Finite state and action MDPs

Chapter · August 2002

DOI: 10.1007/978-1-4615-0805-2_2

1 author:

Lodewijk C. M. Kallenberg
Leiden University
79 PUBLICATIONS **862** CITATIONS

SEE PROFILE

2 FINITE STATE AND ACTION MDPS

Lodewijk Kallenberg

Abstract: In this chapter we study Markov decision processes (MDPs) with finite state and action spaces. This is the classical theory developed since the end of the fifties. We consider finite and infinite horizon models. For the finite horizon model the utility function of the total expected reward is commonly used. For the infinite horizon the utility function is less obvious. We consider several criteria: total discounted expected reward, average expected reward and more sensitive optimality criteria including the Blackwell optimality criterion. We end with a variety of other subjects.
The emphasis is on computational methods to compute optimal policies for these criteria. These methods are based on concepts like value iteration, policy iteration and linear programming. This survey covers about three hundred papers. Although the subject of finite state and action MDPs is classical, there are still open problems. We also mention some of them.

2.1 INTRODUCTION

2.1.1 *Origin*

Bellman's book [13], can be considered as the starting point of Markov decision processes (MDPs). However, already in 1953, Shapley's paper [221] on stochastic games includes as a special case the value iteration method for MDPs, but this was recognized only later on. About 1960 the basics for the other computational methods (policy iteration and linear programming) were developed in publications like Howard [121], De Ghellinck [42], D'Epenoux [55], Manne [164] and Blackwell [27]. Since the early sixties, many results on MDPs are published in numerous journals, monographs, books and proceedings. Thousands of papers were published in scientific journals. There are about fifty books on MDPs. Around 1970 a first series of books was published. These books (e.g. Derman [58], Hinderer [107], Kushner [148], Mine and Osaki [167] and Ross [198]) contain the fundamentals of the theory of finite MDPs. Since that time nearly

every year one or more MDP-books appeared. These books cover special topics (e.g. Van Nunen [250], Van der Wal [246], Kallenberg [134], Federgruen [69], Vrieze [260], Hernández-Lerma [102], Altman [2] and Sennott [218]) or they deal with the basic and advanced theory of MDPs (e.g. Bertsekas [15], Whittle [289], [290], Ross [200], Dietz and Nollau [63], Bertekas [17], Denardo [50], Heyman and Sobel [106], White [285], Puterman [186], Bertsekas [18], [19], Hernández-Lerma and Lasserre [103], [104], and Filar and Vrieze [79].

2.1.2 *The model*

We will restrict ourselves to discrete, finite Markovian decision problems, i.e. the *state space* $\mathbb{X}$ and the *action sets* $\mathbb{A}(i), i \in \mathbb{X}$, are finite, and the decision time points t are equidistant, say $t = 1, 2, \ldots$. If, at time point t, the system is in state i and action $a \in \mathbb{A}(i)$ is chosen, then the following happens independently of the history of the process:
(1) a *reward* $r(i,a)$ is earned immediately;
(2) the process moves to state $j \in \mathbb{X}$ with *transition probability* $p(j|i,a)$, where $p(j|i,a)) \geq 0$ and $\sum_j p(j|i,a) = 1$ for all i, j and a.

The objective is to determine a policy, i.e. a rule at each decision time point, which optimizes the performance of the system. This performance is expressed as a certain *utility function*. Such utility function may be the expected total (discounted) reward over the planning horizon or the average expected reward per unit time. The decision maker has to find the optimal balance between immediate reward and future reward: a high immediate reward may bring the process in a bad situation for later rewards.

In Chapter 1 several classes of *policies* are introduced: general policies, Markov policies and stationary policies. There are randomized and nonrandomized (pure) policies. Denote the set of pure stationary policies by F and a particular policy of that set by f. Let $\mathbb{X} \times \mathbb{A} = \{(i,a) \mid i \in \mathbb{X},\ a \in \mathbb{A}(i)\}$, let the random variables X_t and Y_t denote the state and action at time t and let $\mathbb{P}_{\beta,\pi}[X_t = j, Y_t = a]$ be the notation for the probability that at time t the state is j and the action is a, given that policy π is used and β is the initial distribution. The next theorem shows that for any initial distribution β, any sequence of policies $\pi_1, \pi_2, \ldots$ and any convex combination of the marginal distributions of $\mathbb{P}_{\beta,\pi_k}, k \in \mathbb{N}$, there exists a Markov policy with the same marginal distribution.

Theorem 2.1 *Given any initial distribution β, any sequence of policies π_1, $\pi_2, \ldots$ and any sequence of nonnegative real numbers $p_1, p_2 \ldots$ with $\sum_k p_k = 1$, there exists a Markov policy π_* such that for every $(j,a) \in \mathbb{X} \times \mathbb{A}$*

$$\mathbb{P}_{\beta,\pi_*}[X_t = j, Y_t = a] = \sum\nolimits_k p_k \cdot \mathbb{P}_{\beta,\pi_k}[X_t = j, Y_t = a], \quad t \in \mathbb{N}. \tag{1.1}$$

Corollary 2.1 *For any starting state i and any policy π, there exists a Markov policy π_* such that*

$$\mathbb{P}_{i,\pi_*}[X_t = j, Y_t = a] = \mathbb{P}_{i,\pi}[X_t = j, Y_t = a], \quad t \in \mathbb{N}, (j,a) \in \mathbb{X} \times \mathbb{A}. \tag{1.2}$$

The results of Theorem 2.1 and Corollary 2.1 imply the sufficiency of Markov policies for performance measures which only depend on the marginal distributions. Corollary 2.1 is due to Derman and Strauch [61] and the extension to Theorem 2.1 was given by Strauch and Veinott [237]. The result is further generalized to more general state and actions spaces by Hordijk [112] and Van Hee [247].

2.1.3 *Optimality criteria*

Let $v(i,\pi)$ be the utility function if policy π is used and state i is the starting state, $i \in \mathbb{X}$. The *value vector* v of this utility function is defined by

$$v(i) := \sup_{\pi} v(i,\pi), \quad i \in \mathbb{X}. \tag{1.3}$$

A policy π is an *optimal policy* if $v(i,\pi) = v(i), i \in \mathbb{X}$. In Markov decision theory the existence and the computation of optimal policies is studied. For this purpose a so-called *optimality equation* is derived, i.e. a functional equation for the value vector. Then a solution of this equation is constructed which produces both the value vector and an optimal policy. There are three standard methods to perform this: value iteration, policy iteration and linear programming.

In *value iteration* the optimality equation is solved by successive approximation. Starting with some v^0, v^{t+1} is computed from $v^t, t = 0, 1, \ldots$. The sequence $v^0, v^1, \ldots$ converges to the solution of the optimality equation. In *policy iteration* a sequence of improving policies $f_0, f_1, \ldots$ is determined, i.e. $v(f_{t+1}) \geq v(f_t)$ for all t, until an optimal policy is reached. The *linear programming* method can be used because the value vector is the smallest solution of a set of linear inequalities; an optimal policy can be obtained from its dual program.

In this survey we consider the following utility functions:
(1) total expected reward over a finite horizon;
(2) total expected discounted reward over an infinite horizon;
(3) average expected reward over an infinite horizon;
(4) more sensitive optimality criteria for the infinite horizon.

Suppose that the system has to be controlled over a finite planning horizon of T periods. As performance measure we use the *total expected reward* over the planning horizon, i.e. for policy π we will consider for starting state i

$$v^T(i,\pi) := \sum_{t=1}^{T} \mathbf{E}_{i,\pi}[r(X_t, Y_t)] = \sum_{t=1}^{T} \sum_{j,a} \mathbf{P}_{i,\pi}[X_t = j, Y_t = a] \cdot r(j,a). \tag{1.4}$$

A matrix $P = (p_{ij})$ is called a *transition matrix* if $p_{ij} \geq 0$ for all (i,j) and $\sum_j p_{ij} = 1$ for all i. Markov policies, and consequently also stationary policies, induce transition matrices. For the randomized Markov policy $\pi = (\pi^1, \pi^2, \ldots)$ we define, for every $t \in \mathbb{N}$, the transition matrix $P(\pi^t)$ by

$$[P(\pi^t)]_{ij} := \sum_a p(j|i,a)\pi^t(i,a) \text{ for all } i,j \in \mathbb{X}, \tag{1.5}$$

and the reward vector $r(\pi^t)$ by

$$r_i(\pi^t) := \sum\nolimits_a \pi^t(i,a) r(i,a) \text{ for all } i \in \mathbb{X} \tag{1.6}$$

Hence the total expected reward for the Markov policy π can be written in vector notation as

$$v^T(R) = \sum\nolimits_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t). \tag{1.7}$$

It can be shown that an optimal Markov policy $\pi_* = (f_*^1, f_*^2, \ldots, f_*^T)$ exists, where f_*^t is a pure decision rule $1 \leq t \leq T$. The nonstationarity is due to the finiteness of the planning horizon.

Next, we consider an infinite planning horizon. In that case there is no unique optimality criterion. Different optimality criteria are meaningful: discounted reward, total reward, average reward or more sensitive criteria.

The *total expected α-discounted reward*, given *discount factor* $\alpha \in [0,1)$, initial state i and policy π, is denoted by $v^\alpha(i,\pi)$ and defined by

$$\begin{aligned} v^\alpha(i,\pi) &:= \sum\nolimits_{t=1}^{\infty} \mathbf{E}_{i,\pi}[\alpha^{t-1} r(X_t,(Y_t)] \\ &= \sum\nolimits_{t-1}^{\infty} \alpha^{t-1} \sum\nolimits_{j,a} \mathbb{P}_{i,\pi}[X_t = j, Y_t = a] r(j,a). \end{aligned} \tag{1.8}$$

In section 1.3.1 it will be shown that there exists an optimal policy $f \in F$ and that any stationary policy π satisfies

$$v^\alpha(\pi) = \sum\nolimits_{t-1}^{\infty} \alpha^{t-1} P(\pi)^{t-1} r(\pi) = [I - \alpha P(\pi)]^{-1} r(\pi). \tag{1.9}$$

When there is no discounting, i.e. the discount factor α equals 1, then - for instance - we may consider the total expected reward and the average expected reward criterion. In the total expected reward criterion the utility function is $\lim_{T\to\infty} \sum_{t=1}^{T} \mathbf{E}[r(X_t, Y_t)]$. Without further assumptions, this limit can be infinite or the limsup can be unequal to the liminf. When the average reward criterion is used, the limiting behavior of the expectation of $\frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)$ is considered. Since $\lim_{T\to\infty} \frac{1}{T}\sum_{t-1}^{T} \mathbf{E}[r(X_t, Y_t)]$ or $\mathbf{E}[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)]$ does not exist, in general, and interchanging limit and expectation may not be allowed, there are four different evaluation measures, which can be considered for a given policy:
(a) the lower limit of the average expected reward:

$$\phi(i,\pi) := \liminf_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathbf{E}_{i,\pi}[r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.10}$$

(b) the upper limit of the average expected reward:

$$\Phi(i,\pi) := \limsup_{T\to\infty} \frac{1}{T} \sum\nolimits_{t=1}^{T} \mathbf{E}_{i,\pi}[r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.11}$$

(c) the expectation of the lower limit of the average reward:

$$\psi(i,\pi) := \mathbf{E}_{i,\pi}[\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)], i \in \mathbb{X}; \tag{1.12}$$

(d) the expectation of the upper limit of the average reward:

$$\Psi(i,\pi) := \mathbf{E}_{i,\pi}[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(X_t, Y_t)], i \in \mathbb{X}. \tag{1.13}$$

Lemma 2.1
(i) $\psi(\pi) \leq \phi(\pi) \leq \Phi(\pi) \leq \Psi(\pi)$ *for every policy* π*;*
(ii) $\psi(\pi) = \phi(\pi) = \Phi(\pi) = \Psi(\pi)$ *for every stationary policy* π.

Remark
In Bierth [26] it is shown that the four criteria are equivalent in the sense that the value vectors can be attained for one and the same deterministic policy. Examples can be constructed in which for some policy π the inequalities of Lemma 2.1 part (i) are strict.

The long-run average reward criterion has the disadvantage that it does not consider rewards earned in a finite number of periods. Hence, there may be a preference for more selective criteria. There are several ways to be more selective. One way is to consider discounting for discount factors that tend to 1. Another way is to use more subtle kinds of averaging. We will present some criteria and results. For all criteria it can be shown that optimal policies in class F exist and that these policies are (at least) average optimal.

A policy π_* is called *n-discount optimal* for some integer $n \geq -1$, if $\liminf_{\alpha\uparrow 1}(1-\alpha)^{-n}[v^\alpha(\pi_*) - v^\alpha(\pi)] \geq 0$ for all policies π. 0-discount optimality is also called *bias-optimality*. There is also the concept of *n-average optimality*. For any policy π, any $t \in \mathbb{N}$ and for $n = -1, 0, 1, \ldots$, let the vector $v^{n,t}(\pi)$ be defined by

$$v^{n,t}(\pi) := \begin{cases} v^t(\pi) & for \quad n = -1 \\ \sum_{s=1}^{t} v^{n-1,s}(\pi) & for \quad n = 0, 1 \ldots \end{cases} \tag{1.14}$$

π_* is said to be n-average optimal if $\liminf_{T\to\infty} \frac{1}{T}[v^{n,T}(\pi_*) - v^{n,T}(\pi)] \geq 0$ for all policies π.

A policy π_* is said to be *Blackwell optimal* if π_* is α-discounted optimal for all discount factors $\alpha \in [\alpha_0, 1)$ for some $0 \leq \alpha_0 < 1$. In a fundamental paper Blackwell [27] presented a mathematically rigorous proof for the policy iteration method to compute an α-discounted optimal policy. He also introduced the concept of bias-optimality (Blackwell called it *nearly optimality)* and established the existence of a discounted optimal policy for all discount factors sufficiently close to 1. In honor of Blackwell, such policy is called a Blackwell optimal policy.
It can be shown that n-discount optimality is equivalent to n-average optimality, that (-1)-discount optimality is equivalent to average optimality, and

that Blackwell optimality is n-discount optimality for all $n \geq N-1$, where $N = \#\mathbb{X}$ (in this chapter we will always use the notation N for the number of states).
The n-discount optimality criterion and the policy iteration method for finding an n-discount optimal policy, were proposed by Veinott [257]. He also showed that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. Sladky [223] has introduced the concept of n-average optimality; furthermore, he also showed the equivalence between this criterion and n-discount optimality. More details on bias optimality and Blackwell optimality can be found in Chapter 3 and Chapter 8.

2.1.4 Applications

White has published three papers on 'real applications' of Markov decision theory (White [280], [281] and [284]). Many stochastic optimization problems can be formulated as MDPs. In this section we shortly introduce the following examples: routing problems, stopping and target problems, replacement problems, maintenance and repair problems, inventory problems, the optimal control of queues, stochastic scheduling and multi-armed bandit problems. In this book there are also chapters on applications in finance (Chapter 15) and in telecommunication (Chapter 16). We also mention the contribution Chapter 17 on water reservoir applications.

Routing problems
In routing problems the problem is to find an optimal route through a network. Well known is the shortest path problem. A shortest path problem in a layered network can be formulated as an MDP over a finite horizon. Another application of this kind is the maximum reliability problem. In this network the connections are unreliable: let p_{ij} be the probability of reaching node j when the arc from node i to node j is chosen. The objective is to maximize the probability of reaching a terminal node n when the process is started in some node, say node 1. Results for a stochastic version of the shortest path problem can for instance be found in Bertsekas and Tsitsiklis [23]. The maximum reliability problem is discussed in Roosta [194].

Optimal stopping problems
In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If we continue in state i, a cost c_i is incurred and the probability of being in state j at the next time point is p_{ij}. If the stopping action is chosen in state i, then a final reward r_i is earned and the process terminates. In an optimal stopping problem, in each state one has to determine which action is chosen with respect to the total expected reward criterion. This kind of problem often has an optimal policy that is a so-called *control limit policy*.
The original analysis of optimal stopping problems appeared in Derman and Sacks [60], and Chow and Robbins [36]. A dynamic programming approach can be found in Breiman [28] who showed the optimality of a control limit policy.

Target problems
In a target problem one wants to reach a distinguished state (or a set of states) in some optimal way, where in this context optimal means, for instance, at minimum cost or with maximum probability. The target states are absorbing, i.e. there are no transitions to other states and the process can be assumed to terminate in the target states. These target problems can be modeled as MDPs with the total expected reward as optimality criterion. To the class of target problems we may count the so-called *first passage problem*. In this problem there is one target state and the objective is to reach this state (for the first time) at minimum cost. A second class of target problems are *gambling problems* (the gambler's goal is to reach a certain fortune N and the problem is to determine a policy which maximizes the probability to reach this goal). For more information about MDPs and gambling problems we refer to Chapter 13. The first passage problem was introduced by Eaton and Zadeh [67] under the name "pursuit problem". The dynamic programming approach was introduced in Derman [56]. A standard reference on gambling is Dubins and Savage [64]. Dynamic programming approaches are given in Ross [199] and Dynkin [66].

Replacement problems
Consider an item which is in a certain state. The state of the item describes its condition. Suppose that in each period, given the state of the item, the decision has to be made whether or not to replace the item by a new one. When an item of state i is replaced by a new one, the old item is sold at price s_i, a new item is bought at price c, and the transition to the new state is instantaneous. In case of nonreplacement, let p_{ij} be the probability that an item of state i is at the beginning of the next period in state j, and suppose that c_i is the maintenance cost—during one period for an item of state i. This problem can be modeled as an MDP. It turns out that for the computation of an optimal policy an efficient algorithm, with complexity $\mathcal{O}(N^3)$, exists (see Gal [83]).
Next, we mention the model of deterioration with failure. In this model the states are interpreted as 'ages'. In state i there is a failure probability p_i and, when failure occurs, there is an extra cost f_i and the item has to be replaced by a new one. If there is no failure the next state is state $i+1$. It can be shown that, under natural assumptions about the failure probabilities and the costs, a control limit policy is optimal, i.e. there is an age i_* and the item is replaced by a new one if its age exceeds i_*. This property holds for the discounted reward criterion as well as for the average reward criterion.
There are a lot of references on replacement models. The early survey of Sherif and Smith [222] contained already over 500 references. Results on the optimality of control limit policies for replacement problems can be found in Derman [57, 58], Kolesar [146], Ross [198] and Kao [138].

Maintenance and repair problems
In maintenance and repair problems there is a system which is subject to deterioration and failure. Usually, the state is a characterization of the condition of the system. When the state is observed, an action has to be chosen, e.g. to keep the system unchanged, to execute some maintenance or repair, or to replace one or more components by new ones. Each action has corresponding

costs. The objective is to minimize the total costs, the discounted costs or the average costs. These problems can easily be modeled as an MDP.
A one-component problem is described in Klein [145]. The two-component maintenance problem was introduced by Vergin and Scriabin [259]. Other contributions in this area are e.g. Oezekici [173], and Van der Duyn Schouten and Vanneste [244]. An n-component series system is discussed in Katehakis and Derman [139]. Asymptotic results for highly reliable systems can be found in Smith [226], Katehakis and Derman [140], and Frostig [81].

Inventory problems
In inventory problems an optimal balance between inventory costs and ordering costs has to be determined. We assume that the probability distribution of the demand is known. There are different variants of the inventory problem. They differ, for instance, in the following aspects:
- stationary or nonstationary costs and demands;
- a finite planning horizon or an infinite planning horizon;
- backlogging or no backlogging.
For all these variants different performance measures may be considered.
In many inventory models the optimal policy is of (s, S)-type, i.e. when the inventory is smaller than or equal to s, then replenish the stock to level S. The existence of optimal (s, S)-policies in finite horizon models with fixed cost K is based on the so-called K-convexity, introduced by Scarf [202]. The existence of an optimal (s, S)-policy in the infinite horizon model is shown by Iglehart [126]. Another related paper is Veinott [255]. For the relation between discounted and average costs we refer to Hordijk and Tijms [119]. For the computation of the values s and S we refer to papers like Federgruen and Zipkin [76], and Zheng and Federgruen [292].

Optimal control of queues
Consider a queueing system where customers arrive according to a Poisson process and where the service time of a customer is exponentially distributed. Suppose that the arrival and service rates can be controlled by a finite number of actions. When the system is in state i, i.e. there are i customers in the system, action a means that the arrival or the service rates are $\lambda_i(a)$ or $\mu_i(a)$, respectively. The arrival and service processes are continuous-time processes. However, by the memoryless property of the exponential distribution, we can find an embedded discrete-time Markov chain which is appropriate for our analysis. This technique is called uniformization (see e.g. Tijms [241]).
A queue, or a network of queues, is a useful model for many applications, e.g. manufacturing, computer, telecommunication and traffic systems. See the survey of MDPs in telecommunication, Chapter 16. Control models can optimize certain performance measures by varying the control parameters of the system. We distinguish between *admission control* and *service rate control*.
In a service rate model, the service rate can be chosen from an interval $[0, \bar{\mu}]$. If rate μ is chosen, there are service costs $c(\mu)$ per period; we also assume that there are holding costs $h(i)$ per period when there are i customers in the system. Under natural conditions it can be shown that a *bang-bang policy* is optimal, i.e. $\mu = 0$ or $\mu = \bar{\mu}$. For details see Weber and Stidham [268]. Surveys of optimal

control of (networks of) queues can be found in the book by Walrand [265] and the papers by Stidham [234] and Stidham and Weber [235].

Stochastic scheduling
In a scheduling problem, jobs are processed on machines. Each machine can process only one job at a time. A job has a given processing time on the machines. In stochastic scheduling, these processing times are random variables. At certain time points decisions have to be made, e.g. which job is assigned to which machine. There are two types of models: the *customer assignment* models, in which each arriving customer has to be assigned to one of the queues (each queue with its own server) and *server assignment* models, where the server has to be assigned to one of the queues (each queue has its own customers).
Also in queueing models optimal policies often have a nice structure. Examples of this structure are:
- *μc-rule* : this rule assigns the server to queue k, with k the queue with $\mu_k c_k = \max_i\{\mu_i c_i \mid \text{queue } i \text{ is nonempty}\}$, where c_i is the cost which is charged per unit of time that the customer is in queue i and the service times in queue i are geometrically distributed with rate μ_i;
- *shortest queue policy (SQP)*: an arriving customer is assigned to the shortest queue;
- *longest expected processing time (LEPT)*: the jobs are allocated to the machines in decreasing order of their expected processing times;
- *shortest expected processing time (SEPT)*: the jobs are allocated to the machines in increasing order of their expected processing times.
The optimality of the μc-rule is established in Baras, Ma and Makowsky [9]. Ephremides, Varayia and Walrand [68] have shown the optimality of the shortest queue policy. The results for the optimality of the LEPT and SEPT policies are due to Bruno, Downey and Frederickson [30]. Related results are obtained by Weber [266] and by Chang, Hordijk, Righter and Weiss [33]. For reviews on stochastic scheduling we refer to Weiss [269], Walrand [265] (chapter 8) and Righter [193].

Multi-armed bandit problem
The multi-armed bandit problem is a model for dynamic allocation of a resource to one of n independent alternative projects. Any project may be in one of a finite number of states. At each period the decision maker has the option of working on exactly one of the projects. When a project is chosen, the immediate reward and the transition probabilities only depend on the active project and the states of the remaining projects are frozen. Applications of this model appear in machine scheduling, in the control of queueing systems and in the selection of decision trials in medicine. It can be shown that an optimal policy is the policy that selects the project which has the largest so-called *Gittins-index*. Fortunately, these indices can be computed for each project separately. As a consequence, the multi-armed bandit problem can be solved by a sequence of n one-armed bandit problems. This is a decomposition result by which the dimensionality of the problem is reduced considerably. Efficient algorithms for the computation of the Gittins indices exist. The most fundamental contribution on multi-armed bandit problems was made by Gittins (cf. Gittins and

Jones [86], and Gittins [85]). In Whittle [288] an elegant proof is presented. Other proofs are given by Ross [200], Varaiya, Walrand and Buyoccoc [254], Weber [267] and Tsitsiklis [243]. Several methods are developed for the computation of the Gittins indices: Varaiya, Walrand and Buyukkoc [254], Chen and Katehakis [35], Kallenberg [135], Katehakis and Veinott [141], Ben-Israel and S.D.Flåm [14], and Liu and Liu [155].

2.2 FINITE HORIZON

Consider an MDP with a finite horizon of T periods. In fact, we can analyze with the same effort a nonstationary MDP, i.e. with rewards and transition probabilities which may depend on the time t ($1 \leq t \leq T$). These nonstationary rewards and transition probabilities are denotated by $r^t(i,a)$ and $p^t(j|i,a)$. By the *principle of optimality*, an optimal policy can be determined by *backward induction* as the next theorem shows. The proof can be given by induction on the length T of the horizon. The use of the principle of optimality and the technique of dynamic programming for sequential optimization was provided by Bellman [13].

Theorem 2.2 *Let $x_i^{T+1} = 0, i \in \mathbb{X}$. Determine for $t = T, T-1, ..., 1$ a pure decision rule f^t such that*

$$[r^t(f^t)]_i + [P(f^t)x^{t+1}]_i = \max_{a \in A(i)} \{r^t(i,a) + \sum\nolimits_j p^t(j|i,a) \cdot x_j^{t+1}\}, i \in \mathbb{X},$$

and let $x^t = r^t(f^t) + P^t(f^t)x^{t+1}$. Then, $R_* = (f^1, f^2, ..., f^T)$ is an optimal policy and x^1 is the value vector.

If $[r^t(f^t)]_i + [P^t(f^t)x^{t+1}]_i = \max_{a \in A(i)} \{r^t(i,a) + \sum_j p^t(j|i,a) \cdot x_j^{t+1}\}, i \in \mathbb{X}$, then we denote $r^t(f^t) + P^t(f^t)x = \max_{\mathbb{X} \times \mathbb{A}} \{r^t + P^t x\}$ and $f^t \in \underset{\mathbb{X} \times \mathbb{A}}{\operatorname{argmax}} \{r^t + P^t x\}$.

Algorithm I (finite horizon)

1. $x := 0$.
2. Determine for $t = T, T-1, \ldots, 1$:
   $f^t \in \underset{\mathbb{X} \times \mathbb{A}}{\operatorname{argmax}} \{r^t + P^t x\}$ and $x := r^t(f^t) + P^t(f^t)x$.
3. $R_* := (f^1, f^2, \ldots, f^T)$ is an optimal policy and x is the value vector.

Remarks

1. It is also possible to include in this algorithm *elimination of suboptimal actions* . Suboptimal actions are actions that will not occur in an optimal policy. References are Hastings and Van Nunen [99] and Hübner [124].
2. A finite horizon nonstationary MDP can be transformed in an equivalent stationary infinite horizon model. In such an infinite horizon model other options, as the treatment of *side constraints*, also called *additional constraints*, are applicable. These results can be found in Derman and Klein [59] and in Kallenberg [131], [132].

2.3 DISCOUNTED REWARD CRITERION

2.3.1 *Introduction*

In order to find an optimal policy and the value vector v^α, the so-called optimality equation

$$v_i^\alpha = \max_a \{r(i,a) + \alpha \sum\nolimits_j p(j|i,a) v_j^\alpha\}, i \in \mathbb{X} \tag{3.1}$$

plays a central role. Consider the mapping $T : \mathbb{R}^N \to \mathbb{R}^N$, defined by

$$[Tz]_i := \max_a \{r(i,a) + \alpha \sum\nolimits_j p(j|i,a) z_j\}, z \in \mathbb{R}^N, i \in \mathrm{X}. \tag{3.2}$$

It turns out that T is a *monotone contraction mapping* with as fixed point the value vector v^α. We also introduce, for any stationary policy π, the mapping $T_\pi : \mathbb{R}^N \to \mathbb{R}^N$, defined by (in vector notation)

$$T_\pi z := r(\pi) + \alpha P(\pi) z. \tag{3.3}$$

Let $f_z(i) \in \underset{\mathrm{A}(i)}{\operatorname{argmax}} \{r(i,a) + \alpha \sum_j p(j|i,a) z_j\}$, then

$$T_{f_z} z = Tz = \max_\pi T_\pi z. \tag{3.4}$$

First, we summarize some well-known results on discounted MDPs. The proofs can be found in the standard MDP textbooks. They are based on the theory of monotone contraction mappings. For this theory we refer to the book written by Stoer and Bulirsch [236].

Theorem 2.3 *With respect to the norm $\|\cdot\|_\infty$, T_π and T are monotone contraction mappings in $\mathbb{R}^N$ with contraction factor α.*

Theorem 2.4 *For any stationary policy π, $v^\alpha(\pi)$ is the unique solution of the functional equation $T_\pi z = z$.*

Corollary 2.2 $v^\alpha(\pi) = \lim_{n\to\infty} T_\pi^n z$ *for any* $z \in \mathbb{R}^N$.

Theorem 2.5 v^α *is the unique solution of the equation* $Tz = z$.

Because, $v^\alpha = Tv^\alpha = T_{f_{v^\alpha}} v^\alpha$, the last equality by (3.4), it follows from Theorem 2.5 that $v^\alpha = v^\alpha(f_{v^\alpha})$, i.e. f_{v^α} is an optimal policy. If f satisfies $r(i,f(i)) + \alpha \sum_j p(j|i,f(i)) v_j^\alpha = \max_a \{r(i,a) + \alpha \sum_j p(j|i,a) v_j^\alpha\}, i \in \mathbb{X}$, then f is called a *conserving* policy. f_{v^α} is a conserving policy and conserving policies are optimal. Therefore, the equation $Tz = z$ is called the optimality equation.

Corollary 2.3
(i) There exists a pure stationary optimal policy.
(ii) $v^\alpha = \lim_{n\to\infty} T^n z$ *for any* $z \in \mathbb{R}^N$.
(iii) Any conserving policy is optimal.

We use the following result from the theory of contracting mappings.

Lemma 2.2 *Let* B *be a monotone contraction in* $\mathbb{R}^N$ *with respect to* $\|\cdot\|_\infty$, *with contraction factor* β, *fixed-point* z^* *and with the property that* $B(z+c\cdot e) = Bz + \beta c\cdot e$ *for every* $z \in \mathbb{R}^N$ *and scalar* c. *Suppose that for some scalars* a *and* b *and for some* $z \in \mathbb{R}^N$, $a\cdot e \le Bz - z \le b\cdot e$. *Then,* $z + (1-\beta)^{-1}a\cdot e \le Bz + \beta(1-\beta)^{-1}a\cdot e \le z^* \le Bz + \beta(1-\beta)^{-1}b\cdot e \le z + (1-\beta)^{-1}b\cdot e$.

Since $T_f(z + c\cdot e) = T_f z + \alpha c\cdot e$ and $T(z+c\cdot e) = Tz + \alpha c\cdot e$ for any $z \in \mathbb{R}^N$ and any scalar c, we can apply Lemma 2.2 to obtain bounds for the fixed points $v^\alpha(f)$ and v^α of the operators T_f and T, respectively.

Lemma 2.3 *For any* $z \in \mathbb{R}^N$ *we have*
(i) $z + (1-\alpha)^{-1}\min_i(Tz-z)_i\cdot e \le Tz + \alpha(1-\alpha)^{-1}\min_i(Tz-z)_i\cdot e \le v^\alpha(f_z) \le v^\alpha \le Tz + \alpha(1-\alpha)^{-1}\max_i(Tz-z)_i\cdot e \le z + (1-\alpha)^{-1}\max_i(Tz-z)_i\cdot e$.
(ii) $\|v^\alpha - v^\alpha(f_z)\|_\infty \le \alpha(1-\alpha)^{-1}\mathrm{span}(Tz-z)$, *where* $\mathrm{span}(z)$ *is defined by* $\mathrm{span}(z) := \max_i z_i - \min_i z_i$.

An action $a \in \mathbb{A}(i)$ is called *suboptimal* if there does not exist an optimal policy f with $f(i) = a$. Because f is optimal if and only if $v^\alpha(f) = v^\alpha$, and because $v^\alpha = Tv^\alpha$, an action $a \in \mathbb{A}(i)$ is suboptimal if and only if

$$v_i^\alpha > r(i,a) + \alpha\sum_j p(j|i,a)v_j^\alpha. \tag{3.5}$$

Suboptimal actions can be excluded. Not directly by (3.5), because v^α is unknown, but by using the bounds on v^α as given by Lemma 2.3. Then, by the monotonicity of T, the next result is obtained.

Theorem 2.6
(i) Suppose that $x \le v^\alpha \le y$. *If* $r(i,a) + \alpha\sum_j p(j|i,a)y_j < (Tx)_i$, *then action* $a \in \mathbb{A}(i)$ *is suboptimal.*
(ii) Suppose that for some scalars b *and* c, $x + b\cdot e \le v^\alpha \le x + c\cdot e$. *If* $r(i,a) + \alpha\sum_j p(j|i,a)x_j < (Tx)_i - \alpha(c-b)$, *then action* $a \in \mathbb{A}(i)$ *is suboptimal.*

Using the bounds for v^α from Lemma 2.3, we obtain suboptimality for an action $a \in \mathbb{A}(i)$ if

$$r(i,a) + \alpha\sum_j p(j|i,a)z_j < (Tz)_i - \alpha(1-\alpha)^{-1}\mathrm{span}(Tz-z) \tag{3.6}$$

or

$$r(i,a) + \alpha\sum_j p(j|i,a)(Tz)_j < (T^2z)_i - \alpha^2(1-\alpha)^{-1}\mathrm{span}(Tz-z) \tag{3.7}$$

Remark
If we relax the property that $\sum_j p(j|i,a) = 1$ to $\sum_j p(j|i,a) \leq 1$ for all (i,a) and require that the model is *transient*, i.e. the matrix $\sum_{t=1}^{\infty} [P(f)]^t$ has finite elements for every policy f, then the total expected reward criterion, i.e. the discounting case with discount factor $\alpha = 1$, is well-defined. For this criterion similar results can be obtained as in the discounted model. The investigation whether an MDP is transient can be done efficiently (cf. Veinott [257] and Kallenberg [134]). Other references on this topic are van Hee, Hordijk and van der Wal [248], Denardo and Rothblum [53], and Hordijk and Kallenberg [115].

Already in 1953, Shapley [221] analyzed contraction properties for stochastic games. In the special case of a one-player game a stochastic game becomes an MDP. A comprehensive treatment of the theory of contraction mappings for discounted Markov decision processes was given by Denardo [45]. The details of the proof of Theorem 2.5 can be found in Ross [198]. An alternative proof that T has a fixed point, based on Brouwer's theorem, was given in Shapiro [220]. The concepts 'conserving' and 'span' were introduced by Dubins and Savage [64] and by Bather [10]. Concerning the bounds of Lemma 2.3, the weakest bounds were proposed by MacQueen [162] and the strongest by Porteus [179]. Related papers are Porteus [180] and Bertsekas [16]. The notion that suboptimal actions can be excluded if bounds on the value vector are available can be found in MacQueen [163], which paper includes the test (3.6). Test (3.7) is proposed in Porteus [179]. Other suboptimality tests can be found in Hastings and Mello [97], White [278] and Thomas [239].

2.3.2 *Policy iteration*

For $x, y \in \mathbb{R}^N$, $x > y$ means that $x_i \geq y_i$ for every i and $x_i > y_i$ for at least one i. In the method of policy iteration a sequence of pure stationary policies $f_1, f_2, \ldots$ is constructed such that

$$v^\alpha(f_{k+1}) > v^\alpha(f_k) \text{ for } k = 1, 2, \ldots \tag{3.8}$$

Because there are finitely many pure stationary policies, the method of policy iteration is finite. Furthermore, it can be shown that the method terminates with an α-discounted optimal policy. We first remark that the following lemma is a consequence of Theorem 2.3.

Lemma 2.4
(i) If $T_f z \leq z$, then $v^\alpha(f) = \lim_{n\to\infty} T_f^n z \leq T_f z \leq z$.
(ii) If $T_f z > z$, then $v^\alpha(f) = \lim_{n\to\infty} T_f^n z \geq T_f z > z$.

For every $i \in \mathbf{X}$ and every $f \in F$, the set $A(i,f)$ is defined by

$$A(i,f) := \{a \in \mathbf{A}(i) \mid r(i,a) + \alpha \sum_j p(j|i,a) v_j^\alpha(f) > v_i^\alpha(f)\}. \tag{3.9}$$

The intuitive idea of policy iteration is that if action $f(i)$ is replaced by an action $a \in A(i,f)$ the resulting policy improves the α-discounted rewards.

Therefore, the actions of $A(i,f)$ are called *improving actions*. The correctness of this idea is established by the following theorem.

Theorem 2.7
(i) If $A(i,f) = \emptyset$ for every $i \in \mathbb{X}$, then f is α-discounted optimal.
(ii) If $A(i,f) \neq \emptyset$ for some $i \in \mathbb{X}$, then $v^\alpha(g) > v^\alpha(f)$ for any $g \in F$ with $g \neq f$ and $g(i) \in A(i,f)$ if $g(i) \neq f(i)$.

Let

$$s_{ia}(f) := r(i,a) + \alpha \sum_j p(j|i,a) v_j^\alpha(f) - v_i^\alpha(f), \quad a \in \mathbb{A}(i) \text{ and } i \in \mathbb{X}. \tag{3.10}$$

Algorithm II (policy iteration; discounted rewards)

1. Start with any $f \in F$.
2. Compute $v^\alpha(f)$ as unique solution of the linear system $T_f z = z$.
3. $A(i,f) := \{a \in \mathbb{A}(i) \mid s_{ia}(f) > 0\}$ for every $i \in \mathbb{X}$.
4. If $A(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that, if $g(i) \neq f(i), g(i) \in A(i,f)$.
5. $f := g$ and go to step 2.
6. f is an α-discounted optimal policy.

The idea to use policy iteration to determine an optimal policy appeared in Howard [121]. Blackwell [27] has provided a strong mathematical treatment of this method. In Porteus [183] and in Hartley, Lavercombe and Thomas [92] efficient ways are analyzed in order to calculate $v^\alpha(f)$ as solution of the linear system $T_f z = z$.

Remarks
1. There is some freedom in the choice of policy g in step 4. A usual choice is to take g such that $s_{ig(i)}(f) = \max_a s_{ia}(f)$, i.e. $g(i) \in \operatorname{argmax}_a s_{ia}(f)$.
2. It can be shown (see Puterman and Brumelle [187]) that the policy iteration method, with the above choice for g, is equivalent to solving the optimality equation $Tz = z$ by Newton's method.
3. Furthermore, we can derive a result on the convergence rate. It can be shown that $x^n = v^\alpha(f_n), n = 1, 2, \ldots$, where x^n are the iterates of the Newton method and f_n the policies of the policy iteration method. Since it can be shown that $\| v^\alpha - v^\alpha(f_{n+1}) \|_\infty \leq 2\alpha(1-\alpha)^{-1} \| v^\alpha - v^\alpha(f_n) \|_\infty$, there is geometric convergence. Already in Pollatschek and Avi-Itzhak [178], in the context of stochastic games, the equivalence between the policy iteration method and Newton's method was noticed. A related paper is Schweitzer [211]. Puterman and Brumelle [187] were the first who derived results for the rate of convergence.
4. One can also exclude suboptimal actions. E.g. by using test (3.6) with $z = v^\alpha(f)$. Since, for $z = v^\alpha(f)$,

$(Tz - z)_i = \max_a \{r(i,a) + \sum_j p(j|i,a)v_j^\alpha(f) - v_i^\alpha(f) = \max_a s_{ia}(f), i \in \mathbb{X}$, we have $\text{span}(Tz - z) = \max_i[\max_a\ s_{ia}(f)] - \min_i[\max_a\ s_{ia}(f)]$. Hence (3.6) becomes:

if $s_{ib}(f) < \max_a\ s_{ia}(f) - \alpha(1-\alpha)^{-1}[\max_i \max_a s_{ia}(f) - \min_i \max_a\ s_{ia}(f)]$, then action $b \in \mathbb{A}(i)$ is suboptimal.

Grinold [91] pointed out that suboptimality tests can be implemented in policy iteration. The above test is stronger than Grinold's test.

5. The following modification, which was shown to be correct by Hastings [94], often gives faster convergence. Instead of the steps 3 and 4, the steps 3' and 4' are used, where:

Step 3':

For i = 1 to N do

a. $d_{ia}(f) := r(i,a) + \alpha\sum_{j=1}^{i-1} p(j|i,a)z_j + \alpha\sum_{j=i}^{N} p(j|i,a)v_j^\alpha(f), a \in \mathbb{A}(i)$;

b. if $d_{ia}(f) \leq v_i^\alpha(f)$ for every $a \in \mathbb{A}(i)$:

$z_i := v_i^\alpha(f)$ and $g(i) := f(i)$;

c. if $d_{ia}(f) > v_i^\alpha(f)$ for some $a \in \mathbb{A}(i)$:

$z_i := \max_a\ d_{ia}(f)$ and take $g(i) \in \text{argmax}\ d_{ia}(f)$.

Step 4':

If $g(i) = f(i)$ for every $i \in \mathbb{X}$, then go to step 6.

6. In Schmitz [204] the question is raised: "Does there exist a polynomial bound for the number of iterations in the policy iteration?". Meister and Holzbaur [165] have shown that this method is polynomial in time. In Ng [171] is shown that the complexity of one iteration is $\mathcal{O}(mN^2)$, where m is the number of states i for which $g(i) \neq f(i)$.

2.3.3 *Linear programming*

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

$$v_i \geq r(i,a) + \alpha\sum_j p(j|i,a)v_j \text{ for every } (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.11}$$

Theorem 2.8 *v^α is the (componentwise) smallest α-superharmonic vector.*

Corollary 2.4 *v^α is the unique optimal solution of the LP-problem*

$$\min\left\{\sum_j \beta_j v_j \mid \sum_j [\delta_{ij} - \alpha p(j|i,a)]v_j \geq r(i,a), (i,a) \in \mathbb{X} \times \mathbb{A}\right\} \tag{3.12}$$

where $\beta_j > 0$ for every $j \in \mathbb{X}$.

By Corollary 2.4, the value vector v^α can be found as the optimal solution of the linear program (3.12). This program does not give an optimal policy. However, an optimal policy can be obtained from the solution of the dual program:

$$\max\left\{\sum_i\sum_a r(i,a)x_{ia} \;\middle|\; \begin{array}{l}\sum_i\sum_a[\delta_{ij} - \alpha p(j|i,a)]\ x_{ia} = \beta_j, j \in \mathbb{X} \\ x_{ia} \geq 0, (i,a) \in \mathbb{X} \times \mathbb{A}\end{array}\right\} \tag{3.13}$$

Theorem 2.9 *Let x^* be an optimal solution of (3.13). Then, a policy f with $x^*_{jf(j)} > 0$ for every $j \in \mathbb{X}$ exists and is an optimal policy.*

There is a one-to-one correspondence between the set of feasible solutions of (3.13) and the set of stationary policies, given by the following relations. For a stationary policy π the feasible solution $x(\pi)$ satisfies

$$x_{ia}(\pi) := [\beta^T(I - \alpha P(\pi))^{-1}]_i \cdot \pi_{ia}, (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.14}$$

Conversely, for a feasible solution x of (3.13), define $\pi(x)$ by

$$\pi_{ia}(x) := x_{ia} / \sum\nolimits_a x_{ia}, (i,a) \in \mathbb{X} \times \mathbb{A}. \tag{3.15}$$

Theorem 2.10 *The mapping (3.14) is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the dual program (3.13) with (3.15) as the inverse mapping; furthermore, the set of extreme feasible solutions of (3.13) corresponds to the set F of pure stationary policies.*

Algorithm III (linear programming; discounted rewards)

1. Take any $\beta \in \mathbf{R}^N$ with $\beta_j > 0, j \in \mathbb{X}$.
2. Compute optimal solutions v^* and x^* of the dual pair LP-problems (3.12) and (3.13).
3. Take any f_* such that $x^*_{if_*(i)} > 0$, $i \in \mathbb{X}$.
4. v^* is the value vector and f_* is an α-discounted optimal policy.

It turns out that the linear programming method is, in some sense, equivalent to policy iteration. This is formulated in the next theorem, in which the term *block-pivoting* simplex algorithm is used. A simplex LP-algorithm, which in one iteration more than one pivot step may use, is called a block-pivoting simplex algorithm (cf. Dantzig [40]).

Theorem 2.11
(i) Any policy iteration algorithm is equivalent to a block-pivoting simplex algorithm.
(ii) Any simplex algorithm is equivalent to a particular policy iteration algorithm.

Remarks
1. Since the LP-method and policy iteration are equivalent, exclusion of suboptimal actions can also be implemented in the LP-method. The relevant data $s_{ia}(f)$ for this test (see (3.10)) are available in the simplex tableaus as the so-called reduced costs.
2. The variables $x_{ia}(\pi)$, defined in (3.14) can be interpreted as *discounted state-action frequencies*, i.e. if policy π is used, then $x_{ia}(\pi)$ is equal to the total

expected discounted number of times that state i is visited and then also action a is chosen, given that the starting state is state j with probability $\beta_j, j \in \mathbb{X}$. 3. The linear programming method is the only method which can easily handle additional constraints. Constrained optimization arises in many MDP applications, e.g. in inventory and queueing models. For examples we refer to Derman [58], chapter 7, and to Puterman [186], section 8.9. The constraints have to be expressed in terms of the state-action frequencies and added to the dual program (3.13). Constrained problems have a stationary, but not necessarily pure optimal policy. For the details we refer to Kallenberg [134], and to Hordijk and Kallenberg [115]. In Altman and Shwartz [4] the sensitivity of constrained MDP's is investigated. Altman, Hordijk and Kallenberg [3] have analyzed the behavior of the value function in constrained MDP.

The idea to use linear programming to compute an optimal policy originated with D'Epenoux [55]. The one-to-one correspondence between the feasible solutions of the dual program and the set of stationary policies can be found in De Ghellinck and Eppen [43]. The equivalence between block-pivoting and policy iteration was mentioned in De Ghellinck [42]. The implementation of the suboptimality tests was proposed by Grinold [91] and by Hordijk and Kallenberg [115]. In Sun [238] an implementation of the LP-method is described, based on the revised simplex method. Stein [233] has investigated the computational aspects of the linear programming method in comparison with other methods as policy iteration, modified policy iteration and value iteration. It turns out that the LP-method is preferable if the discount factor is close to unity and the state space is not too large. Chapter 12 deals with the linear programming approach for MDPs with infinite state and action spaces.

2.3.4 *Value iteration*

In the value iteration method the value vector v^α is approximated by a sequence $\{v^n\}_{n=1}^\infty$, which converges to v^α and in this way a nearly optimal policy is obtained. For $\epsilon > 0$, a vector $v \in \mathbb{R}^N$ is an *ϵ-approximation* of v^α if $\| v^\alpha - v \|_\infty \leq \epsilon$; a policy R is an *ϵ-optimal policy* if $\| v^\alpha - v^\alpha(R) \|_\infty \leq \epsilon$, i.e. $v^\alpha(R)$ is an ϵ-approximation of v^α. From Corollary 2.3 (ii) it follows that $v^\alpha = \lim_{n\to\infty} T^n x$ for every $x \in \mathbb{R}^N$.

Define the sequence $\{v^n\}_{n=1}^\infty$ by

$$\begin{cases} v^1 \in \mathbb{R}^N & \text{arbitrarily chosen} \\ v^{n+1} := Tv^n, & n = 1, 2, \ldots \end{cases} \tag{3.16}$$

with a corresponding sequence $f_1, f_2, \ldots$ of policies where $f_n = f_{v^n}$ for every $n \in \mathbb{N}$, i.e.

$$v^{n+1} = Tv^n = T_{f_n} v^n = r(f_n) + \alpha P(f_n) v^n, n \in \mathbb{N}. \tag{3.17}$$

The next lemma shows that f_n is an ϵ-optimal policy for n sufficiently large. The proof is based on contraction properties.

Lemma 2.5 $\| v^\alpha(f_n) - v^\alpha \|_\infty \leq 2\alpha^n (1-\alpha)^{-1} \cdot \| v^2 - v^1 \|_\infty, n \in \mathbb{N}.$

Algorithm IV (value iteration; discounted rewards)

1. Choose $\epsilon > 0$ and $x \in \mathbb{R}^N$ arbitrarily.
2. Compute $y = Tx$ and take $f = f_x$.
3. If $\| y - x \|_\infty \leq (1-\alpha)\alpha^{-1}\epsilon$, then f is a 2ϵ-optimal policy and y is an ϵ-approximation of v^α (Stop);

   Otherwise: $x := y$ and go to step 2.

The correctness of algorithm IV is a consequence of the next theorem, which also follows from the contraction properties.

Theorem 2.12
(i) $\| v^\alpha(f_x) - v^\alpha \|_\infty \leq 2\alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty$;
(ii) $\| Tx - v^\alpha \|_\infty \leq \alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty$.

In the next theorem we summarize some suboptimality tests.

Theorem 2.13 *An action $a \in \mathbb{A}(i)$ is suboptimal if one of the following tests is satisfied:*

$$r(i,a) + \alpha \sum_j p(j|i,a)x_j < (Tx)_i - 2\alpha(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty \tag{3.18}$$

$$r(i,a) + \alpha \sum_j p(j|i,a)(Tx)_j < (T^2x)_i - 2\alpha^2(1-\alpha)^{-1} \cdot \| Tx - x \|_\infty \tag{3.19}$$

$$r(i,a) + \alpha \sum_j p(j|i,a)x_j < (Tx)_i - \alpha(1-\alpha)^{-1}\operatorname{span}(Tx - x) \tag{3.20}$$

$$r(i,a) + \alpha \sum_j p(j|i,a)(Tx)_j < \begin{array}{l}(Tx)_i + \alpha(1-\alpha)^{-1}\min_i(Tx-x)_i \\ -\alpha^2(1-\alpha)^{-1}\max_i(Tx-x)_i\end{array} \tag{3.21}$$

Remarks
1. In the usual computation scheme of the value iteration algorithm, test (3.20) is the best available test.
2. We also mention two variants of the standard algorithm. In the *Pre-Gauss-Seidel* variant we use for the computation of y_i the components y_j which are already computed, i.e.

$$y_i = \max_a \{ r(i,a) + \alpha \sum_{j=1}^{i-1} p(j|i,a)y_j + \alpha \sum_{j=i}^{N} p(j|i,a)x_j \}, \quad i = 1, 2, \ldots, N \tag{3.22}$$

In the *Gauss-Seidel* variant also the i-th component x_i is replaced by y_i, which gives

$$\begin{aligned} y_i = \max_a [1 - \alpha p(i|i,a)]^{-1} \cdot \{ & r(i,a) + \alpha \sum_{j=1}^{i-1} p(j|i,a)y_j \\ & + \alpha \sum_{j=i+1}^{N} p(j|i,a)x_j \}, \quad i = 1, 2, \ldots, N \end{aligned} \tag{3.23}$$

For both variants it can be shown that the corresponding operators are contraction mappings with fixed point v^σ and with contraction factor at most α. Hence, they may be considered as an acceleration of the basic algorithm. Suboptimaly tests for the exclusion of actions can also be included in the value iteration method.

Value iteration goes back to the seminal paper of Shapley [221] on stochastic games. For a survey of the basic properties of value iteration we refer to Federgruen and Schweitzer [70]. The idea to accelerate the convergence by the pre-Gauss-Seidel method was proposed in Hastings [94]. The Gauss-Seidel method can be found in Kushner and Kleinman [149]. An overview of these variants is presented in Porteus [182]. Other techniques, based on successive overrelaxation and stopping times, in order to accelerate the convergence can be found in Reetz [190] and [191], Schellhaas [203], Wessels [272], Van Nunen [249], Van Nunen and Wessels [252], [253], Porteus and Totten [184], Porteus [181], Herzberg and Yechiali [105], and Bertsekas [20]. Holzbaur [110] has presented a theoretically polynomial bound for the number of steps in the value iteration method.

2.3.5 *Modified policy iteration*

In section 1.3.2 the policy iteration method was discussed. This method, with the usual choice for the improving actions, can be considered as Newton's method for the solution of the optimality equation. A new iterand y is obtained from x by the formula

$$y = x + A(Tx - x), \text{ where } A = [I - \alpha P(g)]^{-1} \text{ with } g \text{ such that } T_g x = Tx \tag{3.24}$$

The determination of $[I-\alpha P(g)]^{-1}$, which is equal to $\sum_{i=0}^{\infty} \alpha^i [P(g)]^i$, requires in general a lot of work. In the ***modified policy iteration method*** the matrix A is truncated by

$$A^{(k)} = \sum_{i=0}^{k-1} \alpha^i [P(g)]^i \text{ for some } 1 \leq k \leq \infty\,. \tag{3.25}$$

For $k = 1$, $A^{(k)} = I$ and the value iteration method is obtained; for $k = \infty$, $A^{(k)} = A$, and we have policy iteration. For $1 < k < \infty$, the modified policy iteration method can be considered as a combination of policy iteration and value iteration, or as an inexact Newton method for the solution of the optimality equation.

We may allow that in each iteration another value of k is chosen, and we denote $k(n)$ for the value in iteration n. Hence, we obtain the following iteration scheme, where f_n is the policy in iteration n, i.e. $T_{f_n} x^n = Tx^n$.

$$\begin{aligned} x^{n+1} &= x^n + A^{(k(n))}(Tx^n - x^n) \\ &= x^n + \textstyle\sum_{i=0}^{k(n)-1} \alpha^i P^i(f_n)[r(f_n) + \alpha P(f_n)x^n - x^n] \\ &= r(f_n) + \alpha P(f_n) r(f) + \cdots + [\alpha P(f_n)]^{k(n)-1} r(f_n) + [\alpha P(f_n)]^{k(n)} x^n \\ &= T_{f_n}^{k(n)} x^n. \end{aligned}$$

Algorithm V (modified policy iteration; discounted rewards)

1. Choose $x \in \mathbf{R}^N$, $\epsilon > 0$ and $f \in F$.
2. a. Choose k with $1 \leq k \leq \infty$;
   b. Determine g such that $T_g x = Tx$, where $g(i) = f(i)$ if possible.
3. If $\| Tx - x \|_\infty \leq (1-\alpha)\epsilon$: g is an 2ϵ-optimal policy and Tx is an $\alpha\epsilon$-approximation of v^α (Stop);
   Otherwise: $x := T_g^k x$, $f := g$ and go to step 2.

Remarks
1. Since $x^{n+1} = T_{f_n}^{k(n)} x^n$, the iteration operator depends on n, and it is not obvious that this operator is monotone and/or contracting. Indeed, in general, this operator is neither a contraction nor monotone. Nevertheless, it can be shown that $v^\alpha = \lim_{n\to\infty} T_{f_n}^{k(n)} x^n$ for any starting vector x^1.
2. Also in this method, it is possible to implement tests for the exclusion of suboptimal actions.

Puterman and Shin [188] and independently Van Nunen [249], [250] and [251] have developed the modified policy iteration method. The first authors have shown the convergence under the assumption that the starting vector x satisfies $Tx \geq x$. The convergence of the method for an arbitrary starting vector was proved by Rothblum [201]. In Van Nunen [249] an example is given which shows that the operator of the modified policy iteration method can be neither contracting nor monotonic. The observation that the modified policy iteration method can be viewed as an inexact Newton method was made by Dembo and Haviv [44]. The exclusion of suboptimal actions for this method was developed by Puterman and Shin [189]. Puterman [185] reviews computational results for the modified policy iteration method.

2.4 AVERAGE REWARD CRITERION

2.4.1 *Introduction*

We start this section with some properties of a transition matrix P. The *stationary matrix* P^* of P is defined by the Cesaro-limit of P^n, i.e.

$$P^* := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P^{k-1} \tag{4.1}$$

The next theorem summarizes some properties of the stationary matrix.

Theorem 2.14
(i) $P^*P = PP^* = P^*P^* = P^*$.
(ii) $[P - P^*]^n = P^n - P^*$, $n \geq 1$.
(iii) $\lim_{\alpha\uparrow 1}(1-\alpha)\sum_{n=0}^{\infty} \alpha^n(P^n - P^*) = 0$.
(iv) $[I - P + P^*]^{-1} = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}[P - P^*]^{i-1}$.
(v) For every stationary policy π, *the average reward* $\phi(\pi)$ *satisfies* $\phi(\pi) = P^*(\pi)r(\pi)$, *where* $P^*(\pi)$ *is the stationary matrix of the transition matrix* $P(\pi)$.

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

$[I - P + P^*]^{-1}$ is denoted by Z and is called the *fundamental matrix*. The *deviation matrix* D is defined by

$$D := Z - P^* \tag{4.2}$$

Theorem 2.15
(i) $D = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}[P^{i-1} - P^*]$
(ii) $P^*D = DP^* = [I - P]D + P^* - I = D[I - P] + P^* - I = 0$.

For the proofs of the Theorems 2.14 and 2.15 we refer to books on Markov chains (e.g. Kemeny and Snell [144]). A treatment, related to MDPs, of the stationary, the fundamental and the deviation matrix can be found in Veinott [258]. For a stationary policy π, the deviation matrix of the transition matrix $P(\pi)$ is denoted by $D(\pi)$.

We continue this section with a classification of MDPs based on the ergodic structure. We distinguish between multichain, unichain and irreducible MDPs. The reason for this distinction is that MDPs can be analyzed easier in case they are unichain or irreducible, which may lead to simplified algorithms for solving these MDPs. We assume the reader is familiar with concepts as recurrent state, transient state, recurrent class, irreducibility, unichain and multichain. The determination whether an MDP is irreducible is an easy, i.e. polynomially solvable, problem (the number of steps is bounded by a polynomial function of the problem's data, see Kallenberg [137]).

Open problem
Does there exist a polynomial algorithm to determine whether an MDP is unichain or multichain?

Next, we will formulate a theorem on the existence of a Blackwell optimal policy f_0, i.e. f_0 is α-discounted optimal for all discount factors $\alpha \in [\alpha_0, 1)$ for some $0 \leq \alpha_0 < 1$. The next theorem shows even more, namely that the interval $[0, 1)$ can be partitioned in a finite number of subintervals such that in each subinterval there exists a policy which is discounted optimal over the whole subinterval. Since a proof of this result cannot be found in the textbooks on MDPs, we include an outline of this proof.

Theorem 2.16 *There are numbers* $\alpha_m, \alpha_{m-1}, ..., \alpha_0, \alpha_{-1}$ *and policies* f_m, $f_{m-1}, ..., f_0$ *such that* $0 = \alpha_m < \alpha_{m-1} < \cdots < \alpha_0 < \alpha_{-1} = 1$ *and* $v^{\alpha}(f_j) = v^{\alpha}$ *for all* $\alpha \in [\alpha_j, \alpha_{j-1}), j = m, m-1, ..., 0$.

Proof (outline). Since $v^{\alpha}(f)$ is the solution of the system $[I - \alpha P(f)]x = r(f)$, each component $v_i^{\alpha}(f)$ is a rational function in α. Suppose that a Blackwell optimal policy does not exist. Since for any fixed α a deterministic α-discounted optimal policy exists, this implies that there are series $\{\alpha_k \mid k = 1, 2, \ldots\}$ and $\{f_k \mid k = 1, 2, \ldots\}$ such that $\alpha_1 \leq \alpha_2 \leq \cdots$ with $\lim_{k\to\infty} \alpha_k = 1$ and $v^{\alpha} = v^{\alpha}(f_k) > v^{\alpha}(f_{k-1})$ for $\alpha = \alpha_k, k = 2, 3, \ldots$. Because F is finite, there

are two pure policies, say f and g, that both are in turn optimal for an infinite number of increasing α's with limit $\alpha = 1$. Let $h(\alpha) = v^\alpha(f) - v^\alpha(g)$, then for any $i \in \mathbb{X}$, $h_i(\alpha)$ is a continuous rational function in α on $[0,1)$, which has an infinite number of zeros. This contradicts the rationality of $h_i(\alpha)$. Hence, there exists a Blackwell optimal policy. With similar arguments, it can be shown that for each fixed $\alpha \in (0,1]$ there is an interval around α and a policy which is optimal in that interval. These intervals are a covering of the closed bounded set $[0,1]$. Hence, by the Heine-Borel-Lebesque theorem, it follows that there is a covering by a finite number of intervals. ∎

We close this section with the *Laurent expansion* of a stationary policy π.

Theorem 2.17 *Let $u^k(\pi), k = -1, 0, \ldots$ be defined by $u^{-1}(\pi) = P^*(\pi)r(\pi)$, $u^0(\pi) = D(\pi)r(\pi)$ and $u^{k+1}(\pi) = -D(\pi)u^k(\pi)$, $k \geq 0$. Then, for $\alpha_0(\pi) < \alpha < 1$, we have $v^\alpha(\pi) = \alpha^{-1}\sum_{k=-1}^{\infty}[(1-\alpha)/\alpha]^k \cdot u^k(\pi)$, where $\alpha_0(\pi) = \| D(\pi) \| / [\| D(\pi) \| + 1]$.*

Corollary 2.5
(i) $\phi(\pi) = \lim_{\alpha \uparrow 1}(1-\alpha)v^\alpha(\pi)$.
(ii) $v^\alpha(\pi) = \frac{\phi(\pi)}{1-\alpha} + u^0(\pi) + \epsilon(\alpha)$, *where* $\lim_{\alpha\uparrow 1}\epsilon(\alpha) = 0$.

The first part of the Laurent expansion as presented in Corollary 2.5 (ii) was derived by Blackwell [27]. The complete Laurent expansion of Theorem 2.17 was proposed by Miller and Veinott [166]. The vector $u^0(\pi)$ is called the *bias vector* of policy π.

2.4.2 *The optimality equation*

A. The multichain case.
Before we introduce the optimality equation, we first give some prerequisites.

Lemma 2.6 $\lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(R) \geq \phi(R)$ *for any policy R.*

The proof of this theorem is based on Tauberian arguments which can be found in Derman [58] or Hordijk [111].

Corollary 2.6 *Any stationary Blackwell optimal policy is average optimal.*

In the discounted case, the value vector is the unique solution of an optimality equation. A similar result holds for the average reward criterion, but the derivation is more complex.

Theorem 2.18 *Consider the system*

$$\begin{cases} x_i & = \max_{a\in\mathbb{A}(i)} \sum_j p(j|i,a)x_j, & i \in \mathbb{X} \\ x_i + y_i & = \max_{a\in\mathbb{A}(i,x)}\{r(i,a) + \sum_j p(j|i,a)y_j\}, & i \in \mathbb{X} \end{cases} \tag{4.3}$$

where $\mathbb{A}(i,x) := \{a \in \mathbb{A}(i) \mid x_i = \sum_j p(j|i,a)x_j\}$, $i \in \mathbb{X}$.

This system has the following properties
(i) With f_0 any Blackwell optimal policy, $x = u^{-1}(f_0)$ and $y = u^0(f_0)$ satisfy (4.3).
(ii) If (x, y) is a solution of (4.3), then x equals the value vector ϕ.

B. The unichain case.
In the unichain case, for every policy f, the stationary matrix $P^*(f)$ has identical components. Hence, the value vector ϕ is a constant vector. We will denote this constant vector by $\phi \cdot e$ (ϕ is a scalar). The first part of the optimality equation is always satisfied and the following result can be derived.

Theorem 2.19 *Consider the system $x+y_i = \max_a\{r(i,a)+\sum_j p(j|i,a)y_j\}, i \in \mathbb{X}$. This system has the following properties:*
(i) With f_0 any Blackwell optimal policy, $x \cdot e = u^{-1}(f_0)$ and $y = u^0(f_0)$, satisfy this system.
(ii) If (x, y) is a solution of the system, then $x = \phi$ and $y = u^0(f_0) + c \cdot e$ for some constant c.

The functional equation (4.3) is extensively investigated in Schweitzer and Federgruen [216]. Another proof for the solution of the optimality equation can also be provided by applying Brouwer's fixed point theorem (see Federgruen and Schweitzer [72], and Schweitzer [212]). In the unichain case the solution of the optimality equation can be exhibited as the fixed point of an N-step contraction (cf. Federgruen, Schweitzer and Tijms [74]).

2.4.3 Policy iteration

In the policy iteration method a sequence of policies $f_1, f_2, \ldots$ is constructed such that $\phi(f_{k+1}) \geq \phi(f_k)$ and $v^\alpha(f_{k+1}) > v^\alpha(f_k)$ for all $\alpha \in (\alpha_k, 1)$. Since F is finite and all policies f_k are different, this method has finite termination with an optimal policy.

A. The multichain case.

Theorem 2.20 *Consider the following system of linear equations*

$$\begin{cases} [I - P(f)]\ x & & & = 0 \\ & x + [I - P(f)]\ y & & = r(f) \\ & & y + [I - P(f)]z & = 0. \end{cases} \tag{4.4}$$

Then, (4.4) has a solution $(x(f), y(f), z(f))$, where $x(f)$ and $y(f)$ are unique with $x(f) = u^{-1}(f)$ and $y(f) = u^0(f)$.

For every $i \in \mathbb{X}$ and every policy f, we define the action subset $B(i, f)$ by

$$B(i,f) := \left\{ a \in \mathbb{A}(i) \;\middle|\; \begin{array}{l} \sum_j p(j|i,a)\phi_j(f) > \phi_i(f) \text{ or} \\ \sum_j p(j|i,a)\phi_j(f) = \phi_i(f) \text{ and} \\ r(i,a) + \sum_j p(j|i,a)u_j^0(f) > \phi_i(f) + u_i^0(f) \end{array} \right\} \tag{4.5}$$

Theorem 2.21
(i) If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$, then f is an average optimal policy.
(ii) If $B(i,f) \neq \emptyset$ for at least one i and the policy $g \neq f$ satisfies for each state i: $g(i) \in B(i,f)$ if $g(i) \neq f(i)$, then $\phi(g) \geq \phi(f)$ and $v^\alpha(g) > v^\alpha(f)$ for α sufficiently close to 1.

Algorithm VI (policy iteration; average reward, multichain case)

1. Start with any $f \in F$.
2. Determine $\phi(f)$ and $u^0(f)$ as the unique (x,y)-part in a solution of the linear system (4.4).
3. For every $i \in \mathbb{X}$: determine $B(i,f)$ as defined in (4.5).
4. If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that $g(i) \in B(i,f)$ if $g(i) \neq f(i)$.
5. $f := g$ and go to step 2.
6. f is an average optimal policy.

B. The unichain case.
In the unichain case, since the average reward vectors are constant, the set $B(i,f)$ can be simplified to

$$B(i,f) := \{a \in \mathbb{A}(i) \mid r(i,a) + \sum_j p(j|i,a)u_j^0 > \phi(f) + u_i^0(f)\} \qquad (4.6)$$

The following result holds.

Theorem 2.22 *The linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$, has a unique solution $x = \phi(f)$ and $y = u^0(f) - u_1^0(f)$.*

Algorithm VII (policy iteration; average reward, unichain case)

1. Start with any $f \in F$.
2. Determine $\phi(f)$ and $u^0(f)$ as the unique solution of the linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$.
3. For every $i \in \mathbb{X}$: determine $B(i,f)$ as defined in (4.6).
4. If $B(i,f) = \emptyset$ for every $i \in \mathbb{X}$: go to step 6.
   Otherwise: take any $g \neq f$ such that $g(i) \in B(i,f)$ if $g(i) \neq f(i)$.
5. $f := g$ and go to step 2.
6. f is an average optimal policy.

The concept of policy iteration is originated by Howard [121] who considered the first two parts of system (4.4). However, in that case, the convergence is

not always guaranteed and cycling can occur. Blackwell [27] has given a convergent version by imposing the constraint $P^*(f)y = 0$; the formulation with system (4.4) was proposed by Miller and Veinott [166]. In Blackwell's version, in order to compute $P^*(f)$, the chain structure of the transition matrix $P(f)$ has to be analyzed. Other anti-cycling rules, which avoid the analysis of the chain structure, are introduced in Schweitzer and Federgruen [215], Federgruen and Spreen [75], and Spreen [232]. Various treatments of the policy iteration method in the unichain case (or other special cases) can be found in Schweitzer [208], Denardo [49], Haviv and Puterman [100], and Lasserre [151].

2.4.4 *Linear programming*

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

$$\begin{cases} v_i & \geq \sum_j p(j|i,a)v_j & \text{for every} (i,a) \in \mathbb{X} \times \mathbb{A} \\ v_i + u_i & \geq r(i,a) + \sum_j p(j|i,a)u_j & \text{for every} (i,a) \in \mathbb{X} \times \mathbb{A} \end{cases} \tag{4.7}$$

Theorem 2.23 *The value vector ϕ is the (componentwise) smallest average-superharmonic vector.*

Proof (outline). Let f_0 be a Blackwell optimal policy. From Theorem 2.18 it follows that $\phi_i \geq \sum_j p(j|i,a)\phi_j, (i,a) \in \mathbb{X} \times \mathbb{A}$, and $\phi_i + u_i^0(f_0) \geq r(i,a) + \sum_j p(j|i,a)u_j^0(f_0)$ for every $i \in \mathbb{X}$ and $a \in \mathbb{A}(i,\phi)$. Then, it can be shown that (ϕ, u) is average-superharmonic, where $u = u^0(f_0) + M \cdot \phi$ with M sufficiently large. Suppose that y is also average-superharmonic with corresponding x. Then, $y \geq P(f_0)y$, implying that $y \geq P^*(f_0)y \geq P^*(f_0)\{r(f_0)+[P(f_0)-I]x\} = P^*(f_0)r(f_0) = \phi(f_0) = \phi$, i.e. ϕ is the smallest average-superharmonic vector.∎

A. The multichain case.

Corollary 2.7 *Let (u, v) be an optimal solution of the linear program*

$$\min \left\{ \sum_j \beta_j v_j \,\middle|\, \begin{array}{ll} \sum_j [\delta_{ij} - p(j|i,a)]v_j \geq 0 & ,(i,a) \in \mathbb{X} \times \mathbb{A} \\ v_i + \sum_j [\delta_{ij} - p(j|i,a)]u_j \geq r(i,a) & ,(i,a) \in \mathbb{X} \times \mathbb{A} \end{array} \right\} \tag{4.8}$$

where $\beta_j > 0$, $j \in \mathbb{X}$, is arbitrarily chosen, then $u = \phi$.

The dual program of (4.8) is

$$\max \left\{ \sum_{i,a} r(i,a)x_{ia} \,\middle|\, \begin{array}{ll} \sum_{i,a} [\delta_{ij} - p(j|i,a)]x_{ia} & = 0 \quad ,j \in \mathbb{X} \\ \sum_a x_{ja} + \sum_{i,a} [\delta_{ij} - p(j|i,a)]y_{ia} & = \beta_j \quad ,j \in \mathbb{X} \\ x_{ia}, y_{ia} \geq 0, (i,a) \in \mathbb{X} \times \mathbb{A} & \end{array} \right\} \tag{4.9}$$

For any feasible solution (x, y) of (4.9) we denote by $\mathbb{X}_x$

$$\mathbb{X}_x := \{j \in \mathbb{X} \mid \sum_a x_{ja} > 0\} \tag{4.10}$$

Theorem 2.24 *Let (x, y) be an extreme optimal solution of (4.9). Then, any policy f such that* $\begin{cases} x_{if(i)} > 0 & if \quad i \in \mathbb{X}_x \\ y_{if(i)} > 0 & if \quad i \notin \mathbb{X}_x \end{cases}$ *is an average optimal policy.*

Proof (outline). Because for every $j \in \mathbb{X}$, $\sum_a x_{ja} + \sum_a y_{ja} = \sum_{i,a} p(j|i,a) y_{ia} + \beta_j > 0$, policy f is well defined. Since $x_{if(i)} > 0$, $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$, $i \notin \mathbb{X}_x$, it follows from the complementary slackness of linear programming that $\phi_i + \sum_j [\delta_{ij} - p(j|i,(f(i))] u_j = r(i, f(i)), i \in \mathbb{X}_x$ and $\sum_j [\delta_{ij} - p(j|i, f(i))]\phi_j = 0$, $i \notin \mathbb{X}_x$. Program (4.8) implies that $\sum_j [\delta_{ij} - p(j|i,a)]\phi_j \geq 0$, $i \in \mathbb{X}$, $a \in \mathbb{A}$. Suppose that $\sum_j [\delta_{kj} - p(j|k, f(k))]\phi_j > 0$ for some $k \in \mathbb{X}_x$. Since $x_{kf(k)} > 0$, $\sum_j [\delta_{kj} - p(j|k, f(k))]\phi_j \cdot x_{kf(k)} > 0$. Furthermore, $\sum_j [\delta_{ij} - p(j|i,a)]\phi_j \cdot x_{ia} \geq 0$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, $\sum_{i,a} \sum_j [\delta_{ij} - p(j|i,a)]\phi_j \cdot x_{ia} > 0$.
On the other hand, this result is contradictory to the constraints of program (4.9) because $\sum_{i,a} \sum_j [\delta_{ij} - p(j|i,a)]\phi_j \cdot x_{ia} = \sum_j \{\sum_{i,a} [\delta_{ij} - p(j|i,a) x_{ia}]\}\phi_j = 0$. Therefore, we have shown that $\sum_j [\delta_{ij} - p(j|i, f(i))]\phi_j = 0$ for every $i \in \mathbb{X}$, i.e. $\phi = P(f)\phi$, and consequently $\phi = P^*(f)\phi$. Next, it can easily be shown that $\mathbb{X}_x$ is closed under $P(f)$. Then, we can prove that the states of $\mathbb{X} \backslash \mathbb{X}_x$ are transient in the Markov chain induced by $P(f)$. This implies that the columns of $\mathbb{X} \backslash \mathbb{X}_x$ in $P^*(f)$ are zero. Now, we can finish the proof as follows. For every $k \in \mathbb{X}$, we can write,
$\phi_k(f) = [P^*(f) r(f)]_k = \sum_i [P^*(f)]_{ki} r(i, f(i)) = \sum_{i \in \mathbb{X}_x} [P^*(f)]_{ki} r(i, f(i)) = \sum_{i \in \mathbb{X}_x} [P^*(f)]_{ki} \{\phi_i + \sum_j [\delta_{ij} - p(j|i, f(i))] u_j\} = [P^*(f)\{\phi + \{(I - P(f))u\}]_k = \phi_k$.
Hence, f is an average optimal policy. ∎

Algorithm VIII (linear programming; average reward, multichain case)

1. Take any β with $\beta_j > 0, j \in \mathbb{X}$, and compute extreme optimal solutions (u, v) and (x, y), e.g. by the simplex method, of the dual pair linear programs (4.8) and (4.9) respectively.
2. Choose f such that $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$ if $i \notin \mathbb{X}_x$. Then, f is an average optimal policy and v is the value vector ϕ.

In the average reward case there is no one-to-one correspondence between the feasible solutions of the dual program (4.9) and the stationary policies. However, there are interesting relations. For a feasible solution (x, y) of (4.9) we define a stationary policy $\pi(x, y)$ by

$$\pi_{ia}(x, y) := \begin{cases} x_{ia} / \sum_a x_{ia} & a \in \mathbb{A}(i), \ i \in \mathbb{X}_x \\ y_{ia} / \sum_a y_{ia} & a \in \mathbb{A}(i), \ i \notin \mathbb{X}_x \end{cases} \tag{4.11}$$

Conversely, consider a stationary policy π, and define $(x(\pi), y(\pi))$ by

$$\begin{cases} x_{ia}(\pi) & := \{\sum_k \beta_k [P^*(\pi)]_{ki}\} \cdot \pi_{ia} & a \in \mathbb{A}(i), \quad i \in \mathbb{X} \\ y_{ia}(\pi) & := \{\sum_k \beta_k [D(\pi)]_{ki} + \sum_k \gamma_k [P^*(\pi)]_{ki}\} \cdot \pi_{ia} & a \in \mathbb{A}(i), \quad i \in \mathbb{X} \end{cases} \tag{4.12}$$

with $\gamma_k := \max_{i \in \mathbb{X}_j} \{-\sum_k \beta_k [D(\pi)]_{ki} / \sum_k [P^*(\pi)]_{ki}\}$, $k \in \mathbb{X}_j$, where $\mathbb{X}_j$ is the j-th ergodic set of the transition matrix $P(\pi)$, and $\gamma_k := 0$ for k a transient state. Then, the following results can be derived (see Kallenberg [134]).

Theorem 2.25
(i) For any stationary policy π, $(x(\pi), y(\pi))$ is feasible for (4.9).
(ii) For any pure stationary policy f, $(x(f), y(f))$ is an extreme point of (4.9).
(iii) If π is an average optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of (4.9) and vice-versa.

Remarks
1. As mentioned before, for MDPs with constraints the linear programming approach is appropriate. For multichain constrained MDPs, there is no optimal stationary policy, in general. The variables x_{ia} of program (4.9) can, analogously to the discounted case, be interpreted as average state-action frequencies, but the analysis is much more complex. For the unichain case, this analysis can be found in Derman [58]; the multichain case is treated by Hordijk and Kallenberg [116]. An interpretation of the second type of variables, the variables y_{ia}, is not obvious. They are related to the deviation matrix (Kallenberg [134]) and can be interpreted as biased deviation measures (Altman and Spieksma [8]). Other contributions in this area, based on a sample path approach, are Ross [195] and Ross and Varadarajan [196]. Beutler and Ross [25] discuss the constrained MDP by a Lagrangean approach. In Altman and Shwartz [4] the sensitivity of constrained MDPs is investigated.
2. MDPs with multi-objectives can be treated as constrained MDPs. For this topic we refer to Hordijk and Kallenberg [116], and to Durinovics, Lee, Katchakis and Filar [65].
3. For a decision maker it can be unsatisfactory to consider only the expectation of the rewards. It may be preferable to consider also the variability. Papers on this subject are Sobel [228], [229], Kawai and Katoh [143], White [282], [283] and [286], Filar, Kallenberg and Lee [78], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Huang and Kallenberg [123].

Open problem
For MDPs with constraints, an interesting question is find the best policy in the class of stationary policies or in the class F of pure stationary policies. In the multichain case, no satisfactory algorithm is known for these problems. For the problem to find the best policy within the class of stationary policies, the natural candidate $\pi(x, y)$, with (x, y) the optimal solution of (4.9) with additional constraints, does not satisfy (see Kallenberg [134]).

B. The unichain case.
Since in the unichain case ϕ is a vector with identical components, the property average-superharmonic is equivalent to the existence of a scalar v and a vector u such that $v + u_i \geq r(i,a) + \sum_j p(j|i,a)u_j$ for every $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, the LP-problem for the smallest average-superharmonic vector becomes

$$\min \left\{ v \mid v + \sum\nolimits_j [\delta_{ij} - p(j|i,a)]u_j \geq r(i,a) \text{ for every } (i,a) \in \mathbb{X} \times \mathbb{A} \right\} \quad (4.13)$$

with dual program

$$\max\left\{\sum_{i,a} r(i,a)x_{ia} \;\middle|\; \begin{array}{lll} \sum_{i,a}[\delta_{ij}-p(j|i,a)]\,x_{ia} & = 0, & j \in \mathbb{X} \\ \sum_{i,a} x_{ia} & = 1 & \\ x_{ia} & \geq 0, & (i,a) \in \mathbb{X}\times\mathbb{A} \end{array}\right\} \tag{4.14}$$

Algorithm IX (linear programming; average reward, unichain case)

1. Compute extreme optimal solutions (u, v) and x of the dual pair LPs (4.13) and (4.14) respectively.
2. Choose f such that $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $f(i)$ arbitrary if $i \notin X_x$. Then, f is an average optimal policy and $v \cdot e$ is the value vector ϕ.

Remarks

1. In the irreducible case any feasible solution of (4.13) satisfies $\sum_a x_{ia} > 0$, $i \in \mathbb{X}$. Furthermore, 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 (4.13) onto the stationary policies with as inverse mapping $\pi \to x_{ia}(\pi)$, where $x_{ia}(\pi) = [p^*(\pi)]_i \cdot \pi_{ia}$ with $p^*(\pi)$ the equilibrium distribution. The set F of pure stationary policies corresponds to the set of extreme solutions of (4.13). In this case, similar to the discounted reward criterion, it can be shown that the linear programming method is equivalent to policy iteration. For the relation between the discounted linear program and the undiscounted linear program in the irreducible case, we refer also to Nazareth and Kulkarni [170].
2. In the unichain case, also a suboptimality test can be implemented, in the policy iteration method as well as in the linear programming method (cf. Hastings [96] and Lasserre [152]). Furthermore, in the unichain case, problems with constraints have a solution in the set of stationary policies: if (x, y) is the optimal solution of the LP-problem with constraints, then $\pi(x, y)$ with $\pi_{ia}(x, y) = x_{ia}/\sum_a x_{ia}$, $a \in \mathbb{A}(i)$, $i \in \mathbb{X}_x$ (and arbitrary decisions in $\mathbb{X}\backslash\mathbb{X}_x$) is a stationary optimal policy (see Derman [58]).

The pioneering work in solving MDPs by linear programming was made by Manne [164] and De Ghellinck [42], who considered the irreducible case. The first analysis in the general multichain case was described in Denardo and Fox [51] and in Denardo [47], who proposed a sequential procedure. Hordijk and Kallenberg [114] have shown that also in the multichain case one linear program suffices. Many results about the linear programming method can be found in the monograph Kallenberg [134].

2.4.5 *Value iteration*

It seems natural to investigate for value iteration the formula of the discounted rewards with discount factor $\alpha = 1$, i.e.

$$\begin{cases} v_i^{n+1} := \max_a\{r(i,a) + \sum_j p(j|i,a)v_j^n\}, & i \in \mathbb{X}, n \geq 0 \\ v_i^0 \text{ arbitrary}, & i \in \mathbb{X} \end{cases} \tag{4.15}$$

with corresponding policies $f_0, f_1, \ldots$ such that $v^{n+1} = r(f_n) + P(f_n)v^n$, $n \geq 0$.

This approach, however, causes difficulties: in general, there is no convergence of the sequence $\{v^n \mid n \geq 0\}$ nor of the sequence $\{v^n - v^{n-1} \mid n \geq 1\}$. Since v^n corresponds to the total reward during n periods, the sequence $\{v^n \mid n \geq 0\}$ is in general unbounded and grows linearly in n. Therefore, it is plausible to consider the sequence $\{v^n - n \cdot \phi \mid n \geq 0\}$. The next lemma, which appeared in Brown [29], shows that this sequence is bounded. The behavior of this sequence is also studied by Lanery [150].

Lemma 2.7 The sequence $\{v^n - n \cdot \phi \mid n \geq 0\}$ is bounded.

Corollary 2.8 $\phi = \lim_{n\to\infty} \frac{1}{n} v^n = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (v^k - v^{k-1})$.

Although Corollary 2.8 shows that ϕ can be approximated by the sequence $\{\frac{1}{n} v^n \mid n \geq 1\}$, this result does not provide sufficient information for the computation of an ϵ-optimal policy or an ϵ-approximation of ϕ. Therefore, we need stronger results, e.g. the convergence of the sequence $\{e^n\}_{n=0}^{\infty}$, where e^n is defined by $e^n := v^n - n \cdot \phi$. In general, however, this sequence may fail to converge if some of the transition matrices $P(f)$ are periodic. Fortunately, periodicity can be avoided by the following *data transformation*, proposed by Schweitzer [209]. Schweitzer and Federgruen [214] have given necessary and sufficient conditions which guarantee the convergence of the sequence $\{e^n\}_{n=0}^{\infty}$.

Consider for an arbitrary $\lambda \in (0, 1)$ the modified transition probabilities

$$p^{\lambda}(j|i,a) = \lambda \delta_{ij} + (1 - \lambda) p(j|i,a), i, j \in \mathbb{X} \text{ and } a \in \mathbb{A}(i) \tag{4.16}$$

Since $p^{\lambda}(i|i, f(i)) \geq \lambda > 0$, the transition matrix $P^{\lambda}(f)$ is aperiodic. Let $\phi^{\lambda}(f)$ be the average reward of policy f with respect to the transitions (4.16), then the next lemma shows that $\phi^{\lambda}(f) = \phi(f)$. Hence, we may assume that for every f the Markov chain with transition matrix $P(f)$ is aperiodic, in which case $P^*(f) = \lim_{n\to\infty} P^n(f)$.

Lemma 2.8 $\phi^{\lambda}(f) = \phi(f)$ for every $f \in F$.

To show that, under the aperiodicity assumption, the sequence $\{e^n\}_{n=0}^{\infty}$ is convergent, we need the following theorem.

Theorem 2.26 *Let* $b(i,a) = r(i,a) - \phi_i + \sum_j p(j|i,a) u_j - u_i, i \in \mathbb{X}, a \in \mathbb{A}_i$; $m_i = \liminf_{n\to\infty} c_i^n$, $i \in \mathbb{X}$; $M_i = \limsup_{n\to\infty} c_i^n$, $i \in \mathbb{X}$, *and* $\mathbb{A}_*(i) = \{a \in \mathbb{A}(i) | \phi_i = \sum_j p(j|i,a)\phi_j\}$, $i \in \mathbb{X}$. *Then,* $\max_{a\in\mathbb{A}_*(i)}\{b(i,a) + \sum_j p(j|i,a) m_j\} \leq m_i \leq M_i \leq \max_{a\in\mathbb{A}_*(i)}\{b(i,a) + \sum_j p(j|i,a) M_j\}$.

Theorem 2.27 *Assume that the aperiodicity assumption holds. Then, the sequence* $\{e^n \mid n \geq 0\}$ *is convergent.*

Lemma 2.9 *Assume that the sequence* $\{e^n \mid n \geq 0\}$ *converges. Then,*

(i) f_n is average optimal for n sufficiently large.
(ii) $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$.

A. The multichain case

Since, for large n, $\phi \approx v^{n+1} - v^n$, $\| v^{n+1} - v^n \|$ and span $(v^{n+1} - v^n)$ do not provide a valid stopping criterion. If ϕ is not a constant vector, no stopping criteria are available. Therefore, another approach is necessary. Schweitzer [210] employs a hierarchical decomposition of the MDP into a set of communicating MDPs. This decomposition was proposed by Bather [11]. Schweitzer and Federgruen [216] have shown that this decomposition is unique.

Open problem
Formulate a value iteration algorithm (without a hierarchical decomposition of the MDP and without chain analysis) for multichain undiscounted MDPs.

Fundamental research of value iteration for undiscounted multichain MDPs was made by Schweitzer and Federgruen. In Schweitzer and Federgruen [217] it is shown, without any assumptions about the periodicity or the chain structure, that if the sequence $\{v^n - n \cdot \phi\}_{n=0}^{\infty}$ is convergent, the convergence rate is geometric. This is surprising because the operator of the mapping (4.15) is, in general, not a contraction nor a J-step contraction with respect to any norm or the seminorm *span*. Conditions, other than aperiodicity, for the convergence of $\{v^n - n \cdot \phi\}_{n=0}^{\infty}$ are given by Schweitzer [207], Denardo [49] and Bather [10]. Surveys on value iteration for undiscounted multichain MDPs can be found in Schweitzer and Federgruen [214] and in Federgruen and Schweitzer [70] and [71].

B. The 'constant value vector' case

Assume that the value vector is constant, i.e. $\phi = \phi_0 \cdot e$, where $\phi_0 \in \mathbb{R}$. This assumption is more general than the unichain assumption. Furthermore, we assume aperiodicity, which implies (cf. Lemma 2.9) that $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$. We will formulate an algorithm to compute an ϵ-optimal policy.

Theorem 2.28 *Let* $l_n = \min_i(v_i^n - v_i^{n-1})$ *and* $u_n = \max_i(v_i^n - v_i^{n-1})$, $n \in \mathbb{N}$. *Then,*
(i) $l_n \uparrow \phi_0$ *and* $u_n \downarrow \phi_0$.
(ii) $l_n \cdot e \leq \phi(f_{n-1}) \leq \phi_0 \cdot e \leq u_n \cdot e$, $n \geq 1$.

By the results of Theorem 2.28 an algorithm can be formulated. Since v^n grows linearly in n, a direct application of (4.14) may cause numerical difficulties. Therefore, we use the following transformation, which yields the so-called *relative value algorithm*.
Let $w_i^n := v_i^n - v_N^n$, $i \in \mathbb{X}$, $n \geq 0$; $g^n := v_N^n - v_N^{n-1}$, $n \geq 1$. Then, one can easily show that $\{w^n\}_{n=0}^{\infty}$ and $\{g^n\}_{n=0}^{\infty}$ are bounded sequences. Furthermore, the next iterands can be computed by the formulae $g^{n+1} = \max_{a\in A(N)}\{r(N,a) + \sum_j p(j|N,a)w_j^n\}$, and $w_i^{n+1} = \max_{a\in A(i)}\{r(i,a) + \sum_j p(j|i,a)w_j^n\} - g^{n+1}$, $i \in \mathbb{X}$. For the bounds l_n and u_n, we have $l_n = \min_i(w_i^n - w_i^{n-1}) + g^n$ and $u_n = \max_i(w_i^n - w_i^{n-1}) + g^n$.

Algorithm X (relative value iteration; average reward; aperiodic; constant value vector)

1. Choose $\epsilon > 0$, and take $v \in \mathbb{R}^N$ arbitrarily.
2. Compute:
   a. $s(i,a) := r(i,a) + \sum_j p(j|i,a)v_j$, $(i,a) \in \mathbb{X} \times \mathbb{A}$;
   b. $g := \max_{a \in \mathbb{A}(N)} s(N,a)$;
   c. $w_i := \max_{a \in \mathbb{A}(i)} s(i,a) - g$, $i \in \mathbb{X}$ and take f such that $w = r(f) + P(f)v - g$;
   d. $u := \max_i(w_i - v_i)$; $l := \min_i(w_i - v_i)$.
3. If $u - l \leq \epsilon$: f is an ϵ-optimal policy and $(u+l)/2$ is a $\frac{1}{2}\epsilon$-approximation of the value ϕ_0 (Stop);

   Otherwise: $v := w$ and go to step 2.

One may ask whether exclusion of suboptimal actions can be implemented for the average reward criterion. Similar to the discounted rewards, it can be shown that an action $a \in \mathbb{A}(i)$ is *suboptimal* if

$$\phi + u_i > r(i,a) + \sum_j p(j|i,a)u_j, \tag{4.17}$$

where (ϕ, u) is a solution of the optimality equation of Theorem 2.19. Since such a solution is unknown in advance, in order to apply (4.17) in an algorithm, we need bounds for ϕ and u. Theorem 2.28 provides bounds for ϕ; however, bounds for u are unknown. One may well apply a suboptimality test in one iteration of formula (4.15). In fact, v^n is the total reward over a horizon of n stages. Hence, suboptimality tests for finite horizon models can be used (see Hastings [93], Hastings and Van Nunen [99], and Hübner [124]).

Bounds on the value vector as formulated in Theorem 2.28 can be found in Hastings [95], Odoni [172], Hordijk and Tijms [118], and Platzman [177]. Hordijk and Tijms [120] have proposed an approximation method with a sequence of discounted value iterations with discount factors tending to 1. Algorithm X is established by White [277]. Recently, a new value iteration algorithm was proposed by Bertsekas [21], under the assumption that all policies are unichain and that there exists a state that is recurrent under all policies. This method is inspired by a relation with an associated stochastic shortest path problem.

2.4.6 Modified policy iteration

As in the discounted reward case, modified policy iteration can be applied. However, we need an assumption: we assume that the value vector ϕ is a constant vector: $\phi = \phi_0 \cdot e$. Furthermore, we assume the *strong aperiodicity assumption*, i.e. for some $0 < \lambda < 1$, $p(i|i,a) \geq \lambda > 0$ for all $i \in \mathbb{X}$, $a \in \mathbb{A}(i)$. As shown in the previous section by Schweitzer's aperiodicity transformation (4.16), any MDP can be transformed to an equivalent MDP which has the strong aperiodicity property.

Let T and T_f be the operators as defined in (3.2) and (3.3), respectively, and with $\alpha = 1$.

Lemma 2.10 *Let $l_n := \min_i(Tx^n - x^n)_i$, $n \in \mathbb{N}$. Then, the sequence $\{l_n\}_{n=1}^{\infty}$ is monotonically nondecreasing.*

Remark
Let $u_n := \max_i(Tx^n - x^n)_i$, $n \in \mathbb{N}$. Then the sequence $\{u_n \mid n \in \mathbb{N}\}$ is in general not monotonically nonincreasing.

Theorem 2.29
(i) Both sequences $\{l_n \mid n \in \mathbb{N}\}$ and $\{u_n \mid n \in \mathbb{N}\}$ converge to the value ϕ_0.
(ii) The convergence of span $(Ux^n - x^n)$ to zero is geometrically fast.
(iii) Algorithm XI (see below) terminates with an ϵ-optimal policy f and $\frac{1}{2}[u+l]$ is an $\frac{1}{2}\epsilon$-approximation of ϕ_0.

Algorithm XI (modified policy iteration; average reward; aperiodic; constant value vector)

1. Choose $x \in \mathbb{R}^N$ and $\epsilon > 0$ arbitrarily.
2. a. Choose k with $1 \leq k \leq \infty$;
   b. Determine f such that $T_f x = Tx$;
   c. Let $l := \min_i(Tx - x)_i$ and $u := \max_i(Tx - x)_i$.
3. If $u - l \leq \epsilon$: f is an ϵ-optimal policy and $(u+l)/2$ is a $\frac{1}{2}\epsilon$-approximation of the value ϕ_0 (Stop);
   otherwise: $x := T_f^k x$ and go to step 2.

Remark
If $k = 1$ the method becomes the standard value iteration method (without White's relative values). We will also argue that, in the unichain case, policy iteration corresponds to $k = \infty$. By Theorem 2.15, we have $\phi(f) + [I - P(f)]u(f) = r(f)$, $T_{f_n}^k x^n = T_{f_n}^k[u(f_n)] + P^k(f_n)[x^n - u(f_n)] = u(f_n) + k \cdot \phi(f_n) + P^k(f_n)[x^n - u(f_{n+1})]$. If k tends to infinity, $P^k(f_n)$ converges to $P^*(f_n)$, a matrix with equal rows, i.e. $P^k(f_n)[x^n - u(f_n)]$ converges to a constant vector. Since, $\phi(f_n)$ is also a constant vector, the difference between $T_{f_n}^k x^n$ and $u(f_n)$ converges to a constant vector. In the policy iteration algorithm VII with best improving actions, a new policy corresponds to maximization of $r(i,a) + \sum_j p(j|i,a)u_j(f_n)$, which is the same as $T[L_{f_n}^k x^n]$. Hence, both methods are very similar.

The modified policy iteration method was first mentioned by Morton [169]. Van der Wal [245] and [246] has analyzed this method extensively under various chain structure assumptions (irreducible case, unichain case, communicating case and simply connected case).

2.5 MORE SENSITIVE OPTIMALITY CRITERIA

2.5.1 *Introduction*

In section 1.1.3 the concepts of n-discount optimality, n-average optimality and Blackwell optimality were introduced. For all these criteria optimal policies, which are pure and stationary, exist.

Theorem 2.30 *For $n = -1, 0, 1, \ldots$ the criteria n-discount optimal and n-average optimal are equivalent.*

Remarks
1. The criterion (-1)-discount optimality, and hence also (-1)-average optimality, is equivalent to average optimality.
2. The criteria 0-discount optimality and 0-average optimality are also called bias optimality. For more details about bias optimality we refer to the Chapter 3 of this book.

In Blackwell [27] the concept of bias optimality was introduced. Veinott [256] presented a policy iteration algorithm for finding a bias optimal policy. In Veinott [256] is also shown that an *average overtaking* pure and stationary policy is bias optimal, and conjectured that the reverse statement is also true. A policy R_* is average overtaking optimal if $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}[v^t(R_*) - v^t(R)] \geq 0$ for every policy R. In the class of stationary policies, the conjecture was proved by Denardo and Miller [52]. Lippman [154] showed the equivalence for general (possibly nonstationary) policies. Other contributions to the computation of a bias optimal policy are Denardo [47] and [49], Fox [80] and Kallenberg [133]. Denardo and Rothblum [54] have studied the stronger criterion of *overtaking optimality*. A policy R_* is overtaking optimal if $\liminf_{T\to\infty} \sum_{t=1}^{T}[v^t(R_*) - v^t(R)] \geq 0$ for every policy R. For this criterion the existence of an optimal policy is not guaranteed, in general, as already was shown in Brown [29]. Denardo and Rothblum [54] provided conditions under which a stationary overtaking optimal policy exists. The n-discount optimality criterion was proposed in Veinott [257]. Sladky [223] has introduced the concept of n-average optimality; furthermore, he showed the equivalence between this criterion and the n-discount optimality.

2.5.2 *n-Discount optimality and policy iteration*

In this section we present a policy iteration algorithm to compute a policy that lexicographically maximizes the vector $(u^{-1}(f), u^0(f), \ldots, u^n(f))$ over the set F. For $n = -1$ an average optimal policy and for $n = 0$ a bias optimal policy is obtained. Furthermore, for all $n \geq N - 1$ an n-discount optimal policy is Blackwell optimal.

Theorem 2.31 *The linear system*

$$\left\{ \begin{array}{ll} [I - P(f)]x^{-1} & = 0 \\ x^{-1} + [I - P(f)]x^0 & = r(f) \\ x^{k-1} + [I - P(f)]x^k & = 0 \;\; 1 \leq k \leq n+1; \;\; P^*(f)x^{n+1} = 0 \end{array} \right\}$$

has the unique solution $(u^{-1}(f), u^0(f), \ldots, u^{n+1}(f))$.

Algorithm XII (policy iteration; n-discount optimality)

1. Take an arbitrary policy f.
2. Determine $(u^{-1}(f), u^0(f), \ldots, u^{n+1}(f))$ as unique solution of the linear system

$$\left\{ \begin{array}{ll} [I - P(f)]x^{-1} & = 0 \\ x^{-1} + [I - P(f)]x^0 & = r(f) \\ x^{k-1} + [I - P(f)]x^k & = 0 \; 1 \leq k \leq n+1; \; P^*(f)x^{n+1} = 0 \end{array} \right\}$$

3.
   a. If $\max_{X \times A}[Pu^{-1}(f) - u^{-1}(f)] > 0$, then $\mathbb{A}^{(-1)} = \underset{X \times A}{\operatorname{argmax}}[Pu^{-1}(f) - u^{-1}(f)]$, choose g from $\mathbb{A}^{(-1)}$ and go to step 5.
   b. If $\max_{X \times A^{(-1)}}[r + Pu^0(f) - u^0(f) - u^{-1}(f)] > 0$, then $\mathbb{A}^{(0)} = \underset{X \times A^{(-1)}}{\operatorname{argmax}}[r + Pu^0(f) - u^0(f) - u^{-1}(f)]$, choose g from $\mathbb{A}^{(0)}$ and go to step 5.
   c. For $k = 0$ until n do:
   If $\max_{X \times A^{(k)}}[Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)] > 0$, then $\mathbb{A}^{(k+1)} = \underset{X \times A^{(k)}}{\operatorname{argmax}}[Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)]$, choose g from $\mathbb{A}^{(k+1)}$ and go to step 5.
4. f is n-discount optimal (Stop).
5. $f(i) := g(i)$, $i \in \mathbb{X}$, and go to step 2.

Remarks
1. Instead of $P^*(f)x^{n+1} = 0$, we can also consider $x^{n+1} + [I - P(f)]x^{n+2} = 0$, since multiplication with $P^*(f)$ gives $P^*(f)x^{n+1} = 0$.
2. For $n = -1$ the algorithm is equivalent to algorithm VI.

Theorem 2.32 *Let f and g be subsequent policies in algorithm XII, then $v^\alpha(g) > v^\alpha(f)$ for α sufficiently close to 1.*

Theorem 2.33 *Algorithm XII terminates in a finite number of iterations with an n-discount optimal policy.*

Finally, we mention that an n-discount optimal policy is a Blackwell optimal policy if $n \geq N - 1$.

Theorem 2.34 *If algorithm XII is used to determine an $n \geq (N-1)$-discount optimal policy f, then f is also a Blackwell optimal policy.*

The policy iteration method of this section was proposed in Veinott [257] and in Miller and Veinott [166]. They have also shown that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. In Veinott [258] refined results are given. In Federgruen and Schweitzer [73] a value iteration method is suggested for solving nested functional equations. These equations arise e.g. when more sensitive discount optimal policies are found. In particular, a method is given to find the optimal bias vector and a bias-optimal policy.

2.5.3 *Blackwell optimality and linear programming*

In this section we show how linear programming in the space of the rational functions can be developed to compute optimal policies over the entire range of the discount factor. Especially, a procedure is presented for the computation of a Blackwell optimal policy.

Let $\mathbf{R}$ be the ordered field of the real numbers with the usual ordering denoted by $>$. By $P(\mathbf{R})$ we denote the set of all polynomials with real coefficients:

$$\mathrm{P}(\mathbf{R}) = \{p(x) \mid p(x) = a_0 + a_1 x + \cdots + a_n x^n, a_i \in \mathbf{R}, 1 \leq i \leq n\}. \qquad (5.1)$$

By p_0 and p_1 we denote the polynomials $p_0(x) = 0$ and $p_1(x) = 1$ for every x. The field $F(\mathbf{R})$ of rational functions with real coefficients consists of the elements $\frac{p(x)}{q(x)}$, where p and q are from $P(\mathbf{R})$ and $q \neq p_0$. The polynomial $p(x)$ is considered as identical to the rational function $\frac{p(x)}{p_1(x)}$. Two rational functions $\frac{p}{q}$ and $\frac{r}{s}$ are considered identical, denoted $\frac{p}{q} =_\ell \frac{r}{s}$, if $p(x)s(x) = q(x)r(x)$. The operations $+$ and $\cdot$ are the natural addition and multiplication, i.e. $\frac{p(x)}{q(x)} + \frac{r(x)}{s(x)} =_\ell \frac{p(x)s(x)+r(x)q(x)}{q(x)s(x)}$ and $\frac{p(x)}{q(x)} \cdot \frac{r(x)}{s(x)} =_\ell \frac{p(x)r(x)}{q(x)s(x)}$. The polynomials p_0 and p_1 are the identities with respect to the operations addition and multiplication. A complete ordering in $F(\mathbf{R})$ is obtained by $\frac{p}{q} >_\ell p_0$ if and only if $d(p)d(q) > 0$, where $d(p)$ is the first nonzero coefficient of $p(x)$. If $\frac{p}{q} >_\ell p_0$, then the rational function $\frac{p}{q}$ is called positive. $\frac{p}{q} \geq_\ell p_0$ means that either $p =_\ell p_0$ or $\frac{p}{q} >_\ell p_0$. $F(\mathbf{R})$ is a non-Archimedean ordered field. The continuity of polynomials implies that the rational function $\frac{p}{q}$ is positive if and only if $\frac{p(x)}{q(x)} > 0$ for all x sufficiently close to 0. Hence, we obtain the following result.

Theorem 2.35 *The rational function $\frac{p}{q}$ is positive if and only if there exists an $x_0 > 0$ such that $\frac{p(x)}{q(x)} > 0$ for every $x \in (0, x_0]$.*

Instead of the discount factor α we can also use the interest rate ρ, where the relation between α and ρ is given by $\rho = (1-\alpha)/\alpha$. Notice that the total expected discounted reward $v^\rho(f)$ is the unique solution of the linear system $[(1+\rho)I - P(f)]x = (1+\rho)r(f)$. Solving this equation by Cramer's rule shows that $v_i^\rho(f)$, $i \in \mathbb{X}$, is an element of $F(\mathbf{R})$, say $\frac{p}{q}$, where the degree of the polynomials p and q is at most N. It is well known (Theorem 2.16) that the interval [0,1) of the discount factor can be divided into a finite number of intervals, say $[0 = \alpha_m, \alpha_{m-1}), \ldots, [\alpha_0, \alpha_{-1} = 1)$, in such a way that there exist

policies f_i, $0 \leq i \leq m$, where f_i is α-optimal for all $\alpha \in [\alpha_i, \alpha_{i-1})$. Hence, on any of these intervals the components of the value vector v^ρ are elements of $F(\mathbf{R})$.

Furthermore, the optimality equation (3.2) implies that $(1+\rho)v_i^\rho \geq (1+\rho)r(i,a) + \sum_j p(j|i,a)v_j^\rho$, $(i,a) \in \mathbb{X} \times \mathbb{A}$, $\rho > 0$. Therefore, in the ordered field $F(\mathbf{R})$, we have $(1+\rho)v_i^\rho \geq_\ell (1+\rho)r(i,a) + \sum_j p(j|i,a)v_j^\rho$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. In general, v_i^ρ is not an element of $F(\mathbb{R})$, but there are elements of $F(\mathbb{R})$ which coincide piecewise with v_i^ρ.

An N-vector $w(\rho)$ with components in $F(\mathbf{R})$ is called superharmonic if $(1+\rho)w_i(\rho) \geq_\ell (1+\rho)r(i,a) + \sum_j p(j|i,a)w_j(\rho)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$. Hence, v^ρ is superharmonic. The concept of superharmonicity is useful to derive linear programs for MDPs.

Lemma 2.11 v^ρ is the smallest superharmonic vector with components in $F(\mathbf{R})$, i.e. for any superharmonic vector $w(\rho)$, $w_i(\rho) \geq_\ell v_i^\rho$, $i \in \mathbb{X}$.

Lemma 2.11 implies that the value vector v^ρ on the interval $(0, \rho_0]$ can be found as optimal solution of the following linear program in $F(\mathbf{R})$:

$$\min\{\sum\nolimits_j w_j(\rho) \mid \sum\nolimits_j [(1+\rho)\delta_{ij} - p(j|i,a)]w_j(\rho) \geq_\ell (1+\rho)r(i,a), (i,a) \in \mathbb{X} \times \mathbb{A}\}. \tag{5.2}$$

Consider also the following linear program in $F(\mathbb{R})$, called the *dual program*:

$$\max\left\{\sum\nolimits_{i,a}(1+\rho)r(i,a)\cdot x_{ia}(\rho) \;\middle|\; \begin{array}{ll} \sum_{i,a}[(1+\rho)\ \delta_{ij} - p(j|i,a)]\cdot & \\ \qquad\qquad x_{ia}(\rho) =_\ell p_1, j \in \mathbb{X} & \\ x_{ia} \geq_\ell p_0, & (i,a) \in \mathbb{X} \times \mathbb{A} \end{array}\right\} \tag{5.3}$$

For a fixed real value of ρ, the linear programs (5.2) and (5.3) are the linear programs (3.12) and (3.13) respectively. Also from section 1.3.3 it is known that there is a one-to-one correspondence between the extreme points of (5.3) and the set F. As in the simplex method, we will rewrite the equalities of (5.3) such that in 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. During any iteration, the set of constraints is written in the special form

$$x_B = B^{-1}e - B^{-1}Ax_N \tag{5.4}$$

where e is the vector with the right-hand-side of (5.3) as components, i.e. p_1; x_B and x_N are the basis and nonbasis variables, B is the basic matrix and A consists of the remaining (nonbasis) columns of (5.3).

We solve the dual program (5.3) in such a way that the optimality of some basic solution, or equivalently some policy f, is shown on a certain interval for the value of ρ. This is possible, because for every fixed ρ in that interval the corresponding simplex tableau is an optimal one. At any iteration of the simplex tableau there is a feasible solution $x(\rho)$ of (5.3) and a corresponding

"trial solution" $w(\rho)$ of (5.2), i.e. $w(\rho)$ satisfies the complementary slackness conditions

$$x_{ia}(\rho) \cdot \left\{ \sum_j [(1+\rho)\delta_{ij} - p(j|i,a)]w_j(\rho) - (1+\rho)r(i,a) \right\} = 0, (i,a) \in \mathbb{X} \times \mathbb{A} \tag{5.5}$$

for all ρ in the interval that is considered. Since any basic solution corresponds to a policy f, in each state i there is exactly one action, namely $f(i)$, such that $x_{if(i)}(\rho) > 0$ for all ρ in the actual interval. Hence, by (5.5),

$$[(1+\rho)I - P(f)]w(\rho) = (1+\rho)r(f), \text{ i.e. } w(\rho) = v^\rho(f). \tag{5.6}$$

The organization of the special simplex method with elements that are rational functions is based on the following theorem.

Theorem 2.36
(i) The elements of the simplex tableau can be written as rational functions with a common denominator, which is the product of all previous pivot elements.
(ii) The numerator and denominator of the rational functions are polynomials with degree N at most, except for the reduced costs where the numerator may have degree $N+1$.
(iii) For ρ sufficiently large, the optimal solution $x(\rho)$ is given by the basic variables $x_{if(i)}(\rho)$, where $f(i)$ is such that $r(i,(f(i)) = \max_a r(i,a)$, $i \in \mathbb{X}$.
(iv) The pivot operations in the simplex tableau are as follows ($n(\rho)$ is the common denominator): (a) the numerator of the pivot becomes the next common denominator, and the last common denominator becomes the new numerator of the pivot; (b) the numerators of the other elements in the pivot row are unchanged and the numerators of the other elements in the pivot column are multiplied by -1; (c) for the other elements, say numerator $p(\rho)$, we replace $p(\rho)$ by $\frac{p(\rho)t(\rho) - r(\rho)s(\rho)}{n(\rho)}$, which is a polynomial where $t(\rho)$ is the numerator of the last pivot and $r(\rho)$ is the numerator of the pivot row which is in the same column as $p(\rho)$, and $s(\rho)$ is the numerator in the pivot column which is in the same row as $p(\rho)$.

Starting with the artificial variables $z_j(\rho)$, $j \in \mathbb{X}$ as basic variables, we can compute the optimal simplex tableau for $\rho = \infty$ by exchanging $x_{1f(1)}$ with z_1, $x_{2f(2)}$ with $z_2, \ldots, x_{Nf(N)}$ with z_N, where $f(i)$ is such that $r(i,(f(i)) = \max_a r(i,a)$, $1 \leq i \leq N$. This tableau is optimal for $\rho \geq \rho_1$, where ρ_1 is the smallest value such that the reduced costs are nonnegative. To compute ρ_1 we have to determine the zeroes of some polynomials. The column that determines ρ_1 becomes the next pivot column. After a pivot transformation the next tableau is optimal for $[\rho_2, \rho_1)$ for some ρ_2. In this way we continue until the last interval $[\rho_m = 0, \rho_{m-1})$.

If we are only interested in computing a Blackwell optimal policy, and not in the computation of the intervals with corresponding optimal policies, the method can be described as follows:

1. Start with any policy f and compute the corresponding tableau.

2. If every reduced cost is nonnegative with respect to the ordering in $F(\mathbb{R})$, i.e. the dominating coefficient of the numerator of any reduced cost is nonnegative, 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. Since in any transformation the value of the objective function increases, none of the basis can return and therefore the method is finite.
2. The complexity of one pivot transformation is of order $N^3[\sum_{i=1}^{N} \#\mathbb{A}(i)]$.

Hordijk, Dekker and Kallenberg [113] have developed the simplex method for rational functions for the computation of discounted optimal policies over the whole range of the discount factors, including the computation of a Blackwell optimal policy. Related works are Smallwood [224], Jeroslow [128] and Holzbaur [108], [109].

2.6 A VARIETY OF OTHER TOPICS

2.6.1 *Mean - variability trade-off*

The standard criteria for MDPs based on average or (sensitive) discounted rewards are not always satisfactory. In this section we consider another approach. This approach is especially suitable for a decision maker who prefers to use a criterion which also considers the *variability* induced by a given policy. How do we measure this variability? We want to have a variability measure which is sensible, mathematically tractable, and for which an optimality concept can be used. It turns out that optimality for all starting states simultaneously is too strong a requirement. Therefore, we consider the criterion for a fixed initial distribution β. Then, as mean of the rewards for a given policy R, we use

$$\phi(\beta, R) := \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\beta,R}[r(X_t, Y_t)] \tag{6.1}$$

If a policy R satisfies $\phi(\beta, R) = \sum_i \beta_i \phi_i$, where ϕ is the value vector, then R is called a β-average-optimal policy. There are several ways to define the variability $v(\beta, R)$, where β is the initial distribution and R the policy. We use the definition

$$v(\beta, R) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\beta,R}[r(X_t, Y_t) - \phi(\beta, R)]^2 \tag{6.2}$$

The quantities $\phi(\beta, R)$ and $v(\beta, R)$ can be expressed in the so-called *state-action frequencies*. For any policy R, any $T \in \mathbb{N}$, and any initial distribution β, we denote the *expected state-action frequencies* in the first T periods by the $x^T(R)$, i.e.

$$x^T_{ja}(R) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\beta,R}[X_t = j,\ Y_t = a], (i, a) \in \mathbb{X} \times \mathbb{A} \tag{6.3}$$

By $X(R)$ we denote the limit points of the vectors $\{x^T(R), T = 1, 2, \dots\}$, and by L, $L(M)$, $L(S)$ and $L(D)$ the elements of $X(R)$ corresponding to general, Markov, stationary and deterministic policies, respectively. For policies R with $\#X(R) = 1$, e.g. stationary policies, we denote the unique element of $X(R)$ by $x(R)$. For such policies the variability satisfies

$$v(\beta, R) = \sum_{j,a} x_{ja}(R)[r(j,a)]^2 - \left[\sum_{j,a} x_{ja}(R) r(j,a)\right]^2 \tag{6.4}$$

Let X be the projection on the x-space of the feasible solutions (x, y) of the linear program (4.9) , i.e.

$$X = \left\{ x \;\middle|\; \begin{array}{lll} \sum_{i,a}[\delta_{ij} - p(j|i,a)]x_{ia} & = 0, & j \in \mathbb{X} \\ \sum_a x_{ja} + \sum_{i,a}[\delta_{ij} - p(j|i,a)]y_{ia} & = \beta_j, & j \in \mathbb{X} \\ x_{ia}, y_{ia} \geq 0, (i,a) \in \mathbb{X} \times \mathbb{A} & & \end{array} \right\} \tag{6.5}$$

Theorem 2.37 $\overline{L(D)} = \overline{L(S)} = L(M) = L = X$, *where* $\overline{S}$ *is the closed convex hull of a set* S.

There are several sensible formulations for the mean-variability problem. We consider the following three formulations.

(1) Maximal mean-variability ratio with lower bound on the mean

$$\max\left\{ \frac{[\phi(\beta, R]^2}{v(\beta, R)} \;\middle|\; \phi(\beta, R) \geq l \right\} \tag{6.6}$$

(2) Minimal variability with lower bound on the mean

$$\min\{v(\beta, R) \mid \phi(\beta, R) \geq l\} \tag{6.7}$$

(3) Variability-penalized formulation

$$\max\{\phi(\beta, R) - \lambda \cdot v(\beta, R)\} \text{ for some penalty } \lambda > 0 \tag{6.8}$$

Using the state-action frequencies, the problems (6.6), (6.7) and (6.8) can be formulated as mathematical programs. These programs are special cases of the following unifying program

$$\max\left\{ \frac{\sum_{j,a} B_{ja} x_{ja}}{D(\sum_{j,a} R_{ja} x_{ja})} + C\left(\sum_{j,a} R_{ja} x_{ja}\right) \;\middle|\; \begin{array}{l} x \in X \\ l \leq \sum_{j,a} R_{ja} x_{ja} \leq u \end{array} \right\} \tag{6.9}$$

with (a) C is a convex function; (b) if D is not a constant, then: (i) D is positive, convex and nondecreasing; (ii) C is nondecreasing; (iii) $\sum_{j,a} B_{ja} x_{ja} \leq 0$ for every $x \in X$.

In order to solve (6.9), we consider a parametric version of (4.9) with $B_{ia} + \vartheta R_{ia}$ instead of $r(i,a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$, i.e.

$$\max\left\{\sum_{i,a}x_{ia}[B_{ia}+\vartheta R_{ia}] \;\middle|\; \begin{array}{ll} \sum_{i,a}[\delta_{ij}-p(j|i,a)]x_{ia} & =0, \quad j\in\mathbb{X} \\ \sum_{a}x_{ja}+\sum_{i,a}[\delta_{ij}-p(j|i,a)]y_{ia} & =\beta_j, \quad j\in\mathbb{X} \\ x_{ia},y_{ia}\geq 0,(i,a)\in\mathbb{X}\times\mathbb{A} & \end{array}\right\} \tag{6.10}$$

with $\vartheta \in (-\infty,+\infty)$ as the parameter. The optimal solution $x(\vartheta)$ is a piecewise constant function of ϑ with values being extreme points of X, and the optimal value is a piecewise linear, convex function of ϑ. Thus, there exist $\vartheta_0 \equiv -\infty < \vartheta_1 < \cdots < \vartheta_{m-1} < \vartheta_m \equiv +\infty$ such that $x(\vartheta) = x^n$ for $\vartheta \in [\vartheta_{n-1},\vartheta_n]$, $1 \leq n \leq m$, with x^n an extreme point of X.

Let $k+1$ and $j+1$ be respectively the smallest integers among $0,1,\ldots,m$ such that $\sum_{i,a}R_{ia}x_{ia}^{k+1} > u$ and $\sum_{i,a}R_{ia}x_{ia}^{j+1} \geq l$. Furthermore, let $\alpha \in (0,1]$ and $\beta \in [0,1)$ be such that $x^u = \alpha x^k + (1-\alpha)x^{k+1}$ and $x^l = \beta x^j + (1-\beta)x^{j+1}$ satisfy $\sum_{i,a}R_{ia}x_{ia}^u = u$ and $\sum_{i,a}R_{ia}x_{ia}^l = l$.
Let $G(x) = \sum_{j,a}B_{ja}x_{ja}$, $g(x) = \sum_{j,a}R_{ja}x_{ja}$ and $V(x) = \frac{G(x)}{D(g(x))} + C(g(x))$ for $x \in X$, and let $G_n = G(x^n)$, $g_n = g(x^n)$ and $V^n = V(x^n)$, $1 \leq n \leq m$. Furthermore, define $V_{\text{opt}} = \max\{\max_{j+1\leq n\leq k} V^n, V(x^l), V(x^u)\}$.

Theorem 2.38
(i) Program (6.9) is feasible if and only if $g(x^m) \geq l$ and $g(x^1) \leq u$.
(ii) If program (6.9) is feasible, then V_{opt} is the optimal value of (6.9), and the maximizing x is the optimal solution x_{opt}.

If $l = -\infty$ and $u = +\infty$, then $V_{\text{opt}} = V(x^n)$ for some extreme point x^n of X. Theorem 2.38 provides a way to find an optimal solution for program (6.9), but it does not provide a procedure to construct an optimal policy. The next two theorems show how an optimal policy can be obtained.

Theorem 2.39 *If (x,y) is an extreme optimal solution for (6.10) for all ϑ in an open interval, then any $f \in F$ with $x_{if(i)} > 0$ if $i \in \mathbb{X}_x$ and $y_{if(i)} > 0$ if $i \notin \mathbb{X}_x$ has a state-action frequency vector $x(f)$ satisfying $\sum_{i,a}B_{ia}x_{ia}(f) = \sum_{i,a}B_{ia}x_{ia}$, $\sum_{i,a}R_{ia}x_{ia}(f) = \sum_{i,a}R_{ia}x_{ia}$ and $V(x(f)) = V(x)$.*

Theorem 2.40 *If program (6.9) is feasible, then either $x_{\text{opt}} = x^n$ for some $j+1 \leq n \leq k$ and there exists an optimal deterministic policy, or $x_{\text{opt}} = x^l$ (or x^u) and an initial randomization of two deterministic policies is optimal. These policies can be determined analogously to the policy in Theorem 2.39.*

Corollary 2.9 *For an unconstrained problem, i.e. $l = -\infty$ and $u = +\infty$, there exists an optimal policy $f \in F$.*

Remarks
1. The discounted and the average-unichain case can be treated in the same way as above. In fact, these cases are more simple.
2. The optimal policy R_* is also *Pareto-optimal* with respect to the pair $(\phi(R), -V(R))$.

State-action frequencies for the unichain case are discussed in Derman [58]. The multichain case is analyzed in Kallenberg [134] and Hordijk en Kallenberg [116], who have shown Theorem 2.37. State-action frequencies play also an important role in multiple-objective MDP and for MDPs with additional constraints. Contributions in this area are made by Derman and Veinott [62], Thomas [240], Ross [195], Ross and Varadarajan [196], Altman and Shwartz [7] and by Liu and Ohno [156]. The formulations (6.6), (6.7) and (6.8) were proposed by Sobel [228], by Kawai [142] and by Filar, Kallenberg and Lee [78], respectively. Other contributions in this area are Kawai and Katoh [143], White [282], [283] and [286], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Sobel [229]. The unifying framework is proposed by Huang and Kallenberg [123], who also have shown the Theorems 2.38, 2.39 and 2.40. Another model with a different criterion is an MDP in which a (weighted) sum of a number of discounting rewards, each with a different discount factor, has to be maximized. This model is studied in Feinberg and Shwartz [77]: see Chapter 7.

2.6.2 *Optimal stopping*

Optimal stopping problems were introduced in section 1.1.4. In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If the stopping action is chosen in state i, then a final reward r_i is earned and the process terminates. If the second action is chosen in state i, then a cost c_i is incurred and the probability of being in state j at the next time point is p_{ij}. Therefore the MDP model is:
$\mathbb{X} = \{1, 2, \ldots, N\}$; $\mathbb{A}(i) = \{1, 2\}, i \in \mathbb{X}; r(i,1) = r_i, i \in \mathbb{X}$;
$r(i,2) = -c_i, i \in \mathbb{X}; p(j|i,1) = 0, i, j \in \mathbb{X}; p(j|i,2) = p_{ij}, i, j \in \mathbb{X}$.

We are interested in finding an optimal stopping rule, i.e. we consider only transient policies. A policy R is called *transient* if $\sum_{t=1}^{\infty} \mathbf{P}_{i,R}[X_t \in \mathbb{X}] < \infty$ for all $i \in \mathbb{X}$, i.e. for any starting state i the process terminates in a finite time with probability 1. As optimality criterion the total expected reward is considered i.e.

$$v_i(R) := \sum_{t=1}^{\infty} \sum_{j,a} \mathbf{P}_{i,R}[X_t = j, Y = a] \cdot r(j,a) \tag{6.11}$$

For the computation of an optimal transient policy, the usual properties of discounted MDPs hold (cf. Kallenberg [134] chapter 3). Let v be the value vector, i.e. $v = \sup\{v(R) \mid R \text{ is transient}\}$. Then, similar to the discounted reward criterion, it can be shown that v is the smallest superharmonic vector, i.e. the smallest vector that satisfies

$$\begin{cases} v_i \geq r_i & , i \in \mathbb{X} \\ v_i \geq -c_i + \sum_j p_{ij} v_j & , i \in \mathbb{X}. \end{cases} \tag{6.12}$$

Hence, the value vector is the unique solution of the linear program

$$\min\left\{\sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i & ,i \in \mathbb{X} \\ v_i \geq -c_i + \sum_j p_{ij} v_j & ,i \in \mathbb{X} \end{array}\right\} \tag{6.13}$$

As in the discounted case, an optimal policy can be obtained by the dual program. Therefore, the following algorithm can be used.

Algorithm XIII (optimal stopping; linear programming)

1. Determine an optimal solution (x, y) of the dual program

$$\max\left\{\sum_i r_i x_i - \sum_i c_i y_i \;\middle|\; \begin{array}{lll} x_j + y_j - \sum_i p_{ij} y_i & = 1, & j \in \mathbb{X} \\ x_j, y_j & \geq 0, & j \in \mathbb{X} \end{array}\right\}$$

2. Choose f such that $f(i) = \begin{cases} 1 \text{ if } x_j > 0 \\ 2 \text{ if } x_j = 0. \end{cases}$

Let $S := \{i \in \mathbb{X} \mid r_i \geq -c_i + \sum_j p_{ij} r_j\}$, i.e. S is the set of states in which immediate stopping is not worse than continuing for one period and than choose the stopping action. An optimal stopping problem is *monotone* if $p_{ij} = 0$ for all $i \in S$, $j \notin S$, i.e. S is closed under P.

Theorem 2.41 *In a monotone optimal stopping problem the policy f, where $f(i) = 1$ if and only if $i \in S$, is optimal.*

For monotone stopping problems it is sufficient to determine the set S. The determination of S has complexity of order $\mathcal{O}(N)$.

A classical paper on optimal stopping problems is Breiman [28]. Other papers in this area are Chen [34], Ross [197], Yasuda [291] and Sonin [231].

2.6.3 *Multi-armed bandit problems*

We have introduced this model in section 1.1.4. At each decision time point the decision maker has the option to work on exactly one project. Any project may be in a finite number of states, say project j in the set X_j, $1 \leq j \leq n$. Hence, the state space is the Cartesian product: $\mathbb{X} = X_1 \times X_2 \times \cdots \times X_n$. Each state has the same action set $\mathbb{A} = \{1, 2, \ldots, n\}$, where action a means that project a is chosen, $1 \leq a \leq n$. When project a is chosen, i.e. project a is the active project, the immediate reward and the transition probabilities only depend on project a and the state $i \in X_a$. Let $r(i, a)$ and $p(j|i, a)$, $j \in X_a$, denote these quantities. The states of the inactive projects are frozen. As utility function the discounted reward is used.

The one-armed bandit stopping problem

Consider the one-armed bandit stopping problem, i.e. in each state there are two actions: action 1 is the stopping action where we earn a final reward M and by action 2 the process continues with immediate reward r_i and transition

probabilities p_{ij}. Let $v^\alpha(M)$ be the value vector of this optimal stopping problem. In the previous section it was discussed how this vector $v^\alpha(M)$ and an optimal policy can be computed by the linear programming programs

$$\min\left\{\sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i + \alpha\sum_j p_{ij}v_j, & i \in \mathbb{X} \\ v_i \geq M, & i \in \mathbb{X} \end{array}\right\} \tag{6.14}$$

and its dual

$$\max\left\{\sum_i r_i x_i + M \cdot \sum_i y_i \;\middle|\; \begin{array}{ll} \sum_i(\delta_{ij} - \alpha p_{ij})\ x_i + y_j & = 1, j \in \mathbb{X} \\ x_i, y_i & \geq 0, i \in \mathbb{X} \end{array}\right\} \tag{6.15}$$

Lemma 2.12 *For all $i \in \mathbb{X}$, $v_i^\alpha(M) - M$ is a nonnegative continuous nonincreasing function in M.*

Let $M_i^\alpha = \min\{M \mid v_i^\alpha(M) = M\}$, $i \in \mathbb{X}$, called the *Gittins indices*.

Theorem 2.42 *The policy f, which chooses the stopping action in state i if and only if $M_i^\alpha \leq M$, is optimal.*

For $M = M_i^\alpha$ both actions (stop or continue) are optimal in state i. Hence, an interpretation of the Gittins index M_i^α is the value of M where both actions are simultaneously optimal, and therefore M_i^α is also called the *indifference value*.

Multi-armed bandits
Next, we assume that there are in each state $n+1$ actions: action a, $1 \leq a \leq n$, means continue with project a, and action 0 stops the process with a terminal reward M. Let $v^\alpha(M)$ be the value vector, f_M the optimal policy and $T(M)$ the stopping time, i.e. the expected time before the process terminates with the final reward M. Let $C = (1-\alpha)^{-1} \cdot \max_{i,a}|r(i,a)|$, then C is an upper bound of the total discounted rewards (without the terminal rewards). Hence, if $M \geq C$, then immediate stopping is optimal in all states. The following result is in some sense obvious: $v^\alpha(M)$ is nondecreasing in M and a small change in M will change the value (per unit change) with the discounted (unit) terminal reward $\alpha^{T(M)}$.

Lemma 2.13
(i) $v_i^\alpha(M)$ is a nondecreasing, convex function in M, for all $i \in \mathbb{X}$.
(ii) $\frac{\partial}{\partial M} v_i^\alpha(M) = \mathbb{E}_{i,f_M}[\alpha^{T(M)}]$, for all $i \in \mathbb{X}$.

The next theorem is the key theorem for the multi-armed bandit problem. It says that an optimal action in a state $i = (i_1, i_2, \ldots, i_n)$ is to choose that project which has, for the given state of the project, the smallest Gittins index. This is an interesting result. It is surprising that these indices depend only on the individual project and not on the other projects. Hence, they can be computed independently for each project. By this property, the dimensionality of the problem is considerably reduced.

Theorem 2.43 *In state $i = (i_1, i_2, \ldots, i_n)$ the optimal policy chooses action a, where a is such that $M^\alpha_{i_a} = \max_j M^\alpha_{i_j}$.*

Alternative interpretation of the Gittins index
Consider the one-armed bandit process with initial state i. If $M = M^\alpha_i$ the optimal policy is indifferent between stopping and continuing, so that for any stopping time T, $M^\alpha_i \geq \mathbb{E}$[discounted reward before T] + $M^\alpha_i \cdot \mathbb{E}[\alpha^T]$, with equality for the optimal policy. Hence,

$(1-\alpha)M^\alpha_i = \max_{T\geq 1} \mathbb{E}$[discounted reward before T]$/[\{1 - \mathbb{E}(\alpha^T)\}/(1-\alpha)]$
$= \max_{T\geq 1} \mathbb{E}$[discounted reward before T]$/\mathbb{E}$[discounted time before T],
where the expectations are conditional on the initial state i. Thus, another way to describe the optimal policy in the multi-armed bandit problem is as follows. For each individual project look for the stopping time T whose ratio of expected discounted reward and expected discounted time prior to T is maximal. Then work on the project with the largest ratio. In the case there also is the extra option of stopping, one should stop if all ratios are smaller than $(1-\alpha)$M.

Computation of the Gittins indices by parametric linear programming
We have already seen that for one project the Gittins index is related to the linear programs (6.14) and (6.15). For M big enough, an optimal solution (x, y) of (6.15) will satisfy $y_i > 0$, $i \in \mathbb{X}$. Decreasing M will give that some y_i becomes 0 for a certain value of M. For this M there is indifference between stopping and continuing, i.e. this M is the Gittins index in state i. By further decreasing M one can compute the next Gittins index, and so on. Hence, by parametric linear programming with parameter M which goes from $+\infty$ to $-\infty$, all Gittins indices can be computed for one project. The complexity of this approach is $\mathcal{O}(N^3)$.

Interpretation as restart-in-k problem
There is also another interpretation for the Gittins index M^α_k in a fixed state k. For any terminal value M, we have

$$v^\alpha_i(M) = \max\{M, r_i + \alpha \sum\nolimits_j p_{ij} v^\alpha_j(M)\}, i \in \mathbb{X} \tag{6.16}$$

and in state k, for $M = M^\alpha_k$,

$$v^\alpha_k(M) = M^\alpha_k = r_k + \alpha \sum\nolimits_j p_{kj} v^\alpha_j(M^\alpha_k) \tag{6.17}$$

Substituting (6.17) in (6.18) gives

$$v^\alpha_i(M^\alpha_k) = \max\left\{r_k + \alpha \sum\nolimits_j p_{kj} v^\alpha_j(M^\alpha_k),\ r_i + \alpha \sum\nolimits_j p_{ij} v^\alpha_j(M^\alpha_k)\right\}, i \in \mathbb{X} \tag{6.18}$$

Hence, M^α_k is the k-th component of the value vector of the MDP where there are in each state two actions. By the first action the process is restarted in state k, and the second action continues the process. Since M^α_k can be found as the k-th component of the value vector of the restart-in-k problem, it can be computed by the following linear program

$$\max\left\{\sum\nolimits_j v_j \;\middle|\; \begin{array}{l} \sum_j (\delta_{ij} - \alpha p_{ij}) v_j \geq r_i, \quad i \neq k \\ \sum_j (\delta_{ij} - \alpha p_{kj}) v_j \geq r_k, \ i \in \mathbb{X} \end{array}\right\} \tag{6.19}$$

For this restart-in-k problem, one can also characterize the states where it is optimal to choose action 'continue'.

Theorem 2.44 *Let $C_k = \{i \mid$ for the restart-in-k problem it is optimal to continue in state $i\}$. Then, $C_k = \{i \in \mathbb{X} \mid M_i^\alpha \geq M_k^\alpha\}$.*

Largest remaining index
In the largest remaining index approach the indices can be computed in a sequence, as in parametric linear programming, starting with the largest index.

Theorem 2.45 *Suppose that, for some k, $M_1^\alpha \geq M_2^\alpha \geq \cdots \geq M_k^\alpha$, and $M_k^\alpha \geq M_i^\alpha$ for all $i > k$. Let l_k be such that $M_{l_k}^\alpha = \max_{i>k} M_i^\alpha$ (the largest remaining index). Then, we have,*
$(1-\alpha)M_{l_k}^\alpha = \max_{i>k} \frac{[(I-\alpha P^k)^{-1}r]_i}{[(I-\alpha P^k)^{-1}e]_i}$, where $[P^k]_{ij} = \begin{cases} p_{ij}, & j \leq k \\ 0, & j > k. \end{cases}$

In order to find $M_{l_k}^\alpha$, we have to invert $[I-\alpha P^k]$. Since successive P^k matrices are similar, this can be done efficiently in a recursive way. The computations can be done in $\mathcal{O}(k^2)$. Hence, the overall complexity is $\sum_{k=1}^{N} \mathcal{O}(k^2) = \mathcal{O}(N^3)$.

The basic results on the multi-armed bandit problem are originated by Gittins (Gittins and Jones [86] and Gittins [85]). Other proofs of the optimality of the index rule can be found in Whittle [288] and [289], Ross [200], Tsitsiklis [242] and [243], Katehakis and Veinott [141], Weber [267] and Ishikida and Varaiya [127]. In honor of Gittins, Whittle has introduced the term Gittins indices. A first linear programming method of $\mathcal{O}(N^4)$ is proposed by Chen and Katehakis [35]. Kallenberg [135] has improved this method to $\mathcal{O}(N^3)$. The interpretation as restart-in-k problem is made by Katehakis and Veinott [141]. The method of the largest remaining index rule is due to Varaiya, Walrand and Buyokkoc [254]. A method based on bisection was proposed in Ben-Israel and Flåm [14]. Extension are made in various directions. Branching bandits were studied, e.g. by Weiss [270]; generalized bandits, e.g. in Glazebrook and Owen [89], and in Glazebrook and Greatrix [88]. Bertsimas and Niño-Mora [24] have proposed a new approach by generalizing the theory of extended polymatroids. Other papers based on this new approach are Glazebrook and Garbe [87], and Garbe and Glazebrook [84].

2.6.4 *Separable Markov decision problems*

Separable MDPs have the property that for certain pairs (i,a) of a state i and an action a: (i) the immediate reward is the sum of terms due to the current state and action, i.e. $r(i,a) = s(i)+l(a)$, (ii) the transition probabilities depend only on the action and not on the state from which the transition occurs, i.e. $p(j|i,a) = p(j|a)$. For separable problems an LP formulation can be given, which involves a smaller number of variables than in the general LP formulation. In this section we consider the multichain undiscounted case. For the discounted case and the unichain undiscounted case we refer to De Ghellinck and Eppen [43] and to Denardo [46], respectively.

A *separable Markov decision problem* has the following structure:
(1) In some states, say the states of $\mathbb{X}_1 = \{1, 2, \ldots, m\}$, there are subsets of the action sets, say subset $\mathbb{A}_1(i)$ in state $i \in \mathbb{X}_1$, such that:
(i) $r((i,a) = s(i) + t(a)$, $i \in \mathbb{X}_1$, $a \in \mathbb{A}_1(i)$;
(ii) $p(j|i,a)$ is independent of i : $p(j|i,a) = p(j|a)$, $i \in \mathbb{X}_1$, $a \in \mathbb{A}_1(i)$, $j \in \mathbb{X}$.
(2) The action subsets are nested: $\mathbb{A}_1(1) \supseteq \mathbb{A}_1(2) \supseteq \cdots \supseteq \mathbb{A}_1(m) \neq \emptyset$.
Let $\mathbb{X}_2 := \mathbb{X} \backslash \mathbb{X}_1$, $\mathbb{A}_2(i) := \mathbb{A}(i) \backslash \mathbb{A}_1(i)$, $1 \leq i \leq m$, $\mathbb{A}_2(i) := \mathbb{A}(i)$, $m+1 \leq i \leq N$, and $B(i) := \mathbb{A}_1(i) - \mathbb{A}_1(i+1)$, $1 \leq i \leq m-1$, $B(m) := \mathbb{A}_1(m)$. Then $\mathbb{A}_1(i) = \cup_{j=i}^{m} B(j)$, and the sets $B(j)$ are disjoint. $\mathbb{X}_1$, $\mathbb{X}_2$, $\mathbb{A}_2(i)$ or $B(i)$ may be empty.

If the system is observed in a state $i \in \mathbb{X}_1$, and the decision maker will choose an action from $\mathbb{A}_1(i)$, the decision process can be considered as follows. First a reward $s(i)$ is earned and the system makes a zero-time transition to an additional state $N+i$. In this state there are two options: either to take an action $a \in B(i)$ or to take an action from $\mathbb{A}_1(i+1)$. In the first case reward $t(a)$ is earned and the process moves to state j with probability $p(j|a)$, $j \in \mathbb{X}$; in the second case we have the same situation as in state $N+i+1$, i.e. a zero-time transition is made from state $N+i$ to state $N+i+1$. This formulation can be interpreted as a semi-Markov decision process (see section 1.6.5). It can be shown that a linear program, which directly provides an average optimal policy, can be formulated. This linear program is based on linear programming for semi-Markov decision problems (cf. Kallenberg [134], chapter 7).

Consider the linear program

$$\text{minimize} \sum_{j=1}^{N} g_j + \sum_{j=1}^{m} h_j \tag{6.20}$$

$$\sum_{j=1}^{N} [\delta_{ij} - p(j|i,a)] g_j \geq 0, 1 \leq i \leq N, a \in \mathbb{A}_2(i); g_i - h_i \geq 0, 1 \leq i \leq m$$

$$-\sum_{j=1}^{N} p(j|a) g_j + h_i \geq 0, 1 \leq i \leq m, a \in B(i); h_i - h_{i+1} \geq 0, 1 \leq i \leq m-1$$

$$g_i + \sum_{j=1}^{N} [\delta_{ij} - p(j|i,a)] u_j \geq r(i,a), 1 \leq i \leq N, a \in \mathbb{A}_2(i); u_i - v_i \geq s_i, 1 \leq i \leq m$$

$$h_i - \sum_{j=1}^{N} p(j|a) u_j + v_i \geq t(a), 1 \leq i \leq m, a \in B(i); v_i - v_{i+1} \geq 0, 1 \leq i \leq m-1$$

The corresponding dual program is (the dual variables corresponding to the constraints of (6.20) are y_{ia}, μ_i, z_{ia}, σ_i, x_{ia}, λ_i, w_{ia} and ρ_i, respectively:)

$$\text{maximize} \sum_{i=1}^{N} \sum_{a \in \mathbb{A}(i)} r(i,a) x_{ia} + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} t(a) w_{ia} \tag{6.21}$$

$$\sum_{i=1}^{N}\sum_{a\in A(i)}[\delta_{ij}-p(j|i,a)]y_{ia}+\sum_{i=1}^{m}\delta_{ij}\mu_i\sum_{i=1}^{m}\sum_{a\in B(i)}p(j|a)z_{ia}$$
$$+\sum_{a\in A(j)}x_{ja}=1,\quad 1\le j\le N$$
$$\sigma_j-\sigma_{j-1}+\sum_{a\in B(j)}w_{ja}-\mu_j+\sum_{a\in B(j)}z_{ja}=1,\quad 1\le j\le m$$
$$\sum_{i=1}^{N}\sum_{a\in A_2(i)}[\delta_{ij}-p(j|i,a)]x_{ia}+\sum_{i=1}^{m}\delta_{ij}\lambda_i$$
$$-\sum_{i=1}^{m}\sum_{a\in B(i)}p(j|a)w_{ia}=0,\quad 1\le j\le N$$
$$\rho_j-\rho_{j-1}+\sum_{a\in B(j)}w_{ja}-\lambda_j=0, 1\le j\le m;\quad \rho_0=\rho_m=\sigma_0=\sigma_m=0;$$
$$x_{ia},y_{ia},z_{ia},w_{ia},\lambda_i,\mu_i,\rho_i,\sigma_i\ge 0 \text{ for all } i \text{ and } a.$$

A proof for the next result can be found in Kallenberg [136].

Theorem 2.46
(i) The linear programs (6.20) and (6.21) have finite optimal solutions.
(ii) If (g,h,u,v) is an optimal solution of program (6.20), then g is the value vector.
(iii) Let $(y,\mu,z,\sigma,x,\lambda,w,\rho)$ be an extreme optimal solution of program (6.21). Define m_i and n_i by $m_i=\min\{j\ge i \mid \sum_a w_{ja}>0\}$ and $n_i=\min\{j\ge i \mid \sum_a(w_{ja}+z_{ja})>0\}$, $i\in\mathbb{X}$. Take a policy f such that in state i : $x_{if(i)}>0$ if $\sum_a x_{ia}>0$; $w_{m_i f(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i>0$; $y_{if(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}>0$; $w_{n_i f(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}=0\wedge\sum_a w_{n_i a}>0$; $z_{n_i f(i)}>0$ if $\sum_a x_{ia}=0\wedge\lambda_i=0\wedge\sum_a y_{ia}=0\wedge\sum_a w_{n_i a}=0$. Then, f is well defined and an average optimal policy.

There are many applications which can be formulated as separable MDPs. We mention some of them.

Replacement problem (cf. Howard's [121] automobile problem).
The decision maker has two options in each state i: either to continue or to replace the item by another of a certain state $j\in\{1,2,\dots,N\}$. The linear program to solve this problem as 'normal' Markov decision problem contains $2N(N+1)$ variables and $2N$ constraints. The reduced linear programming formulation has only $6N$ variables and $2N+1$ constraints.

Inventory problem
Consider the following inventory model. At the end of each period, the amount i of inventory is observed, where $0\le i\le N$. The possible actions are: either to order nothing or to order $a-i$ items, where $i+1\le a\le N$, with fixed ordering costs K and cost c for each ordered item. We assume that the delivery is instantaneous and that there is no backlogging. The linear program to solve this problem as 'normal' Markov decision problem has $(N+1)(N+2)$ variables and $2(N+1)$ constraints. In the reduced formulation as separable problem, we have $8N-2$ variables and $2(2N+1)$ constraints. In the case that the optimal policy is an (s,S)-policy the underlying Markov chain is unichained. Then a linear program with $3(N-1)$ variables and $N+2$ constraints suffices.

Totally separable problem
Suppose that the Markov decision problem has the following structure:
$\mathbb{X} = \{1, 2, \ldots, N\}$; $\mathbb{A}(i) = \{1, 2, \ldots, M\}$, $i \in \mathbb{X}$; $r(i,a) = s(i) + t(a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$; $p(j|i,a) = p(j|a)$, $(i,a) \in \mathbb{X} \times \mathbb{A}$ and $j \in \mathbb{X}$.
Examples of this model can be found in Sobel [227]. Without exploiting the structure, the linear program has $2NM$ variables and $2N$ constraints. It can be shown that an optimal myopic solution exists, i.e. the action a_* is optimal in state i, where a_* is determined by:

$$t(a_*) + \sum_{j=1}^{N} p(j|a) s_j = \max_{1 \le a \le M} \left\{ t(a) + \sum_{j=1}^{N} p(j|a) s_j \right\} \tag{6.22}$$

This result is a special case of the stochastic game studied in Sobel [227] and Parthasarathy, Tijs and Vrieze [176].

2.6.5 *Further subjects*

In this chapter some of the main topics of finite MDPs are discussed. In this section we shortly mention some other aspects of MDPs without going into detail.

Semi-Markov decision models
In many applications the times between consecutive decision time points are not identical but random. Such processes are called semi-Markov decision processes if the time until the next decision depends only on the present state i and the action a chosen in state i. We assume that the distribution function $F_{ij}^{a}(t)$ for the random variable $\tau_{ij}(a)$, which is the sojourn time until the next decision point if decision a is chosen when the system is in state i and the transition is into state j, is known for all $i, j \in E$ and $a \in A(i)$.

Semi-Markov decision models are also called *Markov renewal programs*. The essential results of MDPs can be generalized to semi-MDPs. The semi-MDP model was introduced by Jewell [129], [130], Howard [122], De Cani [41] and Schweitzer [206]. Contributions for discounted rewards are e.g. Denardo [45], De Ghellinck and Eppen [43], Kallenberg [134], Wessels and Van Nunen [271], Ohno [174] and Schweitzer [213].

In the average reward case, there is a very elegant data transformation, proposed by Schweitzer [209], which converts a semi-MDP into an equivalent MDP. Let $\tau_i(a)$ be the expected time until the next decision epoch if action a is chosen when the system is in state i. For $0 < \tau \le \min_{i,a} \tau(i,a)$, let

$$\begin{cases} \overline{r}(i,a) = r(i,a)/\tau(i,a) & ,i \in \mathbb{X}, a \in \mathbb{A}(i) \\ \overline{p}(j|i,a) = \delta_{ij} - [\delta_{ij} - p(j|i,a)] \cdot \tau/\tau(i,a) & ,i,j \in \mathbb{X}, a \in \mathbb{A}(i) \end{cases} \tag{6.23}$$

Then, $\phi(\pi) = \overline{\phi}(\pi)$, where $\phi(\pi)$ is the average reward per unit time of the semi-MDP and $\overline{\phi}(\pi)$ the average reward of the discrete-time MDP with rewards $\overline{r}(i,a)$ and transition probabilities $\overline{p}(j|i,a)$ as defined in (6.23).

Other papers on average reward MDPs are Schweitzer and Federgruen [216], Federgruen and Spreen [75], Denardo and Fox [51], Osaki and Mine [175],

Kallenberg [134], Schweitzer and Federgruen [217], Schweitzer [210] and Denardo [48].

MDPs with partial information, partial observation and adaptive control
In an MDP with *partial information* the exact state of the process cannot be observed at decision epochs. The only information available about the state is a subset of the state space to which the state belongs. Formally, an MDP has partial information if the state space is partitioned into subsets $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_m$ such that at each decision epoch the only available information is the subset $\mathbb{X}_k$ to which the state belongs. In the partial information case not all decision rules are feasible: in all states of a subset $\mathbb{X}_k$ $(1 \leq k \leq m)$ the same decision has to be chosen. Such decision rule is called an *admissible* decision rule. The objective is to find an optimal admissible policy for some optimality criterion with respect to a given initial distribution. Papers on MDPs with partial information are e.g. Smallwood and Sondik [225], Hastings and Sadjani [98], Hordijk and Loeve [117], and Loeve [158].

A related model is an MDP with *partial observation*. In this model there is probabilistic information about the state. Using Bayes' rules this model can be translated in a model with full information but with a continuous state space, which incorporates the complete history of the process. Papers in this area are Sondik [230], Albright [1], Monahan [168], Altman and Shwartz [5] and [6], Lovejoy [159], [160] and [161], Rieder [192], Sernik and Markus [219], White III [273] and [274], White III and Scherer [275] and [276], and White [287].

In *adaptive control* models the transition probabilities $p(j|i,a)$ and the rewards $r(i,a)$ depend on an unknown parameter ϑ from a parameter space Θ. About these parameters increasing information is obtained when observing the ongoing process. At each decision epoch the decision maker must estimate the true parameter and then adapt the policy to the estimated value. Further literature about this topic is e.g. Kurano [147], Hübner [125], Hernandez-Lerma [102], Cavazos-Cadena [32] and Burnetas and Katehakis [31].

Vector-valued MDPs
In vector-valued MDPs, when the system is in state i and action a is chosen, there is not a single reward $r(i,a)$, but a vector $r^k(i,a)$, $1 \leq k \leq m$, of rewards. For this model, the concept of optimality is not unambiguous. Given an initial distribution β, a policy R and a utility function u (e.g. discounted or average expected reward), there is an m-vector $u(\beta,R)$ of returns, where the k-th component $u_k(\beta,R)$ corresponds to the rewards $r^k(i,a)$. Optimality is defined with respect to a cone $C \subseteq \mathbb{R}^m$. Such cone defines a partial ordering in $\mathbb{R}^m$: $x \geq_C y$ iff $x - y \in C$. A policy R^* is optimal if $u(\beta,R^*) \geq_C u(\beta,R)$ for all policies R. In general, there does not exist an optimal policy. Therefore, we use the concept of an *efficient policy*. A policy R^* is efficient if there is no 'better' policy, i.e. there is no policy R with $u(\beta,R) > u(\beta,R^*)$. If the cone $C = \mathbb{R}^m_+$, then efficient policies are also called Pareto-optimal policies. For vector-valued MDPs also the term *multi-objective MDPs* is used.

Papers about vector-valued MDPs are e.g. Furakawa [82], White [279], Henig [101], Kallenberg [134], Durinovic, Lee, Katehakis and Filar [65], Ghosh [90], Liu, Ohno and Nakayama [157], and Wakuta [261], [262], [263] and [264].

Acknowledgment

I am grateful to Arie Hordijk for introducing me in the interesting subject of MDPs as well as for the cooperation during a long period.

References

[1] S.C. Albright, [1979]: "Structural results for partially observable Markov decision processes", Operations Research *27*, 1041–1053.

[2] E. Altman, [1999]: "Constrained Markov decision processes", Chapman & Hall/CRC, Boca Raton, Florida.

[3] E. Altman, A. Hordijk and L.C.M. Kallenberg [1996]: "On the value function in constrained control of Markov chains", Mathematical Methods of Operations Research *44*, 387–399.

[4] E. Altman and A. Shwartz [1991a]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1–22.

[5] E. Altman and A. Shwartz [1991b]: "Adaptive control of constrained Markov chains", IEEE-Transactions on Automatic Control *36*, 454–462.

[6] E. Altman and A. Shwartz [1991c]: "Adaptive control of constrained Markov decision chains: criteria and policies", Annals of Operations Research *28*, 101–134.

[7] E. Altman and A. Shwartz [1991]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1–22.

[8] E. Altman and F.M. Spieksma [1995]: "The linear program approach in Markov decision processes", Mathematical Methods of Operations Research *42*, 169–188.

[9] J.S. Baras, D.J. Ma and A.M. Makowsky [1985]: "K competing queues with linear costs and geometric service requirements: the μc-rule is always optimal" Systems Control Letters *6*, 173–180.

[10] J. Bather [1973a]: "Optimal decision procedures for finite Markov chains. Part II: Communicating systems", Advances in Applied Probability *5*, 521–540.

[11] J. Bather [1973b]: "Optimal decision procedures for finite Markov chains. Part III: General convex systems", Advances in Applied Probability *5*, 541–553.

[12] M. Bayal-Gursoy and K.W. Ross [1992]: "Variability-sensitivity Markov decision processes", Mathematics of Operations Research *17*, 558–571.

[13] R. Bellman [1957]: "Dynamic programming", Princeton University Press, Princeton.

[14] A. Ben-Israel and S.D. Flam [1990]: "A bisection/successive approximation method for computing Gittins indices", Zeitschrift für Operations Research *34*, 411–422.

[15] D.P. Bertsekas [1976]: "Dynamic programming and stochastic control", Academic Press, New York.

[16] D.P. Bertsekas [1976b]: "On error bounds for successive approximation methods", IEEE Transactions on Automatic Control *21*, 394–396.

[17] D.P. Bertsekas [1987]: "Dynamic programming: deterministic and stochastic models", Prentice-Hall, Englewood Cliff.

[18] D.P. Bertsekas [1995]: "Dynamic programming and optimal control I", Athena Scientific, Belmont, Massachusetts.

[19] D.P. Bertsekas [1995]: "Dynamic programming and optimal control II", Athena Scientific, Belmont, Massachusetts+.

[20] D.P. Bertsekas [1995c]: "Generic rank-one corrections for value iteration in Markovian decision problems", OR Letters *17*, 111–119.

[21] D.P. Bertsekas [1998]: "A new value iteration method for the average cost dynamic programming problem", SIAM Journal on Control and Optimization *36*, 742–759.

[22] D.P. Bertsekas and S.E. Shreve [1978] "Stochastic Optimal Control", Academic Press, New York.

[23] D.P. Bertsekas and J.N. Tsitsiklis [1991]: "An analysis of stochastic shortest path problems", Mathematics of Operations Research *16*, 580–595.

[24] D. Bertsimas and J. Niño-Mora [1996]: "Conservations laws, extended polymatroids and multi-armed bandit problems; a polyhedral approach to indexable systems", Mathematics of Operations Research *21*, 257–306.

[25] F.J. Beutler and K.W. Ross [1985]: "Optimal policies for controlled Markov chains with a constraint", Journal of Mathematical Analysis and Applications *112*, 236–252.

[26] K.-J. Bierth [1987]: "An expected average reward criterion", Stochastic Processes and Applications *26*, 133–140.

[27] D. Blackwell [1962]: "Discrete dynamic programming", Annals of Mathematical Statistics, 719–726.

[28] L. Breiman [1964]: "Stopping-rule problems", in: E.F. Beckenbach (ed.), Applied Combinatorial Mathematics", Wiley, New York, 284–319.

[29] B.W. Brown [1965]: "On the iterative method of dynamic programming on a finite space discrete time Markov process", Annals of Mathematical Statistics *36*, 1279–1285.

[30] J. Bruno, P. Downey and G.N. Frederickson [1981]: "Sequencing tasks with exponential service times to minimize the expected flowtime or makespan", Journal of the Association for Computing Machinery *28*, 100–113.

[31] A.N. Burnetas, and M.N. Katehakis [1997]: "Optimal adaptive policies for Markov decision processes", Mathematics of Op. Research *22*, 222–255.

[32] R. Cavazos-Cadena [1991]: "Nonparametric estimation and adaptive control in a class of finite Markov decision chains", Annals of Operations Research *28*, 169–184.

[33] C.-S. Chang, A. Hordijk, R. Righter and G. Weiss [1994]: "The stochastic optimality of SEPT in parallel machine scheduling", Probability in the Engineering and Information Sciences *8*, 179–188.

[34] M.C. Chen, Jr. [1973]: "Optimal stopping in a discrete search problem", Operations Research *21*, 741–747.

[35] Y.-R. Chen and M.N. Katehakis [1986] : "Linear programming for finite state bandit problems", Mathematics of Operations Research *11*, 180–183.

[36] Y.S. Chow and H. Robbins [1961]: "A martingale system theorem and applications" in: J. Neyman (ed), "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability", Vol.1, University of Berkeley Press, Berkeley, 93–104.

[37] K.-J. Chung [1989]: "A note on maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *8*, 201–204.

[38] K.-J. Chung [1992]: "Remarks on maximal mean/standard deviation ratio in an undiscounted MDPs", Opimization *26*, 385–392.

[39] K.-J. Chung [1994]: "Mean-variance trade-offs in an undiscounted MDP: the unichain case", Operations Research *42*, 184–188.

[40] G.B. Dantzig [1963]: "Linear programming and extensions", Princeton University Press, Princeton, New Jersey.

[41] J.S. De Cani [1964]: "A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity", Management Science *10*, 716–733.

[42] G.T. De Ghellinck [1960]: "Les problèmes de décisions séquentielles", Cahiers du Centre de Recherche Opérationelle, 161–179.

[43] G.T. De Ghellinck and G.D. Eppen [1967]: "Linear programming solutions for separable Markovian decision problems", Management Science *13*, 371–394.

[44] R.S. Dembo and M. Haviv [1984]: "Truncated policy iteration methods", OR Letters *3*, 243–246.

[45] E.V. Denardo [1967]: "Contraction mappings in the theory underlying dynamic programming", SIAM Review *9*, 165–167.

[46] E.V. Denardo [1968]: "Separable Markovian decision problems", Management Science *14*, 451–462.

[47] E.V. Denardo [1970]: "Computing a bias-optimal policy in a discrete-time Markov decision problem", Operations Research *18*, 279–289.

[48] E.V. Denardo [1971]: "Markov renewal programs with small interest rates", Annals of Mathematical Statistics *42*, 477–496.

[49] E.V. Denardo [1973]: "A Markov decision problem", in: T.C. Hu and S.M. Robinson (eds.), "Mathematical Programming", Academic Press, 33–68.

[50] E.V. Denardo [1982]: "Dynamic programming: models and applications", Prentice-Hall, Englewood Cliff.

[51] E.V. Denardo and B.L. Fox [1968]: "Multichain Markov renewal programs", SIAM Journal on Applied Mathematics *16*, 468–487.

[52] E.V. Denardo and B.L. Miller [1968]: "An optimality condition for discrete dynamic programming with no discounting", Annals of Mathematical Statistics *39*, 1220–1227.

[53] E.V. Denardo and U.G. Rothblum [1979a]: "Optimal stopping, exponential utility and linear programming", Mathematical Programming *16*, 228–244.

[54] E.V. Denardo and U.G. Rothblum [1979b]: "Overtaking optimality for Markov decision chains", Mathematics of Operations Research *4*, 144–152.

[55] F. D'Epenoux [1960]: "Sur un problème de production et de stockage dans l'aléatoire", Revue Française de Recherche Opérationelle, 3–16.

[56] C. Derman [1962]: "On sequential decisions and Markov chains", Management Science *9*, 16–24.

[57] C. Derman [1963]: "Optimal replacement rules when changes of states are Markovian", in: R. Bellman (ed.), "Mathematical optimization techniques", The Rand Corporation, R-396-PR, 201–212.

[58] C. Derman [1970]: "Finite state Markovian decision processes", Academic Press, New York.

[59] C. Derman and M. Klein [1965]: "Some remarks on finite horizon Markovian decision models", Operations Research *13*, 272–278.

[60] C. Derman and J. Sacks [1960]: "Replacement of periodically inspected equipment (an optimal stopping rule)", Naval Research Logistics Quarterly *7*, 597–607.

[61] C. Derman and R. Strauch [1966]: "A note on memoryless rules for controlling sequential control problems", Annals of Mathematical Statistics *37*, 276–278.

[62] C. Derman and A.F. Veinott, Jr. [1972]: "Constrained Markov decision chains", Management Science *19*, 389–390.

[63] H.M. Dietz and V. Nollau [1983]: "Markov decision problems with countable state space", Akademie-Verlag, Berlin.

[64] L. Dubins and L.J. Savage [1965]: "How to gamble if you must", McGraw-Hill, New York.

[65] S. Durinovics, H.M. Lee, M.N. Katchakis and J.A. Filar [1986]: "Multiobjective Markov decision processes with average reward criterion", Large Scale Systems *10*, 215–226.

[66] E.B. Dynkin [1979]: "Controlled Markov process", Springer-Verlag, New York.

[67] J.H. Eaton and L.A. Zadeh [1962]: "Optimal pursuit strategies in discrete state probabilistic systems", Transactions ASME Series D, Journal of Basic Engineering *84*, 23–29.

[68] A. Ephremides, P. Varaiya and J. Walrand [1980]: "A simple dynamic routing problem", IEEE Transactions on Automatic Control AC-25, 690–693.

[69] A. Federgruen [1984]: "Markovian control problems: functional equations and algorithms", Mathematical Centre Tract 97, Mathematical Centre, Amsterdam.

[70] A. Federgruen and P.J. Schweitzer [1978]: "Discounted and undiscounted value iteration in Markov decision problems: a survey", in: M.L. Puterman (ed), "Dynamic programming and its applications", Academic Press, New York, 23–52.

[71] A. Federgruen and P.J. Schweitzer [1980]: "A survey of asymptotic value-iteration for undiscounted Markovian decision processes", in: R. Hartley, L.C. Thomas and D.J. White (eds.), "Recent development in Markov decision processes", Academic Press, New York, 73–109.

[72] A. Federgruen and P.J. Schweitzer [1984a]: "A fixed-point approach to undiscounted Markov renewal programs", SIAM Journal on Algebraic Discrete Methods *5*, 539–550.

[73] A. Federgruen and P.J. Schweitzer [1984b]: "Successive approximation methods for solving nested functional equations in Markov decision problems", Mathematics of Operations Research *9*, 319–344.

[74] A. Federgruen, P.J. Schweitzer and H.C. Tijms [1978]: "Contraction mappings underlying undiscounted Markov decision problems", Journal of Mathematical Analysis and Applications *65*, 711–730.

[75] A. Federgruen and D. Spreen [1980]: "A new specification of the multichain policy iteration algorithm in undiscounted Markov renewal programs", Management Science *26*, 1211–1217.

[76] A. Federgruen and P. Zipkin [1984]: "An efficient algorithm for computing optimal (s, S) policies", Operations Research *34*, 1268–1285.

[77] E.A. Feinberg and A. Shwartz [1994]: "Markov decision models with weighted discounted criteria", Mathematics of Operations Research *19*, 152–168.

[78] J.A. Filar, L.C.M. Kallenberg and H.M. Lee [1989]: "Variance-penalized Markov decision processes", Mathematics of Operations Research *14*, 147–161.

[79] J.A. Filar and O. J. Vrieze [1997]: "Competitive Markov decision processes", Springer-Verlag, New York.

[80] B.L. Fox [1968]: "(g, w)-optima in Markov renewal programs", Management Science *15*, 210–212.

[81] E. Frostig [1993]: "Optimal policies for machine repairmen problems", Journal of Applied Probability *30*, 703–715.

[82] N. Furakawa [1980]: "Characterization of optimal policies in vector-valued Markovian decision processes", Mathematics of Operations Research *5*, 271–279.

[83] S. Gal [1984]: "An $\mathcal{O}(N^3)$ algorithm for optimal replacement problems", SIAM Journal on Control and Optimization *22*, 902–910.

[84] R. Garbe and K.D. Glazebrook [1998]: "On a new approach to the analysis of complex multi-armed bandit problems", Mathematical Methods of Operations Research *48*, 419–442.

[85] J.C. Gittins [1979]: "Bandit processes and dynamic allocation indices", Journal of the Royal Statistic Society Series B *14*, 148–177.

[86] J.C. Gittins and D.M. Jones [1974]: "A dynamic allocation index for the sequential design of experiments", in J.Gani (ed.) "Progress in Statistics", North Holland, Amsterdam, 241–266.

[87] K.D. Glazebrook and R. Garbe [1996]: "Reflections on a new approach to Gittins indexation", Journal of the Operational Research Society *47*, 1301–1309.

[88] K.D. Glazebrook and S. Greatrix [1995]: "On transforming an index for generalized bandit problems", J. of App. Prob. *32*, 168–182.

[89] K.D. Glazebrook and R.W. Owen [1991]: "New results for generalized bandit problems", International Journal of System Sciences *22*, 479–494.

[90] M.K. Ghosh [1990]: "Markov decision processes with multiple costs", OR Letters *9*, 257–260.

[91] R. Grinold [1973]: "Elimination of suboptimal actions in Markov decision problems", Operations Research *21*, 848–851.

[92] R. Hartley, A.C. Lavercombe and L.C. Thomas [1986]: "Computational comparison of policy iteration algorithms for discounted Markov decision processes", Computers and Operations Research *13*, 411–420.

[93] N.A.J. Hastings [1968]: "Some notes on dynamic programming and replacement", Operational Research Quarterly *19*, 453–464.

[94] N.A.J. Hastings [1969]: "Optimization of discounted Markov decision problems", Operations Research Quarterly *20*, 499–500.

[95] N.A.J. Hastings [1971]: "Bounds on the gain of a Markov decision process", Operations Research *19*, 240–243.

[96] N.A.J. Hastings [1976]: "A test for nonoptimal actions in undiscounted finite Markov decision chains", Management Science *23*, 87–92.

[97] N.A.J. Hastings and J.M.C.Mello [1973]: "Tests for nonoptimal actions in discounted Markov decision problems", Management Science *19*, 1019–1022.

[98] N.A.J. Hastings and D.Sadjani [1979]: "Markov programming with policy constraints", European Journal of Operations Research *3*, 253–255.

[99] N.A.J. Hastings and J.A.E.E. Van Nunen [1977]: "The action elimination algorithm for Markov decision processes", in H.C. Tijms and J. Wessels (eds), "Markov decision theory", Mathematical Centre Tract 100, 161–170, Mathematical Centre, Amsterdam.

[100] M. Haviv and M.L. Puterman [1991]: "An improved algorithm for solving communicating average reward Markov decision processes", Annals of Operations Research *28*, 229–242.

[101] M.I. Henig [1983]: "Vector-valued dynamic programming", SIAM Journal on Control and Optimization *21*, 490–499.

[102] O. Hernández-Lerma [1987]: "Adaptive Markov control processes", Springer-Verlag, New York.

[103] O. Hernández-Lerma and J. B. Lasserre [1996]: "Discrete-time Markov control processes: Basic optimality criteria", Springer-Verlag, New York.

[104] O. Hernández-Lerma and J. B. Lasserre [1999]: "Further topics on discrete-time Markov control processes", Springer-Verlag, New York.

[105] M. Herzberg and U. Yechiali [1994]: "Accelerating procedures of the value iteration algorithm for discounted Markov decision processes, based on a one-step look-ahead analysis", Operations Research *42*, 940–946.

[106] D.P. Heyman and M. J. Sobel [1984]: "Stochastic models in Operations Research, Volume II, MacGraw-Hill, New York.

[107] K. Hinderer [1970]: "Foundations of non-stationary dynamic programming with discrete time parameter", Springer-Verlag, New York.

[108] U.D. Holzbaur [1986a]: "Entscheidungsmodelle über angeordneten Körpern", Optimization *17*, 515–524.

[109] U.D. Holzbaur [1986b]: "Sensitivitätsanalysen in Entscheidungsmodellen", Optimization *17*, 525–533.

[110] U.D. Holzbaur [1994]: "Bounds for the quality and the number of steps in Bellman's value iteration algorithm", OR Spektrum *15*, 231–234.

[111] A. Hordijk [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.

[112] A. Hordijk [1974]: "Dynamic programming and Markov potential theory", Mathematical Centre Tract 51, Amsterdam.

[113] A. Hordijk, R. Dekker and L.C.M. Kallenberg [1985]: "Sensitivity-analysis in discounted Markovian decision problems", OR Spektrum *7*, 143–151.

[114] A. Hordijk and L.C.M. Kallenberg [1979]: "Linear programming and Markov decision chains", Management Science *25*, 352–362.

[115] A. Hordijk and L.C.M. Kallenberg [1984a]: "Transient policies in discrete dynamic programming: linear programming including suboptimality tests and additional constraints", Mathematical Programming *30*, 46–70.

[116] A. Hordijk and L.C.M. Kallenberg [1984b]: "Constrained undiscounted stochastic dynamic programming", Mathematics of Operations Research *9*, 276–289.

[117] A. Hordijk and J.A. Loeve [1994]: "Undiscounted Markov decision chains with partial information; an algorithm for computing a locally optimal periodic policy", Mathematical Methods of Operations Research *40*, 163–181.

[118] A. Hordijk and H.C. Tijms [1974]: "The method of successive approximations and Markovian decision problems", Operations Research *22*, 519–521.

[119] A. Hordijk and H.C. Tijms [1975]: "A modified form of the iterative method of dynamic programming", Annals of Statictics *3*, 203–208.

[120] A. Hordijk and H.C. Tijms [1975]: "On a conjecture of Iglehart", Management Science *11*, 1342–1345.

[121] R.A. Howard [1960]: "Dynamic programming and Markov processes", MIT Press, Cambridge.

[122] R.A. Howard [1963]: "Semi-Markovian decision processes", Proceedings International Statistical Institute, Ottawa, Canada.

[123] Y. Huang and L.C.M. Kallenberg [1994]: "On finding optimal policies for Markov decision chains: a unifying framework for mean-variance tradeoffs", Mathematics of Operations Research *19*, 434–448.

[124] G. Hübner [1977]: "Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties", Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions, Reidel, Dordrecht, 257 263.

[125] G. Hübner [1988]: "A unified approach to adaptive control of average reward Markov decision processes", OR Spektrum *10*, 161–166.

[126] D. Iglehart [1963]: "Optimality of (s, S)-policies in the infinite horizon dynamic inventory problem", Management Science *9*, 259–267.

[127] T. Ishikida and P. Varaiya [1994]: "Multi-armed bandit problem revisited", Journal of Optimization Theory and Applications *83*, 113-154.

[128] R.G. Jeroslow [1972]: "An algorithm for discrete dynamic programming with interest rates near zero", Management Science Research Report no. 300, Carnegie-Mellon University, Pittsburgh.

[129] W.S. Jewell [1963a]: "Markov renewal programming. I: Formulation, finite return models", Operations Research *11*, 938 948.

[130] W.S. Jewell [1963b]: "Markov renewal programming. II: Infinite return models, example", Operations Research *11*, 949 971.

[131] L.C.M. Kallenberg [1981a]: "Finite horizon dynamic programming and linear programming", Methods of Operations Research *43*, 105 112.

[132] L.C.M. Kallenberg [1981b]: "Unconstrained and constrained dynamic programming over a finite horizon", Report, University of Leiden, The Netherlands.

[133] L.C.M. Kallenberg [1981c]: "Linear programming to compute a bias-optimal policy", in: B. Fleischmann et al. (eds.) "Operations Research Proceedings", 433–440.

[134] L.C.M. Kallenberg [1983]: "Linear programming and finite Markovian control problems", Mathematical Centre Tract 148, Mathematical Centre, Amsterdam.

[135] L.C.M. Kallenberg [1986]: "Note on M.N.Katehakis and Y.-R.Chen's computation of the Gittins index", Mathematics of Operations Research *11*, 184 186.

[136] L.C.M. Kallenberg [1992]: "Separable Markovian decision problem: the linear programming method in the multichain case", OR Spektrum *14*,43 52.

[137] L.C.M. Kallenberg [1999]: "Combinatorial problems in MDPs", Report, University of Leiden, The Netherlands (to appear in the Proceedings of the Changsha International Workshop on Markov Processes & Controlled Markov Chains).

[138] P.C. Kao [1973]: "Optimal replacement rules when the changes of states are semi-Markovian", Operations Research *21*, 1231–1249.

[139] M.N. Katehakis and C. Derman [1984]: "Optimal repair allocation in a series system", Mathematics of Operations Research *9*, 615–623.

[140] M.N. Katehakis and C. Derman [1989]: "On the maintenance of systems composed of highly reliable components", Management Science *35*, 551 560.

[141] M.N. Katehakis and A.F. Veinott, Jr. [1987]: "The multi-armed bandit problem: decomposition and computation", Mathematics of Operations Research *12*, 262–268.

[142] H. Kawai [1987]: "A variance minimization problem for a Markov decision process", European Journal of Operational Research *31*, 140–145.

[143] H. Kawai and N. Katoh [1987]: "Variance constrained Markov decision process", Journal of the Operations Research Society of Japan *30*, 88–100.

[144] J.G. Kemeny and J.L. Snell [1960]: "Finite Markov chains", Van Nostrand, Princeton.

[145] M. Klein [1962]: "Inspection-maintenance-replacement schedules under Markovian deterioration", Management Science *9*, 25 32.

[146] P. Kolesar [1966]: "Minimum-cost replacement under Markovian deterioration", Management Science *12*, 694 706.

[147] M. Kurano [1983]: "Adaptive policies in Markov decision processes with uncertain transition matrices", Journal of Information and Optimization Sciences *4*, 21–40.

[148] H. Kushner [1971]: "Introduction to stochastic control", Holt, Rineholt and Winston, New York.

[149] H. Kushner and A.J. Keinmann [1971]: "Accelerated procedures for the solution of discrete Markov control problems", IEEE Transactions on Automatic Control *16*, 147–152.

[150] E. Lanery [1967]: "Etude asymptotique des systèmes Markovien à commande", Revue d'Informatique et Recherche Operationelle *1*, 3–56.

[151] J.B. Lasserre [1994a]: "A new policy iteration scheme for Markov decision processes using Schweitzer's formula", Journal of Applied Probability *31*, 268–273.

[152] J.B. Lasserre [1994b]: "Detecting optimal and non-optimal actions in average-cost Markov decision processes", Journal of Applied Probability *31*, 979–990.

[153] W. Lin and P.R. Kumar [1984]: "Optimal control of a queueing system with two heterogeneous servers", IEEE Transactions on Automatic Control AC-29, 696 705.

[154] S.A. Lippman [1969]: "Criterion equivalence in discrete dynamic programming", Operations Research *17*, 920–923.

[155] J.Y. Liu and K. Liu [1994]: "An algorithm on the Gittins index", Systems Science and Mathematical Science 7, 106–114.

[156] Q.-S. Liu and K. Ohno [1992]: "Multiobjective undiscounted Markov renewal program and its application to a tool replacement problem in an FMS", Information and Decision Techniques *18*, 67–77.

[157] Q.-S. Liu, K. Ohno and H. Nakayama [1992]: "Multi-objective discounted Markov processes with expectation and variance criteria", International Journal of System Science *23*, 903–914.

[158] J.A. Loeve [1995]: "Markov decision chains with partial information", PhD dissertation, University of Leiden, The Netherlands.

[159] W.S. Lovejoy [1987]: "Some monotonicity results for partially observed Markov processes", Operations Research *35*, 736–743.

[160] W.S. Lovejoy [1991a]: "Computationally feasible bounds for partially observed Markov decision processes", Operations Research *39*, 162–175.

[161] W.S. Lovejoy [1991b] "A survey of algorithmic methods for partially observed Markov decision processes", Annals of Op. Research *28*, 47–66.

[162] J. Macqueen [1966]: "A modified programming method for Markovian decision problems", Journal of Mathematical Analysis and Applications *14*, 38–43.

[163] J. Macqueen [1967]: "A test for suboptimal actions in Markov decision problems", Operations Research *15*, 559–561.

[164] A.S. Manne [1960]: "Linear programming and sequential decisions", Management Science, 259–267.

[165] U. Meister and U. Holzbaur [1986]: "A polynomial time bound for Howard's policy improvement algorithm", OR Spektrum *8*, 37–40.

[166] B.L. Miller and A.F. Veinott Jr. [1969]: "Discrete dynamic programming with a small interest rate", Annals of Mathematical Statistics *40*, 366–370.

[167] H. Mine and S. Osaki [1970]: "Markov decision processes", American Elsevier, New York.

[168] G.E. Monahan [1982]: "A survey of partially observable Markov decision processes: theory, models and algorithms", Management Science *28*, 1–16.

[169] T. Morton [1971]: "Undiscounted Markov renewal programming via modified successive approximations", Operations Research *19*, 1081–1089.

[170] J.L. Nazareth and R.B. Kulkarni [1986]: "Linear programming formulations of Markov decision processes", OR Letters *5*, 13–16.

[171] M.K. Ng [1999]: "A note on policy iteration algorithms for discounted Markov decision problems", OR Letters *25*, 195–197.

[172] A. Odoni [1969]: "On finding the maximal gain for Markov decision processes", Operations Research *17*, 857–860.

[173] S. Oezekici [1988]: "Optimal periodic replacement of multicomponent reliability systems", Operations Research *36*, 542–552.

[174] K. Ohno [1981]: "A unified approach to algorithms with a suboptimality test in discounted semi-Markov decision processes", Journal of the Operations Research Society of Japan *24*, 296–323.

[175] S. Osaki and H. Mine [1968]: "Linear programming algorithms for semi-Markovian decision processes", Journal of Mathematical Analysis and Applications *22*, 356–381.

[176] T. Parthasarathy, S.H. Tijs and O.J. Vrieze [1984], "Stochastic games with state independent transitions and reparable rewards in: G. Hammer and D. Pallaschke (eds.), Selected Topics in Operations Research and Mathematical Economics.

[177] L.K. Platzman [1977]: "Improved conditions for convergence in undiscounted Markov renewal programming", Op. Research *25*, 529 533.

[178] M.A. Pollatschek and B. Avi-Itzhak [1969]: "Algorithms for stochastic games with geometric interpretation", Management Science *15*, 399–415.

[179] E.L. Porteus [1971]: "Some bounds for discounted sequential decision processes", Management Science *18*, 7 11.

[180] E.L. Porteus [1975]: "Bounds and transformations for discounted finite Markov decision chains", Operations Research *23*, 761–784.

[181] E.L. Porteus [1980a]: "Improved iterative computation of the expected return in Markov and semi-Markov chains", Zeitschrift für Operations Research *24*, 155 170.

[182] E.L. Porteus [1980b]: "Overview of iterative methods for discounted finite Markov and semi-Markov chains", in: R. Hartley, L.C. Thomas and D.J. White (eds.), "Recent development in Markov decision processes", Academic Press, New York, 1–20.

[183] E.L. Porteus [1981]: "Computing the discounted return in Markov and semi- Markov chains", Naval Research Logistics Quarterly *28*, 567 577.

[184] E. L. Porteus and J.C. Totten [1978]: "Accelerated computation of the expected discounted return in a Markov chain", Operations Research *26*, 350 358.

[185] M.L. Puterman [1981]: "Computational methods for Markov decision methods", Proceedings of 1981 Joint Automatic Control Conference.

[186] M.L. Puterman [1994]: "Markov decision processes", Wiley, New York.

[187] M.L. Puterman and S.L. Brumelle [1979]: "On the convergence of policy iteration in stationary dynamic programming", Mathematics of Operations Research *4*, 60–69.

[188] M.L. Puterman and M.C. Shin [1978]: "Modified policy iteration algorithms for discounted Markov decision chains", Management Science *24*, 1127 1137.

[189] M.L. Puterman and M.C. Shin [1982]: "Action elimination procedures for modified policy iteration algorithms" Operations Research *30*, 301–318.

[190] D. Reetz [1973]: "Solution of a Markovian decision problem by successive overrelaxation", Zeitschrift für Operations Research *17*, 29–32.

[191] D. Reetz [1976]: "A decision exclusion algorithm for a class of Markovian decision processes" , Zeitschrift für Operations Research *20*, 125–131.

[192] U. Rieder [1991]: "Structural results for partially observed control problems", Zeitschrift für Operations Research *35*, 473–490.

[193] R. Righter [1994]: "Scheduling", in: M. Shaked and J.G. Shantikumar (eds.), "Stochastic orders and their applications", Academic Press, 381–432.

[194] M. Roosta [1982]: "Routing through a network with maximum reliability", Journal of Mathematical Analysis and Applications *88*, 341–347.

[195] K.W. Ross [1989]: "Randomized and past-dependent policies for Markov decision processes with multiple constraints", Operations Research *37*, 474–477.

[196] K.W. Ross and R. Varadarajan [1991]: "Multichain Markov decision processes with a sample path constraint: a decomposition approach", Mathematics of Operations Research *16*, 195–207.

[197] S.M. Ross [1969]: "A problem in optimal search and stop", Operations Research *17*, 984–992.

[198] S.M. Ross [1970]: "Applied probability models with optimization applications", Holden-Day, San Francisco.

[199] S.M. Ross [1974]: "Dynamic programming and gambling models", Advances in Applied Probability *6*, 593–606.

[200] S.M. Ross [1983]: "Introduction to stochastic dynamic programming", Academic Press, New York.

[201] U.G. Rothblum [1979]: "Iterated successive approximation for sequential decision processes", in J.W.B. van Overhagen and H.C. Tijms (eds.), "Stochastic control and optimization", Free University, Amsterdam, 30–32.

[202] H. Scarf [1960]: "The optimality of (s, S) polices in the dynamic inventory problem", Chapter 13 in: K.J. Arrow, S. Karlin and P. Suppes (eds.), "Mathematical methods in the social sciences", Stanford University Press, Stanford.

[203] H. Schellhaas [1974]: "Zur extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung", Zeitschrift für Operations Research *18*, 91–104.

[204] N. Schmitz [1985]: "How good is Howard's policy improvement algorithm?", Zeitschrift fur Operations Research *29*, 315–316.

[205] L. Schrage [1968]: "A proof of the optimality of the shortest remaining processing time discipline", Operations Research *16*, 687–690.

[206] P.J. Schweitzer [1965]: "Perturbation theory and Markovian decision processes", Ph.D. dissertation, M.I.T., Op. Research Center Report 15.

[207] P.J. Schweitzer [1968]: "Perturbation theory and finite Markov chains" Journal of Applied Probability *5*, 401–413.

[208] P.J. Schweitzer [1971a]: "Multiple policy improvements in undiscounted Markov renewal programming", Operations Research *19*, 784–793.

[209] P.J. Schweitzer [1971b]: "Iterative solution of the functional equations of undiscounted Markov renewal programming", Journal of Mathematical Analysis and Applications *34*, 495–501.

[210] P.J. Schweitzer [1984]: "A value-iteration scheme for undiscounted multichain Markov renewal programs", ZOR – Zeitschrift für Operations Research *28*, 143–152.

[211] P.J. Schweitzer [1985]: "The variational calculus and approximations in policy space for Markov decision processes", Journal of Mathematical Analysis and Applications *110*, 568–582.

[212] P.J. Schweitzer [1987]: "A Brouwer fixed-point mapping approach to communicating Markov decision processes", Journal of Mathematical Analysis and Applications *123*, 117–130.

[213] P.J. Schweitzer [1991]: "Block-scaling of value-iteration for discounted Markov renewal programming", Annals of Op. Research *29*, 603–630.

[214] P.J. Schweitzer and A. Federgruen [1977]: "The asymptotic behavior of value iteration in Markov decision problems", Mathematics of Operations Research *2*, 360–381.

[215] P.J. Schweitzer and A. Federgruen [1978a]: "Foolproof convergence in multichain policy iteration", Journal of Mathematical Analysis and Applications *64*, 360–368.

[216] P.J. Schweitzer and A. Federgruen [1978b]: "The functional equations of undiscounted Markov renewal programming", Mathematics of Operations Research *3*, 308–321.

[217] P.J. Schweitzer and A. Federgruen [1979]: "Geometric convergence of value iteration in multichain Markov decision problems", Advances of Applied Probability *11*, 188–217.

[218] L.I. Sennott [1999]: "Stochastic dynamic programming and the control of queueing systems", Wiley, New York.

[219] E.L. Sernik and S.I. Markus [1991]: "On the computation of the optimal cost function for discrete time Markov models with partial observations", Annals of Operations Research *29*, 471–512.

[220] J.F. Shapiro [1975]: "Brouwer's fixed point theorem and finite state space Markovian decision theory", Journal of Mathematical Analysis and Applications *49*, 710–712.

[221] L.S. Shapley [1953]: "Stochastic games", Proceedings of the National Academy of Sciences, 1095–1100.

[222] Y.S. Sherif and M.L. Smith [1981]: "Optimal maintenance policies for systems subject to failure – A review", Naval Research Logistics Quarterly *28*, 47-74.

[223] K. Sladky [1974]: "On the set of optimal controls for Markov chains with rewards", Kybernatika *10*, 350–367.

[224] R.D. Smallwood [1966]: "Optimum policy regions for Markov processes with discounting", Operations Research *14*, 658–669.

[225] R.D. Smallwood and E.Sondik [1973]: "The optimal control of partially observable Markov processes over a finite horizon", Operations Research *21*, 1071–1088.

[226] D.R. Smith [1978]: "Optimal repairman allocation—asymptotic results", Management Science *24*, 665–674.

[227] M.J. Sobel [1981], "Myopic solutions of Markov decision processes and stochastic games", Operations Research *29*, 995–1009.

[228] M.J. Sobel [1985]: "Maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *4*, 157–159.

[229] M.J. Sobel [1994]: "Mean-variance trade-offs in an undiscounted MDP", Operations Research *42*, 175–183.

[230] E. Sondik [1978]: "The optimal control of partially observable Markov processes over the infinite horizon: discounted costs", Operations Research *26*, 282–304.

[231] I.M. Sonin [1999]: "The elimination algorithm for the problem of optimal stopping", Mathematical Methods of Operations Research *49*, 111–124.

[232] D. Spreen [1981]: "A further anti-cycling rule in multi-chain policy iteration for undiscounted Markov renewal programs", Zeitschrift für Operations Research *25*, 225–234.

[233] J. Stein [1988]: "On efficiency of linear programming applied to discounted Markovian decision problems", OR Spektrum *10*, 153–160.

[234] S.S. Stidham, Jr. [1985]: "Optimal control of admission to a queueing system", IEEE Transactions on Automatic Control *AC-30*, 705–713.

[235] S.S. Stidham, Jr. and R.R. Weber [1993]: "A survey of Markov decision models for control of networks of queues", Queueing Systems *13*, 291–314.

[236] J. Stoer and R. Bulirsch [1980]: "Introduction to numerical analysis", Springer-Verlag, New York.

[237] R. Strauch and A.F. Veinott, Jr. [1966]: "A property of sequential control processes", Report, Rand McNally, Chicago.

[238] M. Sun [1993]: "Revised simplex algorithm for finite Markov decision processes", Journal of Optimization Theory and Applications *79*, 405–413.

[239] L.C. Thomas [1981]: "Second order bounds for Markov decision processes", Journal of Mathematical Analysis and Applications *80*, 294–297.

[240] L.C. Thomas [1983]: "Constrained Markov decision processes as multi-objective problems", in: "Multi-objective decision making", Academic Press, 77–94.

[241] H.C. Tijms [1986]: "Stochastic modelling and analysis: a computational approach", Wiley, Chichester.

[242] J.N. Tsitsiklis [1986]: "A lemma on the multi-armed bandit problem", IEEE Transactions on Automatic Control *31*, 576–577.

[243] J.N. Tsitsiklis [1993]: "A short proof of the Gittins index theorem", Annals of Applied Probability *4*, 194–199.

[244] F.A. Van der Duyn Schouten and S.G. Vanneste [1990]: "Analysis and computation of (n, N)-strategies for maintenance of a two-component system", European Journal of Operations Research *48*, 260–274.

[245] J. Van der Wal [1980]: "The method of value oriented successive approximations for the average reward Markov decision processes", OR Spektrum *1*, 233–242.

[246] J. Van der Wal [1981]: "Stochastic dynamic programming", Mathematical Centre Tract 139, Mathematical Centre, Amsterdam.

[247] K.M. Van Hee [1978]: "Markov strategies in dynamic programming", Mathematics of Operations Research *3*, 191–201.

[248] K.M. Van Hee, A. Hordijk and J. Van der Wal [1977]: "Successive approximations for convergent dynamic programming", in: H.C. Tijms and J. Wessels (eds.), "Markov decision theory", Mathematical Centre Tract no. 93, Mathematical Centre, Amsterdam, 183–211.

[249] J.A.E.E. Van Nunen [1976a]: "A set of successive approximation method for discounted Markovian decision problems", Zeitschrift für Operations Research *20*, 203–208.

[250] J.A.E.E. Van Nunen [1976b]: "Contracting Markov decision processes", Mathematical Centre Tract 71, Mathematical Centre, Amsterdam.

[251] J.A.E.E. Van Nunen [1976c]: "Improved successive approximation methods for discounted Markovian decision processes", in: A. Prekopa (ed.), "Progress in Operations Research", North Holland, Amsterdam, 667–682.

[252] J.A.E.E. Van Nunen and J. Wessels [1976]: "A principle for generating optimization procedures for discounted Markov decision processes", Colloquia Mathematica Societatis Bolyai Janos, Vol. 12, North Holland, Amsterdam, 683–695.

[253] J.A.E.E. Van Nunen and J. Wessels [1977]: "The generation of successive approximations for Markov decision processes using stopping times", in: "Markov decision theory", H. Tijms and J. Wessels (eds.), Mathematical Centre Tract 93, Mathematical Centre, Amsterdam, 25–37.

[254] P.P. Varaiya, J.C. Walrand and C. Buyukkoc [1985]: "Extensions of the multi-armed bandit problem: the discounted case", IEEE Transactions on Automatic Control *30*, 426–439.

[255] A.F. Veinott, Jr. [1966a]: "On the optimality of (s, S) inventory policies: new condition and a new proof", SIAM Journal on Applied Mathematics *14*, 1067–1083.

[256] A.F. Veinott, Jr. [1966b]: "On finding optimal policies in discrete dynamic programming with no discounting", Annals of Math. Stats. *37*, 1284–1294.

[257] A.F. Veinott, Jr. [1969]: "Discrete dynamic programming with sensitive discount optimality criteria", Annals of Math. Stats. *40*, 1635–1660.

[258] A.F. Veinott, Jr. [1974]: "Markov decision chains", in: G.B. Dantzig and B.C. Eaves (eds.), "Studies in Optimization", Studies in Mathematics, Volume 10, The Mathematical Association of America, 124-159.

[259] R.C. Vergin and M. Scribian [1977]: "Maintenance scheduling for multi-component equipment", AIIE Transactions *9*, 297–305.

[260] O.J. Vrieze, [1987]: "Stochastic games with finite state and action spaces", CWI Tract 33, Centre for Mathematics and Computer Science, Amsterdam.

[261] K. Wakuta [1992]: "Optimal stationary policies in the vector-valued Markov decision process", Stochastic Processes and its Applications *42*, 149–156.

[262] K. Wakuta [1995]: "Vector-valued Markov decision processes and the systems of linear inequalities", Stochastic Processes and its Applications *56*, 159–169.

[263] K. Wakuta [1996]: "A new class of policies in vector-valued Markov decision processes", Journal of Mathematical Analysis and Applications *202*, 623–628.

[264] K. Wakuta [1999]: "A note on the structure of value spaces in vector-valued Markov decision processes", Mathematical Methods of Operations Research *49*, 77–86.

[265] J. Walrand [1988]: "An introduction to queueing networks", Prentice-Hall, Englewood Cliffs, New Jersey.

[266] R.R. Weber [1982]: "Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime.

[267] R.R. Weber [1992]: "On the Gittins index for multi-armed bandits", Annals of Applied Probability *2*, 1024–1033.

[268] R.R. Weber and S.S. Stidham, Jr. [1987]: "Optimal control of services rates in networks of queues", Advances in Applied Probability *19*, 202–218.

[269] G. Weiss [1982]: "Multiserver stochastic scheduling", in: M.A.H. Dempster, J.K. Lenstra and A.H.G. Rinnooy Kan (eds.), "Deterministic and stochastic scheduling", Reidel, Dordrecht, Holland, 157–179.

[270] G. Weiss [1988]: "Branching bandit processes", Probability in the Engineering and Information Sciences *2*, 269–278.

[271] J. Wessels and J.A.E.E. Van Nunen [1975]: "Discounted semi-Markov decision processes: linear programming and policy iteration", Statistical Neerlandica *29*, 1–7.

[272] J. Wessels [1977]: "Stopping times on Markov programming", in: Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Academia, Prague, pp. 575–585.

[273] C.C. White, III [1976]: "Procedures for the solution of a finite-horizon, partially observed, semi-Markov optimization problem", Operations Research *24*, 348–358.

[274] C.C. White, III [1991]: "A survey of solution techniques for the partially observed Markov decision process", Annals of Operations Research *33*, 215–230.

[275] C.C. White, III and W.T. Scherer [1989]: "Solution procedures for partially observed Markov decision processes", Op. Research *37*, 791–797.

[276] C.C. White, III and W.T. Scherer [1994]: "Finite-memory suboptimal design for partially observed Markov decision processes", Op. Research *42*, 439–455.

[277] D.J. White [1963]: "Dynamic programming, Markov chains, and the method of successive approximations", Journal of Mathematical Analysis and Applications *6*, 373–376.

[278] D.J. White [1978]: "Elimination of non-optimal actions in Markov decision processes", in: M.L. Puterman (ed.) Dynamic programming and its applications, Academic Press, New York, 131–160.

[279] D.J. White [1982]: "Multi-objective infinite-horizon discounted Markov decision processes", Journal of Mathematical Analysis and Applications *89*, 639–647.

[280] D.J. White [1985]: "Real applications of Markov decision theory", Interfaces *15:6*, 73–83.

[281] D.J. White [1988]: "Further real applications of Markov decision theory", Interfaces *18:5*, 55–61.

[282] D.J. White [1988]: "Mean, variance and probabilistic criteria in finite Markov decision processes: a review", Journal of Optimization Theory and Applications *56*, 1–30.

[283] D.J. White [1992]: "Computational approaches to variance-penalized Markov decision processes", OR Spektrum *14*, 79–83.

[284] D.J. White [1993]: "A survey of applications of Markov decision processes", Journal of the Operational Research Society *44*, 1073–1096.

[285] D.J. White [1993]: "Markov decision processes", Wiley, Chichester.

[286] D.J. White [1994]: "A mathematical programming approach to a problem in variance penalised Markov decision processes", OR Spektrum *15*, 225–230.

[287] D.J. White [1995]: "A superharmonic approach to solving infinite horizon partially observable Markov decision problems", Mathematical Methods of Operations Research *41*, 71–88.

[288] P. Whittle [1980]: "Multi-armed bandits and the Gittins index", Journal of the Royal Statistical Society, Series B *42*, 143–149.

[289] P. Whittle [1982]: "Optimization over time; dynamic programming and stochastic control", Volume I, Wiley, New York.

[290] P. Whittle [1982]: "Optimization over time; dynamic programming and stochastic control", Volume II, Wiley, New York.

[291] M. Yasuda [1988]: "The optimal value of Markov stopping problems with one-step look ahead policy", Journal of Applied Probability *25*, 544–552.

[292] Y.-S. Zheng and A. Federgruen [1991]: "Finding optimal (s, S)-policies is about as simple as evaluating a single policy", Op. Research *39*, 654–665.

Lodewijk Kallenberg
Mathematical Institute
University of Leiden
2300 RA Leiden, The Netherlands
kallenberg@math.leidenuniv.nl