times A \\ v_i + \sum_j (\delta_{ij} - p_{iaj}) u_j \geq r_{ia} & \text{for every } (i, a) \in E \times A \end{array} \right\} . \tag{21}$$

where $\beta_j > 0$, $j \in E$, is arbitrarily chosen.
The dual linear program is

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \ , & j \in E \\ \sum_a x_{ja} + \sum_{i,a} (\delta_{ij} - p_{iaj}) y_{ia} = \beta_j \ , & j \in E \\ x_{ia}, y_{ia} \geq 0 \ , & (i, a) \in E \times A \end{array} \right\} . \tag{22}$$

Theorem 8: Let (x^*, y^*) be an extreme optimal solution of (22). Then, any deterministic and stationary policy f^* with $f^*(i) = a_i$, where $x^*_{ia_i} > 0$ if $\sum_a x^*_{ia} > 0$ and $y^*_{ia_i} > 0$ if $\sum_a x^*_{ia} = 0$, is an average-optimal policy.

Remarks:

1) Any feasible solution (x, y) of (22) satisfies $\sum_a x_{ia} + \sum_a y_{ia} > 0$, $i \in E$.
2) There are examples in which the rule, given in Theorem 7, can yield more than one optimal solution: it is possible to obtain a solution which has in some state i more than one positive x_{ia} or y_{ia} variable.
3) In the average reward case there is no one-to-one correspondence between the feasible solutions of the dual (18) and the stationary policies.
4) The linear programs (21) and (22) were first presented by Denardo/Fox (1968). They showed that the average-value-vector is obtained by (21). However, they had no satisfactory treatment to find an average-optimal policy. The results of Denardo/Fox are slightly refined by Derman (1970). He has shown that, in order to find an average-optimal policy, there have to be solved two sequential linear programs and a simple search problem, in the worst case. Dirickx/Rao (1979) have proved that one linear program and a search procedure are sufficient to compute an optimal policy. Finally, Hordijk/Kallenberg (1979) established that an optimal policy can be found directed from the dual program.

Although there is no one-to-one correspondence between the feasible solutions of the dual program (22) and the stationary policies, there are interesting relations (cf. Hordijk/Kallenberg (1979)). For any feasible solution (x, y) of (22), we define a stationary policy $\pi(x, y)$ by

$$\pi_{ia}(x, y) := \begin{cases} x_{ia} \Big/ \sum_a x_{ia} & a \in A(i) \ , \quad i \in E_x := \left\{ j \,\middle|\, \sum_a x_{ja} > 0 \right\} \\ y_{ia} \Big/ \sum_a y_{ia} & a \in A(i) \ , \quad i \notin E_x \end{cases} \tag{23}$$

Conversely, consider a stationary policy π, and let $P^*(\pi)$ and $D(\pi)$ be the stationary and the derivation matrix, respectively, of $P(\pi)$, i.e. $P^*(\pi) := \frac{1}{T}\sum_{t=1}^{T} P^t(\pi)$ and $D(\pi) := [I - P(\pi) + P^*(\pi)]^{-1} - P^*(\pi)$. $P^*(\pi)$ indicates the average behaviour and $D(\pi)$ the total deviation $\sum_{t=1}^{\infty} [P^t(\pi) - P^*(\pi)]$ (cf. Veinott (1974)).
Define $(x(\pi), y(\pi))$ by

$$\begin{cases} x_{ia}(\pi) := \left[\sum_k \beta_k p^*_{ki}(\pi)\right] \cdot \pi_{ia} & a \in A(i)\ ,\quad i \in E \\ y_{ia}(\pi) := \left[\sum_k \beta_k d_{ki}(\pi) + \sum_k \gamma_k p^*_{ki}(\pi)\right] \cdot \pi_{ia} & a \in A(i)\ ,\quad i \in E \end{cases} \tag{24}$$

where $\gamma \in \mathbb{R}^N$ is a vector which will guarantee the nonnegativity of $y(\pi)$. γ is zero on the transient states and has a constant value on an ergodic set (different ergodic sets will have different values, in general). The value γ_k for the states k of the ergodic set E_j satisfies

$$\gamma_k := \max_{i \in E_j} \left\{ -\sum_k \beta_k d_{ki}(\pi) \Big/ \sum_k p^*_{ki}(\pi) \right\}\ ,\quad k \in E_j\ . \tag{25}$$

Theorem 9: (i) $(x(\pi), y(\pi))$ is a feasible solution of (22); for a determinisitic and stationary policy f is $(x(f), y(f))$ an extreme feasible solution of (22).
(ii) $\pi(x, y)$, defined by (23), is a stationary policy; it is possible that two different feasible solutions are mapped on the same stationary policy. For an extreme feasible solution (x, y) of (22), $\pi(x, y)$ is in general not a deterministic policy.

Theorem 10: (i) If π is a stationary average-optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of (22).
(ii) If (x, y) is an optimal solution of (22), then $\pi(x, y)$ is an average-optimal policy.

Remark: The Blackwell (1962) version of the Howard's policy improvement method is equivalent to a block pivoting simplex method to solve (22) (cf. Osaki/Mine (1969)).

6 Average Reward Criterion (Special Cases)

In all special cases, there are assumptions about the chain structure which imply that the average-value-vector g is a constant vector. Hence, the linear programs (21) and (22) are reduced to

$$\min \left\{ v \mid v + \sum_j (\delta_{ij} - p_{iaj}) u_j \geq r_{ia},\ (i, a) \in E \times A \right\} \tag{26}$$

and

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \;, & j \in E \\ \sum_{i,a} x_{ia} = 1 & \\ x_{ia} \geq 0 \;, & (i,a) \in E \times A \end{array} \right\} . \tag{27}$$

The average-value-vector is obtained as v-part of the optimal solution of the primal program. An optimal policy cannot always directly be obtained from the dual. We will distinguish several cases. The proofs of the results of this section can be found in Kallenberg (1983).

The Irreducible Case

This case is the classical one and linear programming in this case was first presented by Manne (1960) and independently by De Ghellinck (1960). For this case we have the following assumption.

Assumption: For every stationary and deterministic policy f, the Markov chain induced by the transition matrix $P(f)$ is irreducible, i.e. has one ergodic class and no transient states.

Remark: There exists an efficient algorithm (i.e. polynomial in the number of states and actions) to verify the above assumption (cf. Kallenberg (1994)).

All properties of the discounted reward case are also valid in the irreducible case (see also Remark (2) below). We summarize the results.

Theorem 11: Let x^* be an extreme optimal solution of (27). Then, for every $i \in E$, $x^*_{ia} > 0$ for exactly one action, say a_i, and the deterministic and stationary policy f^* with $f^*(i) = a_i$, $i \in E$, is an average-optimal policy.

Theorem 12: The mapping $x_{ia} \to \pi(x)$ with $\pi_{ia}(x) := x_{ia} / \sum_a x_{ia}$ is a one-to-one mapping of the feasible solutions of (27) onto the stationary policies with $\pi \to x_{ia}(\pi)$, where $x_{ia}(\pi) := p_i^*(\pi) \cdot \pi_{ia}$, as inverse mapping. The set of deterministic and stationary policies corresponds to the set of extreme solutions of (27).

Remarks:

1) The variables of the dual program (27) have the interpretation of the stationary state-action frequencies, i.e. $x_{ia}(\pi)$ is the average number of times that state i is entered and action a is chosen, given policy π.
2) Let $x_{ia}^{\alpha}(\pi)$ be the state-action frequencies in the α-discounted case. From Abelian arguments it is well-known that $x_{ia}(\pi) = \lim_{\alpha \uparrow 1}(1-\alpha)x_{ia}^{\alpha}(\pi)$ for every (i, a) and π. With this observation the linear program (27) can be considered as a limiting case of a reformulated linear program for the discounted case. Replace in the usual linear program (5) x_{ia}^{α} by $(1-\alpha)x_{ia}^{\alpha}$ and notice that the equations of (5) add to $(1-\alpha)\sum_{j,a} x_{ja}^{\alpha} = \sum_j \beta_j$. Take $\beta_j > 0$, $j \in E$, such that $\sum_j \beta_j = 1$. With this transformation (5) is equivalent to

$$\max \left\{ \sum_{i,a} r_{ia}(1-\alpha)x_{ia}^{\alpha} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj})(1-\alpha)x_{ia}^{\alpha} = (1-\alpha)\beta_j\,, & j \in E \\ \sum_{j,a} (1-\alpha)x_{ja}^{\alpha} = 1 & \\ (1-\alpha)x_{ia}^{\alpha} \geq 0\,, & (i,a) \in E \times A \end{array} \right\}. \tag{28}$$

Let $\alpha \uparrow 1$, then $(1-\alpha)x_{ia}^{\alpha} \to x_{ia}$ and we obtain

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj})x_{ia} = 0\,, & j \in E \\ \sum_{i,a} x_{ia} = 1 & \\ x_{ia} \geq 0\,, & (i,a) \in E \times A \end{array} \right\},$$

which is program (27). A similar observation can be found in Nazareth/Kalkurni (1986).

The Unichain Case

Assumption: For every stationary and deterministic policy f, the Markov chain chain induced by the transition matrix $P(f)$ has one ergodic set plus a (perhaps empty) set of transient states.

In this case the dual program (27) yields an optimal policy by the following result.

Theorem 13: Let x^* be an extreme optimal solution of (27). Take f^* such that $f^*(i) = a_i$, $i \in E$, where $x^*_{ia_i} > 0$ for every $i \in E^* := \{j | \sum_a x^*_{ja} > 0\}$. Then, f^* is an average-optimal policy.

Remarks:

1) In the unichain case there is in general no one-to-one correspondence between the feasible solutions of (27) and the stationary policies.
2) Denardo (1973) has shown that, with a modified rule for the variable that leaves the basis, the simplex method to solve (27) is equivalent to a version of the policy iteration procedure.

The Communicating Case

We say that a state j is *accessible* from state i if there exists a deterministic and stationary policy f such that $[P^n(f)]_{ij} > 0$ for some $n \in \mathbb{N}_0$, where the policy f may depend on i and j. In the communicating case we have the following condition.

Assumption: State j is accessible from state i (and vice versa) for each two states i and j.

In this case an optimal solution x^* of the dual program (27) yields only optimal decisions for the states of the set E^*. Outside this set we need a search procedure as described in the next algorithm.

Algorithm 1:

1. Let x^* be an extreme optimal solution of program (22) and f^* defined on E^* as in Theorem 13.
2. If $E^* = E$, then f^* is an average-optimal policy (STOP).
   Otherwise: go to step 3.
3. a. Choose a triple (i, a_i, j) with $i \notin E^*$, $j \in E^*$ and $p_{ia_ij} > 0$.
   b. $f^*(i) := a_i$; $E^* := E^* \cup \{i\}$; go to step 2.

Theorem 14: Algorithm 1 is finite and terminates with an average-optimal policy f^*.

Remarks:

1) In fact, the communicating assumption can be relaxed. It is sufficient to require the assumption only for the optimal policies.
2) The verification of the assumption can be executed in polynomial time (cf. Kallenberg (1994)) or by linear programming (cf. Filar/Schultz (1988)).

Denardo (1970a) has shown that (27) can also be used to obtain an average-optimal policy in the multichain case. This is based on the following result.

Theorem 15: Let E^* and x^* as in Theorem 13. Then,

i) For every $i \in E^*$, $x^*_{ia} > 0$ for exactly one action, say a_i.
ii) The policy f^* with $f^*(i) = a_i$, $i \in E^*$ is average-optimal on E^* with value $\sum_{i,a} r_{ia} x^*_{ia}$.
iii) E^* is an ergodic set under $P(f^*)$ for any extension of f^* to E.

Hence, by a search procedure similar to the one in Algorithm 1, the optimal policy f^* can be extended with some transient states. If f^* cannot be extended to the whole state space, then the remaining states are closed under any policy. Therefore, the same approach is repeated on the smaller state space of the remaining states until f^* is defined on the whole state space. Then, f^* is an average optimal policy and we obtain the next algorithm.

Algorithm 2:

1. Define $E_0 := E$.
2. Let x^* be an extreme optimal solution of program (22) with state space E_0, and let f^* and E^* be defined as in Theorem 13.
3. If $E^* = E_0$, then f^* is an average-optimal policy (STOP).
   Otherwise: $E_0 := E_0 \backslash E^*$ and go to step 4.
4. If there exists a triple (i, a_i, j) with $i \in E_0, j \notin E_0$ and $p_{ia_ij} > 0$, then: $f^*(i) := a_i$, $E^* := E^* \cup \{i\}$, $E_0 := E_0 - \{i\}$ and repeat step 4.
5. If $E_0 = \varnothing$, then STOP, else go to step 2.

7 Bias Optimality

In some situations average-optimality is not satisfactory because it only considers the limiting behaviour. Rewards earned in transient states do not influence

the average reward. The concept of *bias-optimality* is more selective. This criterion was introduced by Blackwell (1962) under the term "nearly optimal". Blackwell examined the limiting behaviour of $v^\alpha(f)$ as $\alpha \uparrow 1$ and obtained the following asymptotic expression

$$v^\alpha(f) = (1-\alpha)^{-1} g(f) + u(f) + h(\alpha) \tag{29}$$

where $u(f) := D(f)r(f)$, with $D(f)$ the deviation matrix, is called the bias term and $\lim_{\alpha \uparrow 1} h(\alpha) = 0$.

A policy R^* is called bias-optimal if

$$\liminf_{\alpha \uparrow 1} \{v_i^\alpha(R^*) - v_i^\alpha\} = 0 \ , \quad i \in E \ . \tag{30}$$

Blackwell also showed the existence of a deterministic and stationary bias-optimal policy. As shown by Denardo/Miller (1968) and stated in the following theorem, there are several equivalent formulations for bias-optimality.

Theorem 16: The following statements are equivalent:

i) f^* is bias-optimal.
ii) $\lim_{\alpha \uparrow 1} [v^\alpha(f^*) - v^\alpha(f)] \geq 0$ for every policy f.
iii) $u(f^*) = \max_f \{u(f) | g(f) = g\}$.
iv) $\lim_{T \to \infty} (1/T) \sum_{t=1}^{T} \sum_{s=1}^{t} [P^s(f^*) - P^s(f)] \geq 0$ for every policy f.

$u := \max_f \{u(f) | g(f) = g\}$ is called the *bias-value-vector*. Let F_g be the set of deterministic, stationary average-optimal policies and F_b the subset of F_g consisting of the bias-optimal policies. Then, $u = \max_f \{u(f) | f \in F_g\} = u(h)$ for every $h \in F_b$.

A first algorithm to compute a bias-optimal policy was presented by Veinott (1966). It was based on policy iteration. Denardo (1970b) has refined this method to a three-step procedure, in which each step can be executed by linear programming or by policy iteration. Kallenberg (1981b) has slightly improved Denardo's procedure.

Again, as in average-optimality, we will distinguish between the general case and a special case.

General Case

In the general case we have a *three step procedure* in which the first step is the same as in the average reward case, i.e. the linear programs (21) and (22) are

solved with optimal solutions (g, u^*) and (x^*, y^*), respectively. Then, consider the subsets of the actions which yield equality in (21), i.e.

$$A_1(i) := \left\{ a \in A(i) | g_i = \sum_j p_{iaj} g_j \right\}, \quad i \in E . \tag{31}$$

$$A_2(i) := \left\{ a \in A_1(i) | g_i + u_i^* = r_{ia} + \sum_j p_{iaj} u_j^* \right\}, \quad i \in E . \tag{32}$$

Let R^* be the set of states that are recurrent under at least one bias-optimal policy and $E^* := \{i \in E | A_2(i) \neq \varnothing\}$. Notice that E^* is known, but R^* not. The proofs of the following results can be found in Kallenberg (1983), chapter 5.

Theorem 17: (i) For every $f \in F_g$, we have $f(i) \in A_1(i)$ for every $i \in E$, and $f(i) \in A_2(i)$ for every i that is recurrent under $P(f)$.
(ii) $u_i = u_i^* + \max_f \{[P^*(f)(-u^*)]_i | f \in F_g\}$, $i \in R^*$.

In the second step we try to compute $\max_f \{[P^*(f)(-u^*)]_i | f \in F_g\}$, $i \in R^*$. However, F_g and R^* are unknown. Since Theorem 15 implies that $R^* \subseteq E^*$ and $f(i) \in A_2(i)$, $f \in F_g$, $i \in R^*$, we compute $\max_f \{[P^*(f)(-u^*)]_i | f(i) \in A_2(i)\}$, $i \in E^*$.

Therefore, we will consider the MDP with the state space E^* and with action sets $A_2(i)$. A difficulty is that E^* is, in general, not closed under the actions from $A_2(i)$. To obtain a closed model the actions $a \in A_2(i)$, such that $p_{iaj} > 0$ for at least one $j \notin E^*$, have to be removed from $A_2(i)$. The restricted action sets are again notated by $A_2(i)$.

Since $g(f) = P^*(f)r(f)$, in order to compute $\max_f \{[P^*(f)(-u^*)]_i | f(i) \in A_2(i)\}$, we use the model with immediate rewards $r_{ia}^* := -u_i^*$, $i \in E^*$, $a \in A_2(i)$. In this altered model the average-value-vector is the v-part of the following linear program

$$\min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij} - p_{iaj}) v_j \geq 0 & \text{for every } (i, a) \in E^* \times A_2 \\ v_i + \sum_j (\delta_{ij} - p_{iaj}) w_j \geq -u_i^* & \text{for every } (i, a) \in E^* \times A_2 \end{array} \right\} \tag{33}$$

with

$$\max \left\{ \sum_{i,a} -u_i^* s_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) s_{ia} = 0, & j \in E^* \\ \sum_a s_{ja} + \sum_{i,a} (\delta_{ij} - p_{iaj}) t_{ia} = \beta_j, & j \in E^* \\ s_{ia}, t_{ia} \geq 0, & (i, a) \in E^* \times A_2 \end{array} \right\} . \tag{34}$$

as dual program.

Theorem 18: Let f^* be an average-optimal policy obtained from the optimal solution (s^*, t^*) of (34), and let (v^*, w^*) be an optimal solution of (33). Then,

$$\begin{cases} u_i = u_i^* + v_i^* = u_i^*(f^*) \ , & i \in R^* \\ u_i \geq u_i^* + v_i^* \ , & i \in E^* \setminus R^* \ . \end{cases}$$

Remark: The policy f^* is bias-optimal for the states which are recurrent under at least one bias-optimal policy and $u_i^* + v_i^*$ is the bias-value on R^*. Unfortunately, also at this point we do not know which states are in R^*.

Finally, for the third part of the procedure, we introduce a concept of bias-superharmonicity. A vector $z \in \mathbb{R}^N$ is called *bias-superharmonic* if

$$\begin{cases} z_i \geq (r_{ia} - g) + \sum_j p_{iaj} z_j & \text{for every } (i, a) \in E \times A \\ z_i \geq u_i^* + v_i^* & \text{for every } i \in E^* \end{cases} \tag{35}$$

Theorem 19: The bias-value-vector u is the (componentwise) smallest bias-superharmonic vector.

Corollary: The bias-value-vector u is the unique optimal solution of the linear program

$$\min \left\{ \sum_j \beta_j z_j \;\middle|\; \begin{array}{ll} \sum_j (\delta_{ij} - p_{iaj}) z_j \geq (r_{ia} - g) & \text{for every } (i, a) \in E \times A_1 \\ z_i \geq u_i^* + v_i^* & \text{for every } i \in E^* \end{array} \right\} . \tag{36}$$

The dual program is

$$\max \left\{ \sum_{i,a} (r_{ia} - g_i) x_{ia} + \sum_i (u_i^* + v_i^*) y_i \;\middle|\; \begin{array}{l} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} + \sum_i \delta_{ij} y_i = \beta_j \ , j \in E \\ x_{ia} \geq 0, (i, a) \in E \times A_1 ; \ y_i \geq 0, i \in E^* \end{array} \right\} . \tag{37}$$

Remark: The next theorem shows that a deterministic and stationary bias-optimal policy can be obtained directly from program (37). By the simplex

method both programs (36) and (37) are solved simultaneously and yield the bias-value-vector u as well as a bias-optimal policy.

Theorem 20: Let (x^*, y^*) be an extreme optimal solution of (37), z^* an optimal solution of (36) and let f^* be the policy on E^*, mentioned in Theorem 18. Then, the bias-value-vector $u = z^*$ and the policy f defined by

$$f(i) := \begin{cases} f^*(i) , & i \in \{ j \in E^* | z_j^* = u_j^* + v_j^* \} \\ a_i \in A_1(i) \text{ such that } x_{ia_i}^* > 0 , & i \notin \{ j \in E^* | z_j^* = u_j^* + v_j^* \} \end{cases} \tag{38}$$

is a bias-optimal policy.

Irreducible Case

As special case we will only consider the irreducible case. Other special cases can be found in Kallenberg (1983). Irreducibility implies that the average-value-vector is a constant g and that $R^* = E^* = E$. Hence part 3 of the algorithm is superfluous and the linear programs of the first two parts can be simplified. This results in the following algorithm.

Algorithm 3

1. a. Determine an optimal solution $(h^* = g, u^*)$ of the linear program

$$\min \left\{ h | h + \sum_j (\delta_{ij} - p_{iaj}) u_j \geq r_{ia}, (i, a) \in E \times A \right\} .$$

   b. Let $A_2(i) := \{ a \in A(i) | \sum_j (\delta_{ij} - p_{iaj}) u_j^* = r_{ia} - g \}$.
2. a. Determine optimal solutions (v^*, w^*) and s^* of the linear programs

$$\min \left\{ v | v + \sum_j (\delta_{ij} - p_{iaj}) w_j \geq -u_i^* \text{ for every } (i, a) \in E \times A_2 \right\}$$

   and its dual

$$\max \left\{ \sum_{i,a} -u_i^* s_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) s_{ia} = 0 \;, & j \in E \\ \sum_{i,a} s_{ia} = 1 & \\ s_{ia} \geq 0 \;, & (i,a) \in E \times A_2 \end{array} \right\} .$$

b. $u_i := u_i^* + v^*$, $i \in E$ is the bias-value-vector and f such that $s^*_{if(i)} > 0$, $i \in E$, is a bias-optimal policy.

8 Blackwell Optimality

A policy R^* is called *Blackwell-optimal* if, for some $0 < \alpha_0 < 1$, R^* is α-discounted optimal for every $\alpha \in [\alpha_0, 1)$. The terminology Blackwell-optimal is used in honour to Blackwell (1962) who has shown that a deterministic and stationary Blackwell-optimal policy exists. Blackwell-optimality implies bias-optimality and therefore also average-optimality. Blackwell did not give a method to compute a Blackwell-optimal policy. The first method to compute a Blackwell-optimal policy was presented by Miller/Veinott (1969) and was based on policy iteration. The linear programming approach is due to Hordijk/ Dekker/ Kallenberg (1985), where the proofs can be found of the following statements. Instead of the discount factor α, we can also use the interest rate ρ, defined by $\rho := (1 - \alpha)/\alpha$. Then, $\alpha = (1 + \rho)^{-1}$ and $\alpha \uparrow 1$ is equivalent to $\rho \downarrow 0$. Since the components of the α-discounted reward vector $v^\alpha(f)$, for a deterministic and stationary policy f, can be computed as unique solution of the linear system $[I - \alpha P(f)]x = r(f)$, they are obtained by the unique solution of the system

$$[(1 + \rho)I - P(f)]x = (1 + \rho)r(f) \;. \tag{39}$$

Solving (39) by Cramer's rule shows that $v^\rho(f) := v^\alpha(f)$ is a rational function (in ρ), say $\dfrac{p(\rho)}{q(\rho)}$. Therefore, we consider the following set RF of rational functions in ρ:

$$RF := \left\{ \frac{p(\rho)}{q(\rho)} \;\middle|\; \begin{array}{l} p(\rho) = \sum_{k=0}^{n} a_k \rho^k, \; a_k \in \mathbb{R}, \text{ for some } n \in \mathbb{N}_0 \\ q(\rho) = \sum_{k=0}^{m} b_k \rho^k, \; b_k \in \mathbb{R}, \text{ for some } m \in \mathbb{N}_0 \end{array} \right\} \tag{40}$$

In RF we use the natural operations $+$ and $\cdot$ for addition and multiplication. The *index* $\mu(p)$ of the polynomial $p(\rho) = \sum_{k=0}^{n} a_k \rho^k$ is defined by

$$\mu(p) := \min\{k \in \mathbb{N}_0 | a_k \neq 0\} \ . \tag{41}$$

With this concept a complete ordering (denoted by $>_l$) in RF can be obtained as follows. Let $p(\rho) = \sum_{k=0}^{n} a_k \rho^k$ and $q(\rho) = \sum_{k=0}^{m} b_k \rho^k$, then

$$\frac{p(\rho)}{q(\rho)} >_l p_0(\rho) \quad \text{iff} \quad a_{\mu(p)} \cdot b_{\mu(q)} > 0 \ , \tag{42}$$

where $p_0(\rho) \equiv 0$ is the zero-element of RF.

This ordering means that $\frac{p(\rho)}{q(\rho)} >_l p_0(\rho)$ if $\frac{p(\rho)}{q(\rho)}$ is positive for ρ sufficiently near to 0, i.e. $\frac{p(\rho)}{q(\rho)} > 0$ for all $\rho \in (0, \rho_0]$ for some $\rho_0 > 0$.

Theorem 21: RF is a completely ordered commutative field.

An N-vector $v(\rho)$ with components from RF is called *Blackwell-superharmonic* if

$$(1 + \rho) \cdot v_i(\rho) \geq_l (1 + \rho) r_{ia} + \sum_j p_{iaj} v_j(\rho) \quad \text{for every } (i, a) \in E \times A \tag{43}$$

Let f^* be a Blackwell-optimal policy. Then, $v^\rho := v^\rho(f^*)$, with components in RF, is called the *Blackwell-value-vector*.

Theorem 22: The Blackwell-value-vector v^ρ is the (componentwise) smallest (with respect to the ordering $>_l$) Blackwell-superharmonic vector.

Theorem 22 implies that the Blackwell-value-vector v^ρ on the interval $(0, \rho_0]$ can be found as optimal solution of the following linear program in RF:

$$\min\left\{\sum_j v_j(\rho) \,\middle|\, \sum_j [(1 + \rho)\delta_{ij} - p_{iaj}] v_j(\rho) \geq_l (1 + \rho) r_{ia}, (i, a) \in E \times A\right\} .$$

Consider also the following linear program in RF, called the *dual program*:

$$\max\left\{\sum_{i,a} (1+\rho) r_{ia} \cdot x_{ia}(\rho) \,\middle|\, \begin{array}{l} \sum_{i,a} [(1+\rho)\delta_{ij} - p_{iaj}] \cdot x_{ia}(\rho) =_l p_1(\rho), j \in E \\ \\ x_{ia}(\rho) \geq_l p_0(\rho), (i, a) \in E \times A \end{array}\right\} ,$$

where $p_1(\rho) \equiv 1$ is the identity.

For a fixed real value of ρ, the above linear programs are the linear programs of section 3 with $\beta_j = 1$ for every $j \in E$. Also in section 3, it is mentioned that there is a one-to-one correspondence between the extreme points of the dual program and the set of deterministic and stationary policies.

In RF we can, as is also the case in the usual simplex method, rewrite the equalities $\sum_{i,a}[(1+\rho)\delta_{ij} - p_{iaj}] \cdot x_{ia}(\rho) =_l p_1(\rho)$, $j \in E$, such that at each iteration there is precisely one positive $x(\rho)$ component in each state. The main difference with the usual simplex method for a fixed value of ρ is that instead of real numbers, the elements are rational functions. Hence, a Blackwell-optimal policy can be obtained as follows (for the details we refer to Hordijk/Dekker/Kallenberg (1985).

1. Start with the simplex tableau corresponding to an arbitrary deterministic policy.
2. If every reduced cost is nonnegative with respect to the ordering in RF, then the corresponding policy is Blackwell optimal.
   Otherwise: take any column with a negative reduced cost as pivot column and execute a pivot transformation.
3. Go to step 2.

Remarks:

1) It can be shown that the polynomials in the tableau have a degree of at most N. Furthermore, the elements in tableau have a common denumerator.
2) The computation of each new simplex tableau has complexity $\mathcal{O}(AN^3)$, where $A = \sum_i \# A(i)$.

9 Markov Games

In this section we consider the two-person zero-sum Markov game. For this model there are in each state i two action sets: $A(i)$ and $B(i)$ for player 1 and player 2, respectively. If in state i player 1 chooses action $a \in A(i)$ and player 2 action $b \in B(i)$, then the following occurs:

- player 1 receives an immediate reward r_{iab} from player 2;
- the process moves to state j with probability p_{iabj}, $j \in E$.

The subject of Markov games was introduced by Shapley (1953), even before the first paper on Markov decision processes.

In Markov games the player 1 wants to maximize his rewards and player 2 wants to minimize his payments. We will consider two optimality criteria: discounted rewards and average rewards. The analysis of this section is along the lines of Horkijk/Kallenberg (1981a, 1981b), where also the proofs can be found.

Discounted Rewards

By $v_i^\alpha(R_1, R_2)$, $i \in E$, we denote the total expected discounted reward for player 1, when player 1 uses policy R_1, player 2 policy R_2 and state i is the starting state. We call a policy R_1^* *discounted-optimal for player 1* if

$$v_i^\alpha(R_1^*, R_2) \geq \inf_{R_2} \sup_{R_1} v_1^\alpha(R_1, R_2) \text{ for all policies } R_2 \text{ and all } i \in E \ . \tag{44}$$

The policy R_2^* is *discounted-optimal for player 2* if

$$v_i^\alpha(R_1, R_2^*) \leq \sup_{R_1} \inf_{R_2} v_1^\alpha(R_1, R_2) \text{ for all policies } R_1 \text{ and all } i \in E \tag{45}$$

If $\sup_{R_1} \inf_{R_2} v_i^\alpha(R_1, R_2) = \inf_{R_1} \sup_{R_2} v_1^\alpha(R_1, R_2)$, then we say that the game has a *value-vector* v_i^α, where

$$v_i^\alpha := \sup_{R_1} \inf_{R_2} v_i^\alpha(R_1, R_2) \ , \quad i \in E \ . \tag{46}$$

The usual way to prove optimality for the policies R_1^* and R_2^* is to show that

$$v_i^\alpha(R_1, R_2^*) \leq v_i^\alpha(R_1^*, R_2^*) \leq v_i^\alpha(R_1^*, R_2) \text{ for every pair of policies } (R_1, R_2) \tag{47}$$

Then, R_1^* and R_2^* are optimal policies for player 1 and player 2, respectively, and the game has the value-vector $v_i^\alpha = v_i^\alpha(R_1^*, R_2^*)$, $i \in E$.

A vector $v \in \mathbb{R}^N$ is said to be *discounted-superharmonic* for a two-person zero-sum game if there exists a stationary policy ρ for player 2 such that

$$v_i \geq r_{ia}(\rho) + \alpha \cdot \sum_j p_{iaj}(\rho) v_j \quad \text{for every } (i, a) \in E \times A \ , \tag{48}$$

where $r_{ia}(\rho) := \sum_b r_{iab} \rho_{ib}$ and $p_{iaj}(\rho) := \sum_b p_{iabj} \rho_{ib}$.

Theorem 23: The game has a value-vector and the value-vector is the (componentwise) smallest discounted-superharmonic vector.

Corollary: The value-vector is the unique v-part in the optimal solution of the following mathematical program (cf. Rothblum (1979)) in which the objective function is linear and the constraints are both linear and quadratic:

$$\min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{l} \sum_j \left[\delta_{ij} - \alpha \sum_b p_{iabj} \rho_{ib} \right] v_j - \sum_b r_{iab} \rho_{ib} \geq 0 \;, \quad (i,a) \in E \times A \\ \sum_b \rho_{ib} = 1 \;, \quad i \in E \\ \rho_{ib} \geq 0 \;, \quad (i,b) \in E \times B \end{array} \right\} \tag{49}$$

where $\beta_j > 0, j \in E$.

Pathasarathy/Raghavan (1981) have shown that even when the data (rewards, transition probabilities and discount factor) are rational numbers it is possible that the value-vector has irrational components. Hence, the components of the value-vector are not in the same ordered field as the data. Therefore, the value-vector cannot be computed as solution of a linear program, because in linear programming all elements stay in the same field as the field generated by the data.

In order to have the value-vector in the same order field as the field generated by the data, we need an additional assumption. This assumption and the corresponding linear programming solution were proposed in Parthasarathy/Raghavan (1981), who have shown the result of the next Theorem 24.

Assumption: The transition probabilities $p_{iabj}, j \in E$, do not depend on b.

Because of the above assumption (with writing p_{iaj} instead of p_{iabj}), program (49) becomes the following linear program

$$\min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{l} \sum_j [\delta_{ij} - \alpha p_{iaj}] v_j - \sum_b r_{iab} \rho_{ib} \geq 0 \;, \quad (i,a) \in E \times A \\ \sum_b \rho_{ib} = 1 \;, \quad i \in E \\ \rho_{ib} \geq 0 \;, \quad (i,b) \in E \times B \end{array} \right\} \tag{50}$$

with as dual

$$\max \left\{ \sum_i z_i \;\middle|\; \begin{array}{lll} \sum_{i,a} \left[\delta_{ij} - \alpha \sum_j p_{iaj} \right] x_{ia} & = \beta_j \;, & j \in E \\ -\sum_a r_{iab} x_{ia} + z_i & \leq 1 \;, & i \in E \\ x_{ia} & \geq 0 \;, & (i, a) \in E \times A \end{array} \right\} . \tag{51}$$

Theorem 24: Let (v^*, ρ^*) and (x^*, z^*) be optimal solutions of the linear programs (50) and (51), respectively. Define the stationary policy π^* by

$$\pi^*_{ia} := x^*_{ia} \Big/ \sum_a x^*_{ia} \;, \quad (i, a) \in E \times A \;. \tag{52}$$

Then π^* and ρ^* are stationary optimal policies for player 1 and player 2, respectively, and v^* is the value-vector of the game.

Average Rewards

In the average reward case the treatment is more complicated. The first paper with this criterion is Gillette (1957). He showed that, in general, in the average reward case there are no optimal policies in the class of stationary policies. The question whether or not this game posses a value-vector was open for a long time. Mertens/Neyman (1981) have shown that the value-vector exists.

By $g_i(R_1, R_2)$, $i \in E$, we denote the average expected reward for player 1, when player 1 uses policy R_1, player 2 policy R_2 and state i is the starting state. The concepts of optimality and value-vector are defined as in the discounted case.

A vector $g \in \mathbb{R}^N$ is said to be *average-superharmonic* for the two-person zero-sum game if there exist a vector $t \in \mathbb{R}^N$ and a stationary policy ρ for player 2 such that

$$\begin{cases} g_i \geq \sum_j p_{iaj}(\rho) g_j & \text{for every } (i, a) \in E \times A \\ g_i + t_i \geq r_{ia}(\rho) + \sum_j p_{iaj}(\rho) t_j & \text{for every } (i, a) \in E \times A \end{cases} \tag{53}$$

Theorem 25: If there exist stationary average-optimal policies for both players, then the game has a value-vector and the value-vector is the (componentswise) smallest average-superharmonic vector.

Remarks:

1) In Bewley/Kohlberg (1978) sufficient conditions are given for the existence of stationary average-optimal policies. An example of such a condition is that only the first player controls the transitions.
2) From Theorem 25 a program with linear objective and quadratic constraints can be derived to compute the value-vector (if there are stationary optimal policies).

For the remaining part of this section we assume that the transitions are controlled by player 1. Then, the program becomes linear, namely

$$\min \left\{ \sum_j \beta_j g_j \;\middle|\; \begin{array}{rll} \sum_j [\delta_{ij} - p_{iaj}] g_j \geq 0 , & (i,a) \in E \times A \\ g_i + \sum_j [\delta_{ij} - p_{iaj}] t_j - \sum_b r_{iab} \rho_{ib} \geq 1 , & (i,a) \in E \times A \\ \sum_b \rho_{ib} = 1 , & i \in E \\ \rho_{ib} \geq 0 , & (i,b) \in E \times B \end{array} \right\} \tag{54}$$

with dual program

$$\max \left\{ \sum_i z_i \;\middle|\; \begin{array}{rll} \sum_{i,a} [\delta_{ij} - p_{iaj}] x_{ia} = 0 , & j \in E \\ \sum_a x_{ja} + \sum_{i,a} [\delta_{ij} - p_{iaj}] y_{ia} = \beta_j , & j \in E \\ -\sum_a r_{iab} x_{ia} + z_i \leq 0 , & (i,b) \in E \times B \\ x_{ia}, y_{ia} \geq 0 ; & (i,a) \in E \times A \end{array} \right\} \tag{55}$$

Theorem 26: Let (g^*, t^*, ρ^*) and (x^*, y^*, z^*) be optimal solutions of the linear programs (54) and (55), respectively. Define the stationary policy π^* by

$$\pi^*_{ia} := \begin{cases} x^*_{ia} \Big/ \sum_a x^*_{ia} & a \in A(i) , \quad i \in \left\{ j \;\middle|\; \sum_a x^*_{ja} > 0 \right\} \\ y^*_{ia} \Big/ \sum_a y^*_{ia} & a \in A(i) , \quad i \notin \left\{ j \;\middle|\; \sum_a x^*_{ja} > 0 \right\} \end{cases}$$

Then, π^* and ρ^* are optimal policies for player 1 and player 2, respectively, and g^* is the value-vector of the game.

Remarks:

1) The first finite algorithm for this undiscounted single-controller Markov game was given by Filar (1980). The linear programming approach, as presented above, is due to Hordijk/Kallenberg (1981b) and independently Vrieze (1981).
2) In special cases like the unichain case the algorithm can be simplified similar to the results of section 6.

B Nonstandard Markovian Control Problems

To formulate *MDP*'s only in terms of the standard utility functions can be quite insufficient. One might formulate problems in terms of probabilistics of several characteristics and optimize a certain utility subject to constraints on these probabilities. The first reference in this are is the paper Derman/Klein (1965). They consider an inventory problem in which constraints are formulated on the probability of shortage. Other applications can be found in queueing (e.g. limits on the average throughput), multiple objectives or mean-variance tradeoffs.

10 Additional Constraints

For problems with additional constraints there are in general two approaches: the first based on Lagrangian multipliers (cf. Beutler/Ross (1985, 1986)), the second on linear programming with variables that can be interpreted as state-action frequencies. We will follow the last approach. To be succesfull in this approach it is necessary to formulate the probabilistic constraints in terms of state-action frequencies. For constrained problems the nice property (which hold for standard *MDP*'s) of the existence of a deterministic and stationary optimal policy is no longer valid. Even optimality simultaneously for all starting states does not hold. Therefore, we will optimize with respect to a given initial distribution β, i.e. β_j is the probability that state j is the starting state, $j = 1, 2, \ldots, N$.

Finite Horizon

Let us consider a finite horizon *MDP*. We may assume that the rewards and the transition probabilities are *nonstationary*. Let $r_{ia}(t)$ and $p_{iaj}(t)$ be the immediate reward and the transition probabilities in period t, $1 \leq t \leq T$. As utility we take the *total expected reward* over the T periods, given the initial distribution β. Let $w^T(R)$ be this utility for policy R, i.e.

$$w^T(R) := \sum_j \beta_j \cdot \left\{ \sum_{t=1}^{T} \sum_{i,a} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \cdot r_{ia}(t) \right\} \tag{56}$$

For the additional constraints we assume that besides the reward $r_{ia}(t)$ there are certain costs, say $c_{ia}^k(t)$, when action a is chosen in state i in period t, $k = 1, 2, \ldots, m$.

A policy R is *feasible* if the total expected costs over the T periods, denoted by $c_k^T(R)$, where $c_k^T(R) := \sum_j \beta_j \cdot \{\sum_{t=1}^{T} \sum_{i,a} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \cdot c_{ia}^k(t)\}$, is at most b_k, $k = 1, 2, \ldots, m$.

We want to determine an *optimal policy*, i.e. a policy that maximizes $w^T(R)$ over the feasible policies. $w^T := \sup_R \{w^T(R) | c_k^T(R) \leq b_k, k = 1, 2, \ldots, m\}$ is the *value* of the constrained problem.

The linear program approach for this problem is as follows (cf. Kallenberg (1981a)). Consider

$$\max \left\{ \sum_{i,a,t} r_{ia}(t) x_{ia}(t) \;\middle|\; \begin{array}{ll} \sum_a x_{ja}(1) & = \beta_j, j \in E \\ \sum_a x_{ja}(t+1) - \sum_{i,a} p_{iaj}(t) x_{ia}(t) & = 0, j \in E, 1 \leq t \leq T-1 \\ x_0 - \sum_{i,a} x_{ia}(T) & = 0 \\ \sum_{i,a,t} c_{ia}^k(t) x_{ia}(t) & \leq b_k, 1 \leq k \leq m \\ x_{ia}(t) & \geq 0, (i,a) \in E \times A, 1 \leq t \leq T \end{array} \right\} \tag{57}$$

The variable x_0 comes from an additional state, say 0, where the process is assumed to end after time point T. The next theorem shows that linear programming can be used to obtain an optimal policy.

Theorem 26: (i) Program (57) is infeasible if and only if there is no feasible policy.

(ii) Let $x^*(t)$, $1 \leq t \leq T$, be an optimal solution of the linear program (57). Then, the nonstationary Markov policy $R^* = (\pi^1, \pi^2, \ldots, \pi^T)$ defined by

$$\pi_{ia}^t := \begin{cases} x_{ia}^*(t) \Big/ \sum_a x_{ia}^*(t) & \text{if } \sum_a x_{ia}^*(t) > 0 \\ \text{arbitrary} & \text{if } \sum_a x_{ia}^*(t) = 0 \end{cases} \qquad t = 1, 2, \ldots, T \tag{58}$$

is an optimal policy and $\sum_{t=1}^{T} \sum_{i,a} r_{ia}(t) x_{ia}^*(t)$ is the value.

Remarks:

1) This solution (without the additional constraints) is also valid for the unconstrained case. Then, however, there is no problem of infeasibility. The usual way to solve unconstrained finite horizon *MDP*'s is by backward recursion. In can be shown that a block-pivoting algorithm for the linear program is equivalent to this backward recursion (cf. Hordijk (1978)).
2) $x_{ia}(t)$ has the interpretation of the state-action frequency at time t, namely it is the probability that – given a policy π as defined in (58) – in period t state i is visited and action a is chosen.

Infinite Horizon and Discounted Rewards

For the infinite horizon case we assume (as before) that the immediate rewards and the transition probabilities are stationary. Furthermore, we assume that α is the discount factor and that β is the initial distribution. Let c_{ia}^k be some costs when in state i action a is chosen, $k = 1, 2, \ldots, m$. For a policy R, the total expected discounted reward and costs are denoted by $v^\alpha(R)$ and $c_k^\alpha(R)$, $1 \leq k \leq m$, i.e.

$$v^\alpha(R) := \sum_j \beta_j \cdot \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{i,a} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \cdot r_{ia} \right\}$$

and

$$c_k^\alpha(R) := \sum_j \beta_j \cdot \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{i,a} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \cdot c_{ia}^k \right\}, \quad 1 \leq k \leq m \ .$$

A policy R is *feasible* if $c_k^\alpha(R) \leq b_k$, $k = 1, 2, \ldots, m$. The *value* v^α of the constrained problem is defined by

$$v^\alpha := \sup_R \{v^\alpha(R) | c_k^\alpha(R) \leq b_k, k = 1, 2, \ldots, m\} \ . \tag{59}$$

We want to find an *optimal policy*, i.e. a feasible policy R^* with $v^\alpha(R^*) = v^\alpha$. For any policy R we denote the total expected discounted state-action frequencies, i.e. the number of discounted times of being in state i and then choosing action a, by $x_{ia}^\alpha(R)$, i.e.

$$x_{ia}^\alpha(R) := \sum_j \beta_j \cdot \left\{ \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \right\} \tag{60}$$

Let the sets K, $K(M)$, $K(S)$, $K(DS)$ and P of vectors with components in $E \times A$, be defined by

$$K := \{x^\alpha(R) | R \text{ is arbitrary}\}$$

$$K(M) := \{x^\alpha(R) | R \text{ is a Markov policy}\}$$

$$K(S) := \{x^\alpha(R) | R \text{ is a stationary policy}\}$$

$$K(DS) := \{x^\alpha(R) | R \text{ is a deterministic and stationary policy}\}$$

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

Furthermore, for a set X we denote by $\overline{X}$ the closed convex hull of X.

Theorem 28: $K = K(M) = K(S) = \overline{K(DS)} = P$.

From Theorem 28 it follows that, in order to find an optimal policy for the constrained problem, the linear program

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj}) x_{ia} = \beta_j \ , & j \in E \\ \sum_{i,a} c_{ia}^k x_{ia} \leq b_k \ , & k = 1, 2, \ldots, m \\ x_{ia} \geq 0 \ , & (i, a) \in E \times A \end{array} \right\} \tag{61}$$

has to be considered. Furthermore, it suggestes that an optimal stationary policy can be obtained. The following theorem gives the affirmative answer.

Theorem 29: (i) Program (61) is infeasible if and only if there is no feasible policy. (ii) Let x^* be an optimal solution of (61). Then, the stationary policy π^*, defined by

$$\pi_{ia}^* := \begin{cases} x_{ia}^* \Big/ \sum_a x_{ia}^* & \text{if } \sum_a x_{ia}^* > 0 \\ \text{arbitrary} & \text{if } \sum_a x_{ia}^* = 0 \end{cases} \tag{62}$$

is an optimal policy and $\sum_{i,a} r_{ia} x_{ia}^*$ is the value.

Remark: The above analysis can be found in Hordijk/Kallenberg (1984a). This approach can be extended to the total reward criterion with the restriction that only transient policies may be considered.

Infinite Horizon and Average Reward

Consider the same problem as in the discounted case, but with average rewards and costs. Let $g(R)$ and $c_k(R)$ be the average reward and average costs, respectively, $k = 1, 2, \ldots, m$, defined by

$$g(R) := \liminf_{T\to\infty} \sum_j \beta_j \cdot \left\{ \frac{1}{T} \sum_{t=1}^{T} \sum_{i,a} \mathbb{P}_R[X_t = i,\, Y_t = a | X_1 = j] \cdot r_{ia} \right\}$$

and

$$c_k^a(R) := \liminf_{T\to\infty} \sum_j \beta_j \cdot \left\{ \frac{1}{T} \sum_{t=1}^{T} \sum_{i,a} \mathbb{P}_R[X_t = i,\, Y_t = a | X_1 = j] \cdot c_{ia}^k \right\}, \quad 1 \le k \le m$$

A policy R is *feasible* if $c_k(R) \le b_k$, $1 \le k \le m$. The *value* g of the constrained problem is defined by

$$g := \sup_R \{g(R) | c_k(R) \leq b_k, k = 1, 2, \ldots, m\} \ . \tag{63}$$

We want to find an *optimal policy*, i.e. a feasible policy R^* with $g(R^*) = g$.

In Derman/Veinott (1972) an iterative algorithm, based on the Dantzig-Wolfe decomposition was proposed. Because, at that time, the linear programming polytope for the state-action frequencies was unknown a routine application of the simplex method was not possible. Below, we present this polytope and give a treatment based on the simplex method. The proofs can be found in Hordijk/Kallenberg (1984b).

For any policy R we denote the average expected state-action frequencies in the first T periods by $x_{ia}^T(R)$, i.e.

$$x_{ia}^T(R) := \sum_j \beta_j \cdot \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \right\} \tag{64}$$

Let C^* be the set of policies R for which $\lim_{T\to\infty} x_{ia}^T(R)$ exists for all $(i, a) \in E \times A$. For instance, the stationary policies belong to C^*. For $R \in C^*$ we define $x(R)$ by

$$x_{ia}(R) := \lim_{T\to\infty} x_{ia}^T(R) \ , \quad (i, a) \in E \times A \ . \tag{65}$$

It can be shown that for $R \in C^*$

$$g(R) = \sum_{i,a} r_{ia} x_{ia}(R) \quad \text{and} \quad c_k(R) = \sum_{i,a} c_{ia}^k x_{ia}(R) \ , \quad k = 1, 2, \ldots, m \ . \tag{66}$$

Let the sets L, L^*, $L(M)$, $L(S)$, $L(DS)$ and Q of vectors with components in $E \times A$, be defined by

$$L := \{x(R) | x(R) \text{ is a vector-limit point of } \{x^T(R), T = 1, 2, \ldots\}\}$$

$$L^* := \{x(R) | R \in C^*\}$$

$$L(M) := \{x(R) | R \in C^* \text{ and } R \text{ is a Markov policy}\}$$

$$L(S) := \{x(R) | R \text{ is a stationary policy}\}$$

$$L(DS) := \{x(R) | R \text{ is a deterministic and stationary policy}\}$$

$$Q := \left\{ x \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \ , & j \in E \\ \sum_a x_{ja} + \sum_{i,a} (\delta_{ij} - p_{iaj}) y_{ia} = \beta_j \ , & j \in E \\ x_{ia}, y_{ia} \geq 0 \ , & (i, a) \in E \times A \end{array} \right\} \tag{67}$$

Q is the projection (on the x-space) of the feasible solutions of the dual linear program (22).

Theorem 30: $L = L^* = L(M) = \overline{L(S)} = \overline{L(DS)} = Q$.

Hence, it is plausible to consider

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0 , & j \in E \\ \sum_{a} x_{ja} + \sum_{i,a} (\delta_{ij} - p_{iaj}) y_{ia} = \beta_j , & j \in E \\ \sum_{j,a} c_{ja}^k x_{ja} \leq b_k , & k = 1, 2, \ldots, m \\ x_{ia}, y_{ia} \geq 0 , & (i,a) \in E \times A \end{array} \right\} . \tag{68}$$

Furthermore, Theorem 30 suggests that an optimal stationary policy does not exist generally, but that we need a Markov policy from C^*. The following theorem presents this result.

Theorem 31: (i) Program (68) is infeasible if and only if there is no feasible policy. (ii) Let (x^*, y^*) be an optimal solution of (68) and let R^* be a Markov policy from C^* with $x(R^*) = x^*$. Then, R^* is an optimal policy and $\sum_{i,a} r_{ia} x_{ia}^*$ is the value.

Remarks:

1) One might wonder whether or not a stationary optimal policy exists. In Kallenberg (1983) an example can be found which implies that a stationary optimal policy does not exist, in general.
2) The determination of a Markov policy R^* such that $x(R^*) = x^*$ can be done (cf. Hordijk/Kallenberg (1984b)), but is rather complex and prohibitive.
3) It is an interesting question to find the best policy in the class of stationary policies. The natural candidate for such a stationary policy π^* seems:

$$\pi_{ia}^*(x, y) := \begin{cases} x_{ia}^* \Big/ \sum_a x_{ia}^* & a \in A(i) , \quad i \in \left\{ j \;\middle|\; \sum_a x_{ja}^* > 0 \right\} \\ y_{ia}^* \Big/ \sum_a y_{ia}^* & a \in A(i) , \quad i \in \left\{ j \;\middle|\; \sum_a x_{ja}^* = 0; \sum_a y_{ja}^* > 0 \right\} \\ \text{arbitrary} & \text{elsewhere} \end{cases} \tag{69}$$

However, in general, π^* is not the best policy in the class of the stationary policies (see Kallenberg (1983)).

(4) There are special cases in which the stationary policy defined by (69) is an optimal policy, e.g. when the Markov chain induced by $P(\pi^*)$ is unichained or when $x^*_{ia}/\sum_a x^*_{ia} = y^*_{ia}/\sum_a y^*_{ia}$ for all $a \in A(i)$ and all $i \in \{j | \sum_a x^*_{ja} > 0$ and $\sum_a y^*_{ja} > 0\}$.

(5) In Derman (1970) the unichain case is considered. He showed that in that case $L = L(M) = L(S) = \overline{L(DS)} = X$, where X is the feasible set of (27). Then, the constrained problem can be solved as follows. Let x^* be an optimal solution of the linear program

$$\max \left\{ \sum_{i,a} r_{ia} x_{ia} \;\middle|\; \begin{array}{ll} \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \;, & j \in E \\ \sum_{j,a} x_{ja} = 1 & \\ \sum_{j,a} c^k_{ja} x_{ja} \leq b_k \;, & k = 1, 2, \ldots, m \\ x_{ia} \geq 0 \;, & (i,a) \in E \times A \end{array} \right\} \tag{70}$$

$$\text{and define } \pi^* \text{ by } \pi^*_{ia} := \begin{cases} x^*_{ia} \Big/ \sum_a x^*_{ia} & \text{if } \sum_a x^*_{ia} > 0 \\ \text{arbitrary} & \text{if } \sum_a x^*_{ia} = 0 \end{cases} \tag{71}$$

Then, π^* is an optimal policy.

(6) There is a Tauberian result that gives a clear connection between the discounted state-action frequencies $x^\alpha_{ia}(R)$ and the average state-action frequencies $x_{ia}(R)$, for policies $R \in C^*$:

$$x_{ia}(R) = \lim_{\alpha \uparrow 1} (1 - \alpha) x^\alpha_{ia}(R) \text{ for all } (i,a) \in E \times A \;. \tag{72}$$

11 Multiple Objectives

We will discuss the subject of multiple objectives in the case of discounted rewards. For the average reward criterion the analysis is similar and can be found in Durinovic/Filar/Katehakis/Lee (1986). Assume that we want to maximize the total expected discounted rewards for an m-tuple of immediate rewards, say $r^1_{ia}, r^2_{ia}, \ldots, r^m_{ia}$, given an intial distribution β.

The goal in multi-objective programming is to find an efficient solution, i.e. a solution which is not dominated by another solution with respect to a number of criteria. We will discuss two concepts for a policy to be efficient.

Let $v^\alpha(k, R)$ be the total expected discounted reward with respect to the immediate rewards r_{ia}^k, $k = 1, 2, \ldots, m$. A policy R^* is *efficient* if there does not exist a policy R such that $v^\alpha(k, R) \geq v^\alpha(k, R^*)$ for all k with strict inequality for at least one k. Hence, to find an efficient policy, it is sufficient to determine a policy which maximizes $\sum_{k=1}^m \lambda_k v^\alpha(k, R)$ over the policies R for some λ with strictly positive components.

Since $v^\alpha(k, R) = \sum_{i,a} r_{ia}^k x_{ia}(R)$ and $K = \overline{K(DS)} = P$ (see section 10), it follows that a deterministic and stationary efficient policy can be determined as follows.

Theorem 32: Take any $\lambda \in \mathbb{R}^m$ with $\lambda_k > 0$, $k = 1, 2, \ldots, m$. Let x^* be an extreme optimal solution of the linear program

$$\max \left\{ \sum_{i,a} \left[\sum_{k=1}^m \lambda_k r_{ia}^k \right] x_{ia} \;\middle|\; \begin{array}{l} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj}) x_{ia} = \beta_j \;, \quad j \in E \\ x_{ia} \geq 0 \;, \quad (i, a) \in E \times A \end{array} \right\} . \tag{73}$$

Then, for each $i \in E$, $x_{ia}^* > 0$ for exactly one action, say a_i, and the deterministic and stationary policy f^* with $f^*(i) = a_i$, $i \in E$, is an efficient policy.

Next, suppose that we want to maximize lexicigraphically the functions $v^\alpha(k, R)$ for $k = 1, 2, \ldots, m$. A policy R^* which is lexicographical maximal with respect to $(v^\alpha(1, R), v^\alpha(2, R), \ldots, v^\alpha(m, R))$ is a *lexicographically efficient* policy.

To determine a lexicographically efficient policy, we compute an optimal solution, say x^1 of

$$\max \left\{ \sum_{i,a} r_{ia}^1 x_{ia} \;\middle|\; \begin{array}{l} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj}) x_{ia} = \beta_j \;, \quad j \in E \\ x_{ia} \geq 0 \;, \quad (i, a) \in E \times A \end{array} \right\} . \tag{74}$$

Then, we solve the following linear program which has one additional constraint (with x^2 as optimal solution)

$$\max \left\{ \sum_{i,a} r_{ia}^2 x_{ia} \;\middle|\; \begin{array}{l} \sum_{i,a} (\delta_{ij} - \alpha \cdot p_{iaj}) x_{ia} = \beta_j \;, \quad j \in E \\ \sum_{i,a} r_{ia}^1 x_{ia} = \sum_{i,a} r_{ia}^1 x_{ia}^1 \\ x_{ia} \geq 0 \;, \quad (i, a) \in E \times A \end{array} \right\} . \tag{75}$$

Continuing in this way we either stop when we find that an optimal solution x^k, for some $1 \leq k \leq m$, is unique or when we have solved all programs for $k = 1, 2, \ldots, m$. Let x^* be the finally obtained solution. Then a lexicographically efficient stationary, but in general not deterministic, policy π^* is obtained by the usual rule: $\pi^*_{ia} = x^*_{ia}/\sum_a x^*_{ia}$, $(i, a) \in E \times A$.

12 Mean-Variance Tradeoffs

In the first part of this paper standard utility criteria, like the average reward, are considered. It is conceivable that such a criterion is unacceptable to a risk-sensitive decision maker. Such a decision maker wants also to consider a kind of variance. There is no unique agreement how this variance should be defined. It has to capture both the risk, caused by the probabilistic nature, as the variability in the actual stream of rewards. Furthermore, it is nice to have a mathematically tractable concept, for instance a concept for which an optimal policy exists and can be computed.

Given the initial distribution β, we will define the mean $g(R)$ by

$$g(R) := \liminf_{T\to\infty} \sum_j \beta_j \cdot \left\{ \frac{1}{T} \sum_{t=1}^{T} \sum_{i,a} \mathbb{P}_R[X_t = i, Y_t = a | X_1 = j] \cdot r_{ia} \right\} \tag{76}$$

and the variance $V(R)$ by

$$V(R) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\beta,R}[r_{X_t Y_t} - g(R)]^2 \ . \tag{77}$$

It can be shown that for policies $R \in C^*$, where C^* is defined in section 10, $g(R) = \sum_{i,a} x_{ia}(R) r_{ia}$ and $V(R) = \sum_{i,a} x_{ia}(R) r^2_{ia} - [\sum_{i,a} x_{ia}(R) r_{ia}]^2$.

There are several ways to consider the mean-variance tradeoffs (e.g. Sobel (1985), Filar/Lee (1985), Kawai (1987) and Filar/Kallenberg/Lee (1989)). Sobel (1985) proposed maximizing the mean-variance ratio with constraints on the mean. In policy space this is equivalent to

$$\min \left\{ \frac{V(R)}{[g(R)]^2} \,\middle|\, L \leq g(R) \leq U \right\} . \tag{78}$$

Using the notion of state-action frequencies, (78) is related to the following nonlinear program

$$\min\left\{\frac{\sum_{i,a} r_{ia}^2 x_{ia} - \left[\sum_{i,a} r_{ia} x_{ia}\right]^2}{\left[\sum_{i,a} r_{ia} x_{ia}\right]^2} \;\middle|\; \begin{array}{l} x \in Q \\ L \le \sum_{i,a} r_{ia} x_{ia} \le U \end{array}\right\}, \tag{79}$$

where Q is defined in (67).

Kawai (1987) has considered the problem of minimizing the variance subject to bounds on the mean. Then, the problems become in policy space and as mathematical program

$$\min\{V(R) | L \le g(R) \le U\} \tag{80}$$

and

$$\min\left\{\sum_{j,a} r_{ja}^2 x_{ja} - \left[\sum_{j,a} r_{ja} x_{ja}\right]^2 \;\middle|\; \begin{array}{l} x \in Q \\ L \le \sum_{i,a} r_{ia} x_{ia} \le U \end{array}\right\} \tag{81}$$

respectively.

In Filar/Lee (1985) the problem is considered as a two-objective Markov decision problem. Filar/Kallenberg/Lee (1989) proposed to maximize the mean minus a multiple λ of the variance. Hence, these problems are

$$\max\{g(R) - \lambda \cdot V(R)\} \tag{82}$$

and

$$\max\left\{\sum_{j,a} r_{ja} x_{ja} - \lambda \cdot \left\{\sum_{j,a} r_{ja}^2 x_{ja} - \left[\sum_{j,a} r_{ja} x_{ja}\right]^2\right\} \;\middle|\; x \in Q\right\} \tag{83}$$

respectively.

The paper Huang/Kallenberg (1994) unifies and extends the approaches stated above. Furthermore, a solution method based on parametric linear programming is given which is as least as good as any known method for the particular problems. This unifying framework is given by the nonlinear program

$$\max\left\{\frac{\sum_{j,a} B_{ja} x_{ja}}{D\left(\sum_{j,a} R_{ja} x_{ja}\right)} + C\left(\sum_{j,a} R_{ja} x_{ja}\right) \;\middle|\; \begin{array}{l} x \in Q \\ L \le \sum_{i,a} r_{ia} x_{ia} \le U \end{array}\right\} \tag{84}$$

with (i) the functions $D(\cdot)$ and $C(\cdot)$ convex;
(ii) either $D(\cdot)$ is a constant
or $D(\cdot)$ is positive and nondecreasing, $C(\cdot)$ is nondecreasing and $\sum_{j,a} B_{ja} x_{ja} \geq 0$ for every $x \in Q$.

The following results are taken from Huang/Kallenberg (1994).

Theorem 32: (i) Program (84) is a convex program and it has an extreme optimal solution.
(ii) An optimal solution can be obtained by the following algorithm, which is based on the parametric linear programming problem

$$\max \left\{ \sum_{j,a} [B_{ja} + \vartheta \cdot R_{ja}] x_{ja} | x \in Q \right\} . \tag{85}$$

Algorithm 4:

1. Solve the parametric linear program (85): this yields a finite number of increasing ϑ_n, of extreme points (x^n, y^n) with (x^n, y^n) optimal on the the intervals $(\vartheta_{n-1}, \vartheta_n]$, $n = 1, 2, \ldots, m$, where $\vartheta_0 = -\infty$ and $\vartheta_m = +\infty$, and of corresponding deterministic and stationary policies f^n that are average-optimal policies on $(\vartheta_{n-1}, \vartheta_n]$ with respect to the one-period rewards $B_{ja} + \vartheta \cdot R_{ja}$, $1 \leq n \leq m$.
2. Let k be the smallest integer such that $\sum_{j,a} R_{ja} x_{ja}^{k+1} > U$ and l the smallest integer such that $\sum_{j,a} R_{ja} x_a^l < L$. Let μ and ν be such that x^U and x^L satisfy $\sum_{j,a} R_{ja} x_{ja}^U = U$ and $\sum_{j,a} R_{ja} x_{ja}^L = L$, where $x^U := \mu x^k + (1 - \mu) x^{k+1}$ and $x^L := \nu x^l + (1 - \nu) x^{l+1}$.
3. Let $V(x) := \dfrac{\sum_{j,a} B_{ja} x_{ja}}{D(\sum_{j,a} R_{ja} x_{ja})} + C(\sum_{j,a} R_{ja} x_{ja})$. Compute $V(x)$ for $x = x^L$, x^{l+1}, $x^{l+2}, \ldots, x^k$, x^U and let x^* be the x with the highest V-value. Then, x^* is an optimal solution of (84).

Remarks:

1) If $x^* = x^n$ for some $l + 1 \leq n \leq k$, then any deterministic and stationary policy f^* with $f^*(i) = a_i$, where $x_{ia_i}^* > 0$ for $i \in \{ j \in E | \sum_a x_{ja}^* > 0 \}$ and $y_{ia_i}^* > 0$ for $i \in \{ j \in E | \sum_a x_{ja}^* = 0$ and $\sum_a y_{ja}^* > 0 \}$ is an optimal policy. Hence, in the unconstrained case ($L = -\infty$; $U = +\infty$) an extreme optimal solution and a deterministic and stationary optimal policy are obtained.
2) If $x^* = x^U$ (or x^L), then an optimal policy is an initial randomization of the two deterministic policies corresponding to x^k and x^{k+1} (or x^l and x^{l+1}).

3) We mention that the above approach can also be applied to find optimal policies for discounted mean-variance tradeoff problems.
4) The optimal policy, say R^*, is also Pareto-optimal with respect to the pair $(g(R), -V(R))$, i.e. there does not exist a policy $R \neq R^*$ such that $g(R) \geq g(R^*)$ and $-V(R) \geq -V(R^*)$.

Acknowledgement: I am grateful to Arie Hordijk for introducing me to the subject of linear programming for *MDP*'s and for the many stimulating discussions, especially during the period when he was my thesis advisor.

References

Beutler FJ, Ross KW (1985) Optimal policies for controlled Markov chains with a constraint. Journal of Mathematical Analysis and Applications 112:236–252

Beutler FJ, Ross KW (1986) Time-average optimal constrained semi-Markov decision processes. Advances in Applied Probability 18:341–359

Bewley T, Kohlberg E (1978) On stochastic games with stationary optimal stategies. Mathematics of Operations Research 3:104–125

Blackwell D (1962) Discrete dynamic programming. Annals of Mathematical Statistics 33:719–726

Dantzig GB (1963) Linear programming and extensions. Princeton University Press, Princeton

De Ghellinck GT (1960) Les problèmes de décisions séquentielles. Cahiers du Centre de Recherche Opérationelle 2:161–179

De Ghellinck GT, Eppen GD (1967) Linear programming solutions for separable Markovian decision problems. Management Science 13:371–394

Denardo EV (1967) Contraction mapping in the theory underlying dynamic programming. SIAM Review 9:165–177

Denardo EV (1970a) On linear programming in a Markov decision problem. Management Science 16:281–288

Denardo EV (1970b) Computing a bias-optimal policy in a discrete-time Markov decision problem. Operations Research 18:279–289

Denardo EV (1973) A Markov decision problem. In: Hu TC, Robinson SM (eds) Mathematical Programming, Academic Press 33–68

Denardo EV, Fox BL (1968) Multichain Markov renewal programs. SIAM Journal on Applied Mathematics 16:468–487

Denardo EV, Miller BL (1968) An optimality condition for discrete dynamic programming with no discounting. Annals of Mathematical Statistics 39:1220–1227

Denardo EV, Rothblum UG (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 Opérationelle 14:3–16

D'Epenoux F (1963) A probabilistic production and inventory problem. Management Science 10:98–108

Derman C (1970) Finite state Markovian decision processes. Academic Press, New York

Derman C, Klein M (1965) Some remarks on finite horizon Markovian decision models. Operations Research 13:272–278

Derman C, Strauch R (1966) A note on memoryless rules for controlling sequential control problems. Annals of Mathematical Statistics 37:276–278

Derman C, Veinott AF Jr (1972) Constrained Markov decision chains. Management Science 19:389–390

Dirickx YMI, Rao MR (1979) Linear programming methods for computing gain-optimal policies in Markov decision models. Cahiers du Centre Etudes Recherche Opérationelle 21:133–142

Durinovic S, Filar JA, Katehakis MN, Lee HM (1986) Multi-objective Markov decision process with average reward criterion. Large Scale System 10:215–226

Federgruen A, Schweitzer PJ (1980) A survey of asymptotic value-iteration for undiscounted Markovian decision processes 73–110. In: Hartley R, Thomas LC, White DJ (eds) Recent developments in Markov decision processes, Academic Press

Filar JA (1980) Algorithms for solving some undiscounted stochastic games. PhD Thesis, University of Illinois at Chicago

Filar JA, Kallenberg LCM, Lee HM (1989) Variance-penalized Markov decision Processes. Mathematics of Operations Research 14:147–161.

Filar JA, Lee HM (1985) Gain/Variability tradeoffs in undiscounted Markov decision Processes. Proceedings of 24th Conference on Decision and Control IEEE 1106–1112

Filar JA, Schultz TA (1988) Communicating MDP's: equivalence and LP properties. OR Letters 7:303–307

Gillette D (1957) Stochastic games with zero stop probabilities. In: Dresher M, Tucker AW, Wolfe P (eds) Contributions to the theory of games, Volume 3, Annals of Mathematical Studies 39, Princeton University Press 179–189

Heilmann W-R (1977) Lineare Programmierung stochastischer dynamischer Entscheidungsmodelle. Thesis, Universität Hamburg

Heilmann W-R (1978) Solving stochastic dynamic programming by linear programming – An annotated bibliography. Zeitschrift für Operations Research 22:43–53

Hordijk A (1978) From linear to dynamic programming via shortest paths. In: Baayen, PC et al. (eds) Proceedings of the Bicentennial Congress of the Wiskundig Genootschap, Mathematical Centre, Amsterdam

Hordijk A, Dekker R, Kallenberg LCM (1985) Sensitivity analysis in discounted Markov decision problems. OR Spektrum 7:143–151

Hordijk A, Kallenberg LCM (1979) Linear programming and Markov decision chains. Management Science 25:352–362

Hordijk A, Kallenberg LCM (1980) Linear programming methods for solving finite Markovian decision problems. In: Fandel G et al. (eds) Operations Research Proceedings 1980 468–482

Hordijk A, Kallenberg LCM (1981a) Linear programming and Markov games I. In: Moeschlin O, Pallaschke D (eds) Game Theory and Mathematical Economics. North Holland, Amsterdam 291–305

Hordijk A, Kallenberg LCM (1981b) Linear programming and Markov games II. In: Moeschlin O, Pallaschke D (eds) Game Theory and Mathematical Economics. North Holland, Amsterdam 307–320

Hordijk A, Kallenberg LCM (1984a) Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints. Mathematical Programming 30:46–70

Hordijk A, Kallenberg LCM (1984b) Constrained undiscounted dynamic programming. Mathematics of Operations Research 9:276–289

Howard RA (1960) Dynamic Programming and Markov Processes. Wiley, New York

Huang Y, Kallenberg LCM (1994) On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs. Mathematics of Operations Research 19

Kallenberg LCM (1981a) Unconstrained and constrained dynamic programming over a finite horizon. Report, University of Leiden, The Netherlands

Kallenberg LCM (1981b) Linear programming to compute a bias-optimal policy. In: B. Fleischmenn et al. (eds) Operations Research Proceedings 433–440

Kallenberg LCM (1983) Linear programming and finite Markovian control problems. Mathematical Centre Tracts #148, Amsterdam

Kallenberg LCM (1994) Efficient algorithms to determine the classification of a Markov decision problem. Report, University of Leiden, The Netherlands

Kawai H (1987) A variance minimization problem for a Markov decision process. European Journal of operational Research 31:140–145

Manne AS (1960) Linear programming and sequential decisions. Management Science 6:259–267

Mertens JF, Neyman A (1981) Stochastic games. International Journal of Game Theory 10:53–56

Miller BL, Veinott AF Jr (1969) Discrete dynamic programming with a small interest rate. Annals of Mathematical Statistics 40:366–370

Mine H, Osaki S (1970) Markovian decision processes. Elsevier, New York

Nazareth JL, Kulkarni RB (1986) Linear programming formulations of Markov decision processes. OR Letters 5:13–16

Osaki S, Mine H (1969) Linear programming considerations on Markovian decision processes with no discounting. Journal of Mathematical Analysis and Applications 26:221–232

Parthasarathy T, Raghavan TES (1981) An ordered field property for stochastic games when one player controls transition probabilities. Journal of Optimization Theory and Applications 33:375–392

Porteus EL (1980) Overview of iterative methods for discounted finite Markov and semi-Markov decision chains. In: Hartley R, Thomas LC, White DJ (eds) Recent developments in Markov decision processes, Academic Press 1–20

Puterman ML (1988) Markov decision processes: a survey. In: Heyman DP, Sobel MJ (eds) Handbook of Operations Research and Management Science, Volume 2: Stochastic Models. North Holland, Amsterdam 331–434

Puterman ML, Brumelle SL (1979) On the convergence of policy iteration in stationary dynamic programming. Mathematics of Operations Research 4:60–69

Raghavan TES, Filar JA (1991) Algorithms for stochastic games – A survey. Zeitschrift für Operations Research 35:437–472

Ross SM (1983) Introduction to stochastic programming. Academic Press

Rothblum UG (1979) Solving stopping stochastic games by maximizing a linear function subject to quadratic constraints. In: Moeschlin O, Pallaschke D (eds) Game Theory and Mathematical Economics. North Holland, Amsterdam 103–105

Shapley LS (1953) Stochastic games. Proceedings National Academy of Sciences USA 39:1095–1100

Sobel MJ (1985) Maximal mean/standard deviation ratio in an undiscounted MDP. OR Letters 4:157–159

Stein J (1988) On efficiency of linear programming applied to discounted Markovian decision problems. OR Spektrum 10:153–160

Sutherland WRS (1980) Optimality in transient Markov chains and linear programming. Mathematical Programming 18:1–6

Van Nunen, JAEE (1976) Contracting Markov decision processes. Mathematical Centre Tract #71, Amsterdam

Veinott AF Jr (1966) On finding optimal policies in discrete dynamic programming with no discounting. Annals of Mathematical Statistics 37:1284–1294

Veinott AF Jr (1969) Discrete dynamic programming with sensitive discount optimality criteria. Annals of Mathematical Statistics 40:1635–1660

Veinott AF Jr (1974) Markov decision chains. In: GB Dantzig, BC Eaves (eds) Studies in Mathematics, Volume 10: Studies in Optimization. The Mathematical Association of America 124–159

Vrieze OJ (1981) Linear programming and undiscounted stochastic games in which one player controls the transitions. OR Spektrum 3:29–35

Wessels J, Van Nunen JAEE (1975) Discounted semi-Markov decision processes: linear programming and policy iteration. Statistica Neerlandica 29:1–7
White, DJ (1988) Mean, variance and probabilistic criteria in finite Markov decision processes: a review. Journal of Optimization Theory and Applications 56:1–30

Received: November 1992
Revised version received: January 1994