Markov Decision Processes

Lodewijk Kallenberg
Leiden University
The Netherlands

May 28, 2000

Chapter 1

Finite state and action MDPs

ch:finite

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 rewards is commonly used. For the infinite horizon the utility function is less obvious. We consider several criteria: total discounted expected rewards, average expected rewards and more sentitive 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.

1.1 Introduction

1.1.1 Origin

Bellman's book [13], can be considered as starting point of Markov decision processes (MDPs). However, already in 1953, Shapley's paper [222] 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 [122], De Ghellinck [42], D'Epenoux [55], Manne [165] and Blackwell [27]. Since the early sixties, many results on MDPs are published in numerous journals, monographs, books and proceedings. Over thousand 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 [149], Mine and Osaki [168] and Ross [199] 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 [251], Van der Wal [247], Kallenberg [135], Federgruen [69], Vrieze [261], Hernández-Lerma [102], Altman [2] and Sennott [219] or they deal with the basic and advanced theory of MDPs (e.g. Bertsekas [15], Whittle [290], [291], Ross [201], Dietz and Nollau [63], Bertekas [17], Denardo [50], Heyman and Sobel [106], White [286], Puterman [187], Bertsekas [18], [19], Hernández-Lerma and Lasserre [103], [104], and Filar and Vrieze [79].

1.1.2 The model

We will restrict ourselves to discrete, finite Markovian decision problems, i.e. the *state space* E and the *action spaces* $A(i), i \in E$, are finite, and the decision time points t are equidistant, say $t = 1, 2, \dots$. If, at time point t, the system is in state i and action $a \in A(i)$ is chosen, then the following happens independently of the history of the process:

(1) a *reward* $r_i(a)$ is earned immediately;

(2) the process moves to state $j \in E$ with *transition probability* $p_{ij}(a)$, where $p_{ij}(a) \geq 0$ and $\Sigma_j p_{ij}(a) = 1$ 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) rewards over the planning horizon or the average expected reward per unit time. The decision maker has to find the optimal balance between immediate rewards and future rewards: a high immediate reward may bring the process in a bad situation for later rewards.

A *policy* R is a sequence of decision rules: $R = (\pi^1, \pi^2, \dots, \pi^t, \dots)$, where π^t is the decision rule at time point t. The decision rule at time point t may depend on "all information" of the system until time t, i.e. on the states at the time points $1, 2, \dots, t$ and on the actions at the time points $1, 2, \dots, t-1$. A formal definition of a decision rule is stated below.

Let ExA $= \{(i, a) \mid i \in E,\ a \in A(i)\}$ and let H_t denote the set of the possible histories of the system up to time point t, i.e.

$$H_t = \{(i_1, a_1, \dots, i_{t-1}, a_{t-1}, i_t) | (i_k, a_k) \in E \times A, 1 \leq k \leq t-1; i_t \in E\}. \quad (1.1)$$

A *decision rule* π^t at time point t gives the probability, as function of the history H_t to the action space, of choosing action a_t i.e. $\pi^t_{h_t}(a_t) \geq 0$ for every $a_t \in A(i_t)$ and $\Sigma_{a_t} \pi^t_{h_t}(a_t) = 1$ for every $h_t \in H_t$.

A policy $R = (\pi^1, \pi^2, \dots, \pi^t, \dots)$ is said to be *memoryless* if the decision rule π^t is independent of $(i_1, a_1, \dots, i_{t-1}, a_{t-1})$ for every $t \in \mathbb{N}$. Hence, in a memoryless policy the decision rule at time t only depends on the state i_t. Memoryless policies are also called *Markov* policies. If a policy is memoryless and the decision rules are independent of the time point, i.e. $\pi^1 = \pi^2 = \cdots$, then the policy is called a *stationary* policy. We will denote the stationary policy $R = (\pi, \pi, \dots)$ by π^∞. If the decision rule π of a stationary policy π^∞ is nonrandomized, then the policy is called *deterministic*. A deterministic policy can be described by a function f on E such that $f(i)$ is the action chosen in state $i, i \in$ E. A deterministic policy will be denoted by f^∞.

Any policy R and any *initial distribution* β, i.e. β_i is the probability that the system starts in state i, induce - by a theorem of Ionescu Tulcea (cf. Bertsekas and Shreve bertse78 [22]) - a probability measure $\mathbb{P}_{\beta,R}$ on H_∞, where

$$H_\infty = \{(i_1, a_1, i_2, a_2, \dots) \mid (i_k, a_k) \in E \times A, k = 1, 2, \dots\}. \quad (1.2)$$

Let the random variables X_t and Y_t denote the state and action at time $t (t = 1, 2, \dots)$ and let $\mathbb{P}_{\beta,R}[X_t = j, Y_t = a]$ be the notation for the probability that at time t the state is j and the action is a, given that policy R is used and β is the initial distribution. The next theorem shows that for any initial distribution β, any sequence of policies $R_1, R_2, \dots$ and any convex combination of the marginal distributions of $\mathbb{P}_{\beta,R_k}, k \in \mathbb{N}$, there exists a Markov policy R which has the same marginal distribution.

Theorem 1 *Given any initial distribution β, any sequence of policies $R_1, R_2, \dots$ and any sequence of nonnegative real numbers $p_1, p_2, \dots$ with $\Sigma_k p_k = 1$, there exists a Markov policy R_* such that for every $(j, a) \in E \times A$*

$$\mathbb{P}_{\beta,R_*}[X_t = j, Y_t = a] = \Sigma_k p_k \cdot \mathbb{P}_{\beta,R_k}[X_t = j, Y_t = a], t \in \mathbb{N}. \tag{1.3}$$

Corollary 2 *For any starting state i and any policy R, there exists a Markov policy R_* such that*

$$\mathbb{P}_{i,R_*}[X_t = j, Y_t = a] = \mathbb{P}_{i,R}[X_t = j, Y_t = a], t \in \mathbb{N}, (j, a) \in E \times A. \tag{1.4}$$

The results of Theorem 1 and Corollary 2 imply the sufficiency of Markov policies for performance measures which only depend on the marginal distributions. Corollary 2 is due to Derman and Strauch [61] and the extension to theorem 1 was given by Strauch and Veinott [238]. The result is further generalized to more general state and actions spaces by Hordijk [112] and Van Hee [248].

1.1.3 Optimality criteria

Let $v_i(R)$ be the *utility function* if policy R is used and state i is the starting state, $i \in$ E. The *value vector* v of this utility function is defined by

$$v_i = \sup_R v_i(R), i \in \text{ E}. \tag{1.5}$$

A policy R is an *optimal policy* if $v_i(R) = v_i, i \in$ E. 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 $v^t, t = 0, 1, \dots$. The sequence $v^0, v^1, \dots$ converges to the solution of the optimality equation. In *policy iteration* a sequence of improving policies $f_0^\infty, f_1^\infty, \dots$ is determined, i.e. $v(f_{t+1}^\infty) \geq v(f_t^\infty)$ 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 rewards over a finite horizon;
(2) total expected discounted rewards over an infinite horizon;
(3) average expected rewards 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 rewards* over the planning horizon, i.e. for policy R we will consider for starting state i

$$v_i^T(R) = \Sigma_{t=1}^T \mathbb{E}_{i,R}[r_{X_t}(Y_t)] = \Sigma_{t=1}^T \Sigma_{j,a} \mathbb{P}_{i,R}[X_t = j, Y_t = a] \cdot r_j(a). \tag{1.6}$$

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

$$[P(\pi^t)]_{ij} = \Sigma_a p_{ij}(a) \pi_i^t(a) \text{ for all } i,j \in E, \tag{1.7}$$

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

$$r_i(\pi^t) = \Sigma_a \pi_i^t(a) r_i(a) \text{ for all } i \in \text{ E}. \tag{1.8}$$

Hence the total expected rewards can be written in vector notation as

$$v^T(R) = \Sigma_{t=1}^T P(\pi^1) P(\pi^2) \cdots P(\pi^{t-1}) r(\pi^t). \tag{1.9}$$

It can be shown (we refer to section 2) that an optimal Markov policy $R_* = (f_*^1, f_*^2, \dots, f_*^T)$ exists, where f_*^t is a deterministic decision rule $1 \leq t \leq$ T. Such policy is called nonstationary deterministic. The nonstationarity is due to the finiteness of the planning horizon.

Next, we consider an infinite planning horizon. In that case there is not a unique optimality criterion. Different optimality criteria are meaningful: discounted rewards, total rewards, average rewards or more sensitive criteria. In the discounted reward criterion an amount r that is obtained at time point 1 has at time point 2 the value $(1+\rho) \cdot r$, at time point 3 the value $(1+\rho)^2 \cdot r$, etc., where ρ is the *interest rate*. In the discounted reward criterion the rewards are considered as earned at time point 1. Therefore, the reward $r_{X_t}(Y_t)$ at time point t has at time point 1 the discounted value $\alpha^{t-1}r_{X_t}(Y_t)$, where $\alpha = (1+\rho)^{-1}.\alpha$ is called *the discount factor* and $0 < \alpha < 1$.

The *total expected α-discounted reward*, given initial state i and policy R, is denoted by $v_i^\alpha(R)$ and defined by

$$\begin{aligned} v_i^\alpha(R) &= \Sigma_{t=1}^\infty \mathbb{E}_{i,R}[\alpha^{t-1} r_{X_t}(Y_t)] \\ &= \Sigma_{t=1}^\infty \alpha^{t-1} \Sigma_{j,a} \mathbb{P}_{i,R}[X_t = j, Y_t = a] r_j(a). \end{aligned} \tag{1.10}$$

In section 3 it will be shown that there exists an optimal deterministic policy and that any stationary policy π^∞, and therefore also any deterministic policy, satisfies

$$v^\alpha(\pi^\infty) = \Sigma_{t=1}^\infty \alpha^{t-1} P(\pi)^{t-1} r(\pi) = [I - \alpha P(\pi)]^{-1} r(\pi). \tag{1.11}$$

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 $\Sigma_{t=1}^\infty \mathbb{E}[r_{X_t}(Y_t)]$. However, this infinite sum does not exist, in general. When the average reward criterion is used, the limiting behaviour of the expectation of $\frac{1}{T}\Sigma_{t=1}^T r_{X_t}(Y_t)$ is considered. Since $\lim_{T\to\infty} \frac{1}{T}\Sigma_{t=1}^T \mathbb{E}[r_{X_t}(Y_t)]$ or $\mathbb{E}[lim_{T\to\infty} \frac{1}{T}\Sigma_{t=1}^T r_{X_t}(Y_t)]$ 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:

$$\phi_i(R) = \liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T \mathbb{E}_{i,R}[r_{X_t}(Y_t)], i \in E; \tag{1.12}$$

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

$$\Phi_i(R) = \limsup_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T \mathbb{E}_{i,R}[r_{X_t}(Y_t)], i \in E; \tag{1.13}$$

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

$$\Psi_i(R) = \mathbb{E}_{i,R}[\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T r_{X_t}(Y_t)], i \in E; \tag{1.14}$$

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

$$\Psi_i(R) = \mathbb{E}_{i,R}[\limsup_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^T r_{X_t}(Y_t)], i \in E. \tag{1.15}$$

Lemma 3 *(i)* $\psi(R) \leq \phi(R) \leq \Phi(R) \leq \Psi(R)$ *for every policy* R; *(ii)* $\psi(\pi^\infty) = \phi(\pi^\infty) = \Phi(\pi^\infty) = \Psi(\pi^\infty)$ *for every stationary policy* π^∞.

Remark

In Bierth [26] is shown that the four citeria 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 R the inequalities of Lemma 3 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: the streams of rewards $0, 0, 0, 0, 0, \dots$ and $100, 100, 0, 0, 0, \dots$ are measured as the same. 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 deterministic optimal policies exist and that these policies are (at least) average optimal.

A policy R_* is called *bias-optimal* if $lim_{\alpha \uparrow 1}[v^\alpha(R_*) - v^\alpha] = 0$ and *Blackwell optimal* if $v^\alpha(R_*) = v^\alpha$ for all $\alpha \in [\alpha_0, 1)$ for some α_0. These two criteria are special cases of a class of criteria called n-*discount optimality*,for $n = -1, 0, 1, \dots$.

A policy R_* is called n-discount optimal if $lim_{\alpha \uparrow 1}(1-\alpha)^{-n}[v^\alpha(R_*) - v^\alpha] = 0$. Obviously, 0-discount optimal is the same as bias-optimal. One can show that (-1)-discount optimality is equivalent to average optimality, and that Blackwell optimality is n-discount optimality for all $n \geq N-1$, where $N = \#$E. In this chapter we will always use the notation N for the number of states.

There is also the concept of n-*average optimality.* For any policy $R, t \in \mathbb{N}$ and $n = -1, 0, 1, \dots$, let the vector $v^{n,t}(R)$ be defined by

$$v^{n,t}(R) = \begin{cases} v^t(R) & for \quad n = -1 \\ \Sigma_{s=1}^t v^{n-1,s}(R) & for \quad n = 0, 1 \dots \end{cases} \tag{1.16}$$

R_* is said to be n-average optimal if $\liminf_{T\to\infty} \frac{1}{T}[v^{n,T}(R_*) - v^{n,T}(R)] \geq 0$ for all policies R. (-1)-average optimality is equivalent to average optimality. Furthermore, it can be shown that for all $n \in \{-1, 0, \dots\}$ n-average optimality is equivalent to n-discount optimality.

In a fundamental paper Blackwell [27] presented a mathematical 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 honour of Blackwell, such policy is called a Blackwell optimal policy.
The n-discount optimality criterion, and the policy iteration method for finding an n-discount optimal policy, was proposed by Veinott [258]. He also showed that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. Sladky [224] has introduced the concept of n-average optimality; furthermore, he also showed the equivalence between this criterion and the n-discount optimality.

1.1.4 Applications

White has published three papers on 'real applications' of Markov decision theory (White [281], [282] and [285]. We also mention Lamond's chapter in this book on water reservoir applications. Many 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 (by Schäl) and in telecommunication (by Altman).

Routing problems
In routing problems the problem is to find an optimal route through a network. Well known is the deterministic *shortest path* problem. Consider a layered network, i.e. the nodes V are partitioned into $V_1 = \{v_1\}, V_2, \cdots, V_{m-1}$, $V_m = \{v_n\}$, and if (i,j) is an arc then $i \in V_k$ and $j \in V_{k+1}$ for some $1 \leq k \leq m-1$. In the shortest path problem one has to find the (length of the) shortest path from node v_1 to node v_n. This problem can be modelled as an MDP with a finite horizon (m stages). The state space E consists of the nodes; the action set $A(i)$ in state i corresponds to the arcs with starting point node i; the transitions are deterministic: $p_{ij}(a) = 1$ if action a corresponds to the arc (i,j), and the length l_{ij} is considered as cost when in state i action (i,j) is chosen, i.e. $r_i(a) = -l_{ij}$ if action $a = (i,j)$.
Let u_i be the minimum distance from state i to the destination v_n. Then, by the principle of optimality, we obtain the so-called optimality equation for the shortest path problem, i.e. $u_i = \min\{l_{ij} + u_j | (i,j) \in A\}, i \neq n; u_n = 0$. u_1 is the solution of the shortest path problem. There is also a stochastic version of the shortest path problem.

Another application of this kind is the *maximum reliability* problem. In this network the (direct) 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 obtain the maximum probability of reaching a terminal node n when starting from node 1.

Let P be a path from node 1 to node n. Notice that $\mathbb{P}(P)$, the probability to reach n from 1 when path P is chosen, satisfies $\mathbb{P}(P) = \Pi_{(i,j)\in P} p_{ij}$ and that $\max_P \mathbb{P}(P)$ is equivalent to $\max_P[\log\{\mathbb{P}(P)\}]$ which in turn is equivalent to $\min_P[-\log\{\mathbb{P}(P)\}]$. Since $\log\{\mathbb{P}(P)\} = \Sigma_{(i,j)\in P} \log[p_{ij}]$ the maximum reliability problem is equivalent to the shortest path problem with $l_{ij} = -\log[p_{ij}]$.

Results for the 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 [195]. We also mention White's chapter in this book on transportation applications.

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.

An example of an optimal stopping problem is the *house selling* problem (see Ross [199]). At the beginning of each period a decision has to be taken whether to sell the house for the best offer so far, or to wait for another period (when the house is unsold there are maintenance costs for the next period). We assume that the numbers p_j are known, where p_j is the probability of an offer j during a period (these probabilities are stationary, i.e. they are independent of the period). This house selling problem can be modelled as an optimal stopping problem with the following state descrition: state i means that the best offer so far is i. Selling the house corresponds to the stopping action; otherwise we continue and have probability p_j for a transition from state i to state $j > i$ and the transition probability from i to i is equal to $\Sigma_{j\leq i}\, p_j$.

For this kind of problems it can be shown that a *control limit policy* is optimal (i.e. sell the house at the first offer of at least i_* for a certain amount i_*). It is possible to derive an explicit expression for i_*.

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 control limit policies. We refer also to Sonin's chapter in this book.

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 e.g. 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 modelled as MDPs with the total expected reward as optimality criterion.

To the class of target problems we may count the *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. We assume that the costs are positive. It is easy to see that the first passage problem is equivalent to this MDP.

A second class of target problems are the *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 Maitra and Sudderth's chapter in this book.

The first passage problem was introduced by Eaton and Zadeh [67] under the name "pursuit problem". The dynamic programming approach is introduced in Derman [56]. The standard reference on gambling is Dubins and Savage [64]. Dynamic programming approaches are given in Ross [200] 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 selled at price s_i, a new item is bought at price c, and the transition to the new state is instantaneously. 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 modelled as an MDP. It turns out that an efficient algorithm for the computation of an optimal policy exists, the so-called *myopic algorithm* with complexity $\mathcal{O}(N^3)$, see Gal [83].

Next, we consider the model *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 very 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 structural 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 [223] contained already over 500 references. Results on the optimality of control limit policies for replacement problems can be found in Derman [57], Kolesar [147], Derman [58], Ross [199] and Kao [139].

Maintenance and repair problems

In maintenance and repair problems there is a system which is subject to deterioration and failure. In each period the state of the system is observed. 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 (total costs, discounted costs or average costs) over a finite or infinite horizon.

As an example of a repair problem, consider a system with two components in series. Each component may be either in a functioning or a failed state. At the end of a period a functioning component may be in a failed state with probability p. For the production of one unit operations on both components are needed. When both components are functioning (at the begin of the period) a good unit is produced; if one of the components has failed, then a good unit is produced with probability q; if both components have failed the unit is defective. After inspection of the state of the components, there are the following possible actions:

- if both components are functioning: do nothing;
- if one component has failed: either do nothing or repair the component;
- if both components have failed: either repair one component or repair both components.

The repair of a component costs c and the production of a good unit gives a reward r. The objective is to maximize the average reward over an infinite horizon. This problem can easily be modelled as an MDP.

The one-component problem is described in Klein [146]. The two-component maintenance problem was introduced by Vergin and Scriabin [260]. Other contributions in this area are e.g. Oezekici [174], and Van der Duyn Schouten and Vanneste [245]. An n-component series system is discussed in Katehakis and Derman [140]. Asymptotic results for highly reliable systems can be found in Smith [227], Katehakis and Derman [141], 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.
As an example we will describe an inventory model with backlogging. At the start of each week the inventory manager observes the inventory and decides how many units to order. We assume that orders are delivered instantaneously, and that there is a finite inventory capacity of B. Let D be the stochastic demand during one week and let $\mathbb{P}[D = j]$ be denoted by $p_j, 0 \leq j \leq$ M. If the demand during a period excesses the inventory on hand, then the shortage is backlogged in the next period. If the inventory at the beginning of a week is i (shortages are modelled as negative inventory), the number of units ordered is a and the inventory at the end of the week is j, then the following costs are involved:

$$\text{Ordering costs}: \ K \cdot \delta(a) + k \cdot a, \ \text{where } \delta(a) = \begin{cases} 1 & if \quad a \geq 1 \\ 0 & if \quad a \leq 0 \end{cases}$$

$$\text{Inventory costs}: \ h \cdot \delta(j) \cdot j; \ \text{backlogging costs} \ : q \cdot \delta(-j) \cdot (-j).$$

The data K, k, h, q and $p_j, 0 \leq j \leq M$ are known.
This inventory problem can be modelled as an MDP with:

$$\begin{aligned}
E &= \{-M, \dots, -1, 0, 1, \dots, B\}; \ A(i) = \{a \geq 0 | 0 \leq i + a \leq B\}; \\
p_{ij}(a) &= \begin{cases} p_{i+a-j} & j \leq i + a \\ 0 & j > i + a \end{cases} \qquad i, j \in E, \quad a \in A(i); \\
r_i(a) &= -\big[K \cdot \delta(a) + k \cdot a \ + \ \Sigma_{j=0}^{i+a} p_j \cdot h \cdot (i + a - j) \\
&\qquad + \ \Sigma_{j=i+a+1}^{M} p_j \cdot q \cdot (j - i - a)\big]
\end{aligned}$$

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 [203]. The existence of an optimal (s, S)-policy in the infinite horizon model is shown by Iglehart [127]. Another related paper is Veinott [256]. For the relation between discounted and average costs we refer to Hordijk and Tijms [120]. For the computation of the values s and S we refer to papers like Federgruen and Zipkin [76], and Zheng and Federgruen [293].

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 [242]).

A queue, or a network of queues, is a useful model for many applications, e.g. manufacturing, computer, telecommunication and traffic systems. Control models can optimize certain performance measures by varying the control parameters of the system. We distinguish between *admission control* and *service rate control*.

As an example of admission control, consider an $M/M/1$ queue where customers arrive as a Poisson process and the service time is exponential. One has the option to reject an arrival. However, a cost c must be paid for each rejection. Furthermore, there are holding costs h per period for each customer in the system. Hence, in an optimal situation there is a balance between rejection of new customers and congestion of customers in the system. Assume that the total expected discounted costs have to be minimized. It can be shown that there is a control limit or *threshold policy* which is optimal, i.e. there is a value i_* and an arriving new customer is rejected if there are more than i_* customers in the system. For details we refer to Walrand [266].

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 [269]. Surveys of optimal control of (networks of) queues can be found in the book by Walrand [266] and the papers by Stidham [235] and Stidham and Weber [236].

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 is a utility function by which different policies can be measured, and we want to find a policy that optimizes the utility function. Such problems are also considered in the control of queueing systems. Then, instead of jobs and machines, the terms customers and servers are frequently used. 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).
As a first example we consider *one server allocation (with preemption) to parallel queues*. Customers arrive at a system of m parallel queues and one server. The system operates at discrete time points, i.e. arrival times and service times have values in the set $\{1, 2, \dots\}$. Furthermore, the arrival times are arbitrary and the service time T_i, for a customer in queue i, is geometrically distributed with rate μ_i, which implies that $\mathbb{E}[T_i] = \mu_i^{-1}$. At any time point the server chooses a customer from one of the queues; so, this is a server assignment model. Services which are going on may be interrupted and resumed later on (preemption). For each customer in queue i, a cost c_i is charged per unit of time that this customer is in the system. Which policy minimizes the total costs in T periods? It can be shown that the so-called *μc-rule* is an optimal policy. This rule assigns the server to queue k, where k is the queue with $\mu_k c_k = \max_i\{\mu_i c_i \mid \text{queue } i$ is nonempty$\}$. Note that $c_i\mu_i$ is the expected costs per unit of service for a customer in queue i, and by using the *μc-rule*, the largest reduction of the expected costs in the next period is obtained.
A second example is *customer allocation to parallel identical queues*. Customers arrive at a system of m parallel queues. Each queue has a server and the service times are identically and exponentially distributed with rate μ. At arrival, a customer has to join one of the queues (customer assignment model). Which policy minimizes the total discounted number of customers in the system? It is clear that the policy that allocates an arriving customer to the shortest queue is optimal for this individual customer. It can be shown that this policy, which is called the *shortest queue policy (SQP)*, is also optimal for the overall criterion.
As a final example of stochastic scheduling we consider the problem of *mini-*

mizing the makespan and the total flowtime. Consider a set of n jobs, where each job has to be processed on one of m identical machines. Job j has an exponential distribution with rate μ_j, where $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. Let T_j be the completion time of job $j, 1 \leq j \leq$ n. Which policy minimizes the expected makespan, i.e. $\mathbb{E}[\max(T_1, T_2, \ldots, T_n)]$. This model can be considered as a server assignment model with m servers (machines): which jobs are processed by machine 1, which jobs by machine 2, etc. It can be shown that an optimal policy is the policy which chooses the longest expected processing time *(LEPT)* first. Hence, the jobs are allocated to the machines in the order $1, 2, \ldots$,n.
When the expected total flowtime $\mathbb{E}[\Sigma_{j=1}^n T_j]$ has to be minimized, then the order is reversed, i.e. the shortest expected processing time first *(SEPT)*. In this case the allocation of the jobs to the machines is $n, n-1, \ldots, 1$.
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 [267] and by Chang, Hordijk, Righter and Weiss [33]. For reviews on stochastic scheduling we refer to Weiss [270], Walrand [266] (chapter 8), Righter [194] and Weber's chapter in this book.

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. As utility function the total discounted reward is chosen. There are many applications of this model, e.g. 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.* Surprisingly, these indices can be computed for each project seperately. 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]). The importance of Gittins work was not recognized in the

seventies. The re-discovery is due to Whittle [289], who gave a more easy and natural proof. Other proofs are given by Ross [201], Varaiya, Walrand and Buyoccoc [255], Weber [268] and Tsitsiklis [244]. Several methods are developed for the computation of the Gittins indices: Varaiya, Walrand and Buyukkoc [255], Chen and Katehakis [35], Kallenberg [136], Katehakis and Veinott [142], Ben-Israel and S.D.Flåm [14], and Liu and Liu [156].

1.2 FINITE HORIZON

Consider an MDP with a finite horizon of T periods. In fact, we can analyse 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 notated by $r_i^t(a)$ and $p_{ij}^t(a)$. 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 is provided by Bellman [13].

Theorem 4 *Let $x_i^{T+1} = 0, i \in E$. Determine for $t = T, T-1, ..., 1$ a deterministic decision rule f_t such that*

$$r_i^t(f_t(i)) + [P(f_t)x^{t+1}]_i = \max_{a \in A(i)} \{r_i^t(a) + \Sigma_j p_{ij}^t(a) \cdot x_j^{t+1}\}, i \in E,$$

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 v^T.

If $r_i^t(f_t(i)) + [P^t(f_t)x^{t+1}]_i = \max_{a \in A(i)}\{r_i^t(a) + \Sigma_j p_{ij}^t(a) \cdot x_j^{t+1}\}, i \in E$, then we denote $r^t(f_t) + P^t(f_t)x = \max_{\mathrm{ExA}}\{r^t + P^t x\}$ and $f_t \in argmax_{\mathrm{ExA}}\{r^t + P^t x\}$.

Algorithm I (finite horizon)

1. $x = 0$.
2. Determine for $t = T, T-1, \dots, 1$:

   $f_t \in argmax_{\mathrm{ExA}}\{r^t + P^t x\}$ and $x = r^t(f_t) + P^t(f_t)x$.
3. $R^* = (f_1, f_2, \dots, f_T)$ 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 [125].
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*, are applicable. These results can be found in Derman and Klein [59] and in Kallenberg [132], [133].

1.3 DISCOUNTED REWARD CRITERION

1.3.1 Introduction

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

$$v_i^\alpha = \max_a \{r_i(a) + \alpha \Sigma_j p_{ij}(a) v_j^\alpha\}, i \in E \tag{3.1}$$

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

$$x_i = \max_a \{r_i(a) + \alpha \Sigma_j p_{ij}(a) x_j\}, i \in \text{ E}. \tag{3.2}$$

It turns out that U is a *monotone contraction mapping* with as fixed point the value vector v^α. By the general theory of monotone contraction mappings questions can be answered as:
- Is there a unique fixed-point?
- How can the fixed-point be computed?
- What is the rate of convergence of the computation?
We first introduce some concepts and properties of contraction mappings.

Let X be a normed Banach space and $B : \ X \ \to$ X. The operator B is called a *contraction mapping* if for some $\beta \ \in \ [0,1)$ one has

$$\| Bx - \text{ By } \| \leq \beta \cdot \| x - y \| \text{ for all } x, y \ \in \text{ X}. \tag{3.3}$$

The number β is called the *contraction factor* of B. An element $x^* \ \in \ X$ is said to be a *fixed-point* of B if

$$Bx^* = x^*. \tag{3.4}$$

The next *Fixed-point theorem* shows the existence of a unique fixed point in a Banach space.

Theorem 4 *Let X be a Banach space and suppose that $B : X \to X$ is a contraction. Then, for every $x \in X, x^* = lim_{n\to\infty} B^n x$ exists and x^* is the unique fixed-point of B.*

Hence, a straightforward method to approximate the fixed point is: $x^n = Bx^{n-1} = B^n x^0, n \in \mathbb{N}$ (choose x^0 arbitrarily). The convergence rate of the sequence $x^0, x^1, \ldots$ to the fixed point is based on the following result.

Theorem 5 *Let X be a Banach space and suppose that $B : X \to X$ is a contraction mapping with contraction factor β and fixed-point x^*. Then, for all $x \in X$*

$$
\begin{aligned}
(i) \quad \| x^* - B^n x \| &\leq \beta(1-\beta)^{-1} \cdot \| B^n x - B^{n-1} x \| \\
&\leq \beta^n (1-\beta)^{-1} \cdot \| Bx - x \|, n \in \mathbb{N}; \\
(ii) \quad \| x^* - x \| &\leq (1-\beta)^{-1} \cdot \| Bx - x \| .
\end{aligned}
$$

From the above theorem it follows that the convergence of $B^n x$ to the fixed-point is at least linear. This kind of convergence is also called *geometric convergence.*

Next, we mention some properties of monotone (contraction) mappings. Suppose that $B : X \to X$ is a mapping in a partially ordered set X. B is called *monotone* if $x \leq y$ implies $Bx \leq$ By. The next theorem shows an order property.

Theorem 6 *Let X be a partially ordered Banach space and suppose that $B : X \to X$ is a monotone contraction mapping with fixed-point x^*. Then,*

$$
\begin{aligned}
(i) &\quad Bx \leq x \quad \textit{implies that} \quad x^* \leq Bx \leq x; \\
(ii) &\quad Bx \geq x \quad \textit{implies that} \quad x^* \geq Bx \geq x.
\end{aligned}
$$

It is well-known that $\| x \|_\infty = \max_{1\leq i\leq N} \mid x_i \mid$ is a norm, the *supremum norm*, in the partially ordered Banach space $\mathbb{R}^N$ with ordering $x \leq y$ if $x_i \leq y_i$ for $1 \leq i \leq$ N. Hence, for $x \in \mathbb{R}^N$ we have $x \leq \| x \|_\infty \cdot e$, where e is the vector with all components equal 1.The *subordinate matrix norm* $\| P \|_\infty$for a square matrix P is defined by $\| P \|_\infty = \max_i \Sigma_j |p_{ij}|$(cf. Stoer and Bulirsch [237] p.178). The following results, which we will express in the supremum norm, can be generalized to the so-called *μ-norm*, defined by $\| x \|_\mu = \max_{1\leq i\leq N} \mu_i^{-1} \cdot \mid x_i \mid$, where $\mu \in \mathbb{R}^N$ with $\mu_i > 0$. For $\mu = e$, the μ-norm and the supremum norm coincide. More results are formulated in the following lemmas.

Lemma 7 *Let B be a monotone contraction in $\mathbb{R}^N$ with respect to the supremum norm, with contraction factor β and fixed-point x^*. Suppose that there* *exist scalars a and b such that $a \cdot e \leq Bx - x \leq b \cdot e$ for some $x \in \mathbb{R}^N$. Then,*

$$\begin{aligned} x-(1-\beta)^{-1} \mid a \mid \cdot e \leq\ & Bx - \beta(1-\beta)^{-1} \mid a \mid \cdot e \leq x^* \leq \\ & Bx + \beta(1-\beta)^{-1} \mid b \mid \cdot e \leq x + (1-\beta)^{-1} \mid b \mid \cdot e. \end{aligned}$$

The proof of this lemma can be given by first showing (by induction) that $B^n x \leq Bx + (\beta + \cdots + \beta^{n-1}) \mid b \mid \cdot e \leq x + (1 + \beta + \cdots + \beta^{n-1}) \mid b \mid \cdot e, n \in \mathbb{R}^N$. By letting $n \to \infty$ we obtain $x^* \leq Bx + \beta(1-\beta)^{-1} \mid b \mid \cdot e \leq x + (1-\beta)^{-1} \mid b \mid \cdot e$. The other side of the inequalities can be shown similarly.
Since $- \parallel Bx - x \parallel_\infty \cdot e \leq Bx - x \leq \parallel Bx - x \parallel_\infty \cdot e$, we obtain the following corollary.

Corollary 8 *Let B be a monotone contraction in $\mathbb{R}^N$ with respect to $\parallel x \parallel_\infty$, with contraction factor β and fixed-point x^*. Then,*

$$\begin{aligned} x - (1-\beta)^{-1} \cdot \parallel Bx - x \parallel_\infty \cdot e \leq\ & Bx - \beta(1-\beta)^{-1} \cdot \parallel Bx - x \parallel_\infty \cdot e \leq x^* \leq \\ & Bx + \beta(1-\beta)^{-1} \cdot \parallel Bx - x \parallel_\infty \cdot e \leq \\ & x + (1-\beta)^{-1} \cdot \parallel Bx - x \parallel_\infty \cdot e. \end{aligned}$$

Lemma 9 *Let B be a monotone contraction in $\mathbb{R}^N$ with respect to $\parallel x \parallel_\infty$, with contraction factor β, fixed-point x^* and with the property that $B(x + c \cdot e) = Bx + \beta c \cdot e$ for every $x \in \mathbb{R}^N$ and scalar c. Suppose that for some scalars a and b and for some $x \in \mathbb{R}^N$, $a \cdot e \leq Bx - x \leq b \cdot e$. Then,
$x + (1-\beta)^{-1} a \cdot e \leq Bx + \beta(1-\beta)^{-1} a \cdot e \leq x^* \leq Bx + \beta(1-\beta)^{-1} b \cdot e \leq x + (1-\beta)^{-1} b \cdot e$.*

The proof of lemma 9 can be given analogously to the proof of lemma 7. We will apply these general results of monotone contraction mappings to special mappings in our MDP model. These mappings are U, defined in (3.2) and L_π, for any stationary decision rule π, defined by

$$L_\pi x = r(\pi) + \alpha P(\pi) x. \tag{3.5}$$

Let $f_x(i) \in argmax_{A(i)} \{r_i(a) + \alpha \Sigma_j p_{ij}(a) x_j\}$, then

$$L_{f_x} x = Ux = \max_\pi L_\pi x. \tag{3.6}$$

Theorem 10 *With respect to the norm $\parallel x \parallel_\infty, L_\pi$ and U are monotone contraction mappings in $\mathbb{R}^N$ with contraction factor α.*

Proof. Suppose that $x \geq$ y. Let π be any stationary decision rule. Because $P(\pi) \geq 0$, we may write $L_\pi x = r(\pi) + \alpha P(\pi)x \geq r(\pi) + \alpha P(\pi)y = L_\pi y$, i.e. L_π is monotone. Since $Ux = \max_\pi L_\pi x \geq L_{f_y} x \geq L_{f_y} y = Uy, U$ is also monotone.
Because $\| L_\pi x - L_\pi y \|_\infty = \| \alpha P(\pi)(x-y) \|_\infty \leq \alpha \cdot \| P(\pi) \|_\infty \cdot \| x - y \|_\infty = \alpha \cdot \| x - y \|_\infty$, L_π is a contraction with contraction factor α. For the operator U we obtain
$Ux - Uy = L_{f_x} x - L_{f_y} y \leq L_{f_x} x - L_{f_x} y = \alpha \cdot P(f_x)(x-y) \leq \alpha \cdot \| x - y \|_\infty \cdot e$.
Interchanging x and y yields $Uy - Ux \leq \alpha \cdot \| x - y \|_\infty \cdot e$. Hence,
$\| Ux - Uy \|_\infty \leq \alpha \cdot \| x - y \|_\infty$, i.e. U is a contraction with contraction factor α. ∎

The next theorem shows that, for any stationary policy π^∞, the total α-discounted reward $v^\alpha(\pi^\infty)$ is the fixed-point of the mapping L_π.

Theorem 11 *For any stationary decision rule $\pi, v^\alpha(\pi^\infty)$ is the unique solution of the functional equation $L_\pi x = x$.*

Using the property $v^\alpha(\pi^\infty) = [I - \alpha P(\pi)]^{-1} r(\pi)$, the proof follows directly from the theorems 4 and 10.

Corollary 12 $v^\alpha(\pi^\infty) = \lim_{n\to\infty} L_\pi^n x$ *for any* $x \in \mathbb{R}^N$.

Theorem 13 *v^α is the unique solution of the equation $Ux = x$.*
Proof. An outline of the proof is as follows. First, show that $v^\alpha(R) \leq Uv^\alpha$ for any Markov policy R, implying - by theorem 1 - that $v^\alpha = sup_{R\in C(M)} v^\alpha(R) \leq Uv^\alpha$. Then, prove that for any $\epsilon > 0$, $v_i^\alpha \geq (Uv^\alpha)_i - \epsilon$, $i \in E$, which implies that $v^\alpha \geq Uv^\alpha$. ∎

Because, $v^\alpha = Uv^\alpha = L_{f_v \alpha} v^\alpha$, the last equality by (3.6), it follows from theorem 13 that $v^\alpha = v^\alpha(f_{v^\alpha}^\infty)$, i.e. $f_{v^\alpha}^\infty$ is an optimal policy. If f^∞ satisfies $r_i(f(i)) + \alpha \Sigma_j\, p_{ij}(f(i)) v_j^\alpha = \max_a \{r_i(a) + \alpha \Sigma_j p_{ij}(a) v_j^\alpha\}, i \in E$, then f^∞ is called a *conserving* policy. $f_{v^\alpha}^\infty$ is a conserving policy and conserving policies are optimal. Therefore, the equation $Ux = x$ is called the *optimality equation.*

Corollary 14 *(i) There exists a deterministic and stationary α-discounted optimal policy; (ii) $v^\alpha = \lim_{n\to\infty} U^n x$ for any $x \in \mathbb{R}^N$; (iii) Any conserving policy is α-discounted optimal.*

Since $L_f(x + c\cdot e) = L_f x + \alpha c \cdot e$ and $U(x + c \cdot e) = Ux + \alpha c \cdot e$ for any $x \in \mathbb{R}^N$ and any scalar c, we can apply lemma 9 to obtain bounds for the fixed points $v^\alpha(f^\infty)$ and v^α of the operators L_f and U, respectively.

Lemma 15 *For any $x \in \mathbb{R}^N$ we have*

$$
\begin{aligned}
(i)\quad x+(1-\alpha)^{-1}\min_i(Ux-x)_i\cdot e \;&\leq Ux+\alpha(1-\alpha)^{-1}\min_i(Ux-x)_i\cdot e\\
&\leq v^\alpha(f_x^\infty)\leq v^\alpha\\
&\leq Ux+\alpha(1-\alpha)^{-1}\max_i(Ux-x)_i\cdot e\\
&\leq x+(1-\alpha)^{-1}\max_i(Ux-x)_i\cdot e.
\end{aligned}
$$

$(ii)\quad \parallel v^\alpha - v^\alpha(f_x^\infty)\parallel_\infty \leq \alpha(1-\alpha)^{-1}span(Ux-x),$
where $span(y)=\max_i y_i - \min_i y_i$.

An action $a\in A(i)$ is called *suboptimal* if there does not exist an α-discounted optimal policy f^∞ with $f(i)=a$. Because f^∞ is α-discounted optimal if and only if $v^\alpha(f^\infty) = v^\alpha$, and because $v^\alpha = Uv^\alpha$, an action $a \in A(i)$ is suboptimal if and only if

$$v_i^\alpha \;>\; r_i(a)+\alpha\Sigma_j p_{ij}(a)v_j^\alpha. \tag{3.7}$$

Suboptimal actions can be excluded. Not directly by (3.7), because v^α is unknown, but by using the bounds on v^α as given by lemma 15. Then, by the monotonicity of U, the next result is obtained.

Theorem 16 *(i) Suppose that* $x\leq v^\alpha\leq y$. *If* $r_i(a)+\alpha\Sigma_j p_{ij}(a)y_j < (Ux)_i$, *then action* $a \in A(i)$ *is suboptimal. (ii) Suppose that for some scalars b and c,* $x+b\cdot e\leq v^\alpha\leq x+c\cdot e$. *If* $r_i(a)+\alpha\Sigma_j p_{ij}(a)x_j < (Ux)_i - \alpha(c-b)$, *then action* $a \in A(i)$ *is suboptimal.*

Using the bounds of v^α from lemma 15, we obtain suboptimality for an action $a \in A(i)$ if

$$r_i(a)+\alpha\Sigma_j p_{ij}(a)x_j < (Ux)_i - \alpha(1-\alpha)^{-1}\text{span}(Ux-x) \tag{3.8}$$

or

$$r_i(a)+\alpha\Sigma_j p_{ij}(a)(Ux)_j < (U^2x)_i - \alpha^2(1-\alpha)^{-1}\text{span}(Ux-x) \tag{3.9}$$

Remark
If we relax the property that $\Sigma_j p_{ij}(a)=1$ to $\Sigma_j p_{ij}(a)\leq 1$ for all (i,a) and require that the model is *transient*, i.e. the matrix $\Sigma_{t=1}^\infty[P(f)]^t$ 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 [258] and Kallenberg [135]. Other references on this topic are van Hee, Hordijk and van der Wal [249], Denardo and Rothblum [53], and Hordijk and Kallenberg [116].

Already in 1953, Shapley [222] analysed 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 generalization to the μ-norm was made by Wessels [272] and Van Nunen [251]. They apply this μ-norm to MDPs with countable state space and unbounded rewards. The details of the proof of theorem 13 can be found in Ross [199]. An alternative proof that U has a fixed point, based on Brouwer's theorem, was given in Shapiro [221]. The concepts 'conserving' and 'span' were introduced by Dubins and Savage [64] and Bather [10]. Concerning the bounds of lemma 15, the weakest bounds were proposed by MacQueen [163] and the strongest by Porteus [180]. Related papers are Porteus [181] and Bertsekas [16]. The notion that suboptimal actions can be excluded if bounds on the value vector are available can be found in MacQueen [164], which paper includes the test (3.8). Test (3.9) is proposed by Porteus [180]. Other suboptimality tests can be found in Hastings and Mello [97], White [279] and Thomas [240].

1.3.2 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 deterministic policies $f_1^\infty, f_2^\infty, \ldots$ is constructed such that

$$v^\alpha(f_{k+1}^\infty) > v^\alpha(f_k^\infty) \text{ for } k = 1, 2, \ldots \tag{3.10}$$

Because there are finite deterministic policies f^∞, 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 10.

Lemma 17 *(i) If $L_f x \leq x$, then $v^\alpha(f^\infty) = \lim_{n\to\infty} L_f^n x \leq L_f x \leq x$; (ii) if $L_f x > x$, then $v^\alpha(f^\infty) = \lim_{n\to\infty} L_f^n x \geq L_f x > x$.*

For every $i \in E$ and deterministic policy f^∞, the set $A(i, f)$ be defined by

$$A(i, f) = \{a \in A(i) | r_i(a) + \alpha \Sigma_j p_{ij}(a) v_j^\alpha(f^\infty) > v_i^\alpha(f^\infty)\}. \tag{3.11}$$

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 18 *(i) If $A(i,f)=\emptyset$ for every $i \in E$, then f^∞ is α-discounted optimal; (ii) If $A(i,f)\neq\emptyset$ for some $i \in E$, then $v^\alpha(g^\infty) > v^\alpha(f^\infty)$ for any policy g with $g \neq f$ and $g(i) \in A(i,f)$ if $g(i)\neq f(i)$.*

Proof. (i) $A(i,f)=\emptyset$, $i \in E$, implies $L_g v^\alpha(f^\infty) = r(g) + \alpha P(g) v^\alpha(f^\infty) \leq v^\alpha(f^\infty)$ for every g. By lemma 17(i) $v^\alpha(g^\infty) \leq v^\alpha(f^\infty)$ for every g, i.e. f^∞ is optimal.

(ii) Take any $g \neq f$ such that $g(i) \in A(i,f)$ if $g(i) \neq f(i)$. Then, if $g(i) \neq f(i)$, $r_i(g) + \alpha\Sigma_j p_{ij}(g) v_j^\alpha(f^\infty) > v_i^\alpha(f^\infty)$.

If $g(i) = f(i)$, $r_i(g) + \alpha\Sigma_j p_{ij}(g) v_j^\alpha(f^\infty) = r_i(f) + \alpha\Sigma_j p_{ij}(f) v_j^\alpha(f^\infty) = v_i^\alpha(f^\infty)$. Hence, $L_g v^\alpha(f^\infty) = r(g) + \alpha P(g) v^\alpha(f^\infty) > v^\alpha(f^\infty)$ and, by lemma 17(ii), $v^\alpha(g^\infty) > v^\alpha(f^\infty)$. ∎

Let

$$s_{ia}(f) = r_i(a) + \alpha\Sigma_j p_{ij}(a) v_j^\alpha(f^\infty) - v_i^\alpha(f^\infty),\ a \in A(i) \text{ and } i \in E. \quad (3.12)$$

Algorithm II (policy iteration; discounted rewards)

1. Start with any deterministic policy f^∞.
2. Compute $v^\alpha(f^\infty)$ as unique solution of the linear system $L_f x =$ x.
3. Determine for every $i \in E : A(i,f) = \{a \in A(i) | s_{ia}(f) > 0\}$.
4. If $A(i,f)=\emptyset$ for every $i \in E$: 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 [122]. Blackwell [27] has provided a strong mathematical treatment of this method. In Porteus [184] and in Hartley, Lavercombe and Thomas [92] efficient ways to determine $v^\alpha(f^\infty)$ as solution of the linear system $L_f x = x$ are analysed.

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. $g(i) \in argmax_a s_{ia}(f)$.
2. It can be shown (see Puterman and Brumelle [188]) that the policy iteration method, with the above choice for g, is equivalent to solve the optimality equation $Ux = x$ by Newton's method.

3. Furthermore, we can derive a result on the convergence rate. It can be shown that $x^n = v^\alpha(f_n^\infty), n = 1, 2, \dots,$ 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 $\| v^\alpha - v^\alpha(f_{n+1}^\infty) \|_\infty \leq 2\alpha(1-\alpha)^{-1} \| v^\alpha - v^\alpha(f_n^\infty) \|_\infty$, there is geometric convergence. Already in Pollatschek and Avi-Itzhak [179], in the context of stochastic games, the equivalence between the policy iteration method and Newton's method was noticed. A related paper is Schweitzer [212]. Puterman and Brumelle [188] were the first who derived result for the rate of convergence.
4. We can also include in policy iteration the exclusion of suboptimal actions. We can use e.g. test (3.8) with $x = v^\alpha(f^\infty)$. Since, for $x = v^\alpha(f^\infty)$, $(Ux - x)_i = \max_a\{r_i(a) + \Sigma_j p_{ij}(a) v_j^\alpha(f^\infty) - v_i^\alpha(f^\infty)\} = \max_a s_{ia}(f), i \in E$, we have $span(Ux - x) = \max_i[\max_a s_{ia}(f)] - \min_i[\max_a s_{ia}(f)]$. Hence (3.8) 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 $b \in A(i)$ 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. We can implement a modification of the method. Instead of the steps 3 and 4, we take the steps 3' and 4' which are as follows:
Step 3'. For i = 1 to N do

a. $d_{ia}(f) = r_i(a) + \alpha\Sigma_{j=1}^{i-1} p_{ij}(a) x_j + \alpha\Sigma_{j=i}^{N} p_{ij}(a) v_j^\alpha(f^\infty), a \in A(i)$;

b. if $d_{ia}(f) \leq v_i^\alpha(f^\infty)$ for every $a \in A(i) : x_i = v_i^\alpha(f^\infty)$ and $g(i) = f(i)$;

c. if $d_{ia}(f) > v_i^\alpha(f^\infty)$ for some $a \in A(i) : x_i = \max_a d_{ia}(f)$ and take $g(i) = argmax d_{ia}(f)$.

Step 4'. If $g(i) = f(i)$ for every $i \in E$, then go to step 6.
This modified algorithm, which can be shown to be correct, is proposed in Hastings [94].
6. In Schmitz [205] the question is raised: "Does there exist a polynomial bound for the number of iterations in the policy iteration?". Meister and Holzbaur [166] have shown that this method is polynomially in time. In ng [172] is shown that the complexity of one iteration is $\mathcal{O}(mN^2)$, where m is the number of states i for which $g(i) \neq f(i)$.

1.3.3 Linear programming

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

$$v_i \geq r_i(a) + \alpha\Sigma_j p_{ij}(a) v_j \text{ for every } (i,a) \in E \times A. \tag{3.13}$$

Theorem 19 *v^α is the (componentwise) smallest α-superharmonic vector.*

Proof. Theorem 13 implies that $v_i^\alpha \geq r_i(a) + \alpha\Sigma_j p_{ij}(a)v_j^\alpha$ for every (i,a), i.e. v^α is α-superharmonic. Suppose that $v \in \mathbb{R}^N$ is also α-superharmonic. Then, $v \geq r(f) + \alpha P(f)v$ for every f^∞, which implies that $[I - \alpha P(f)]v \geq r(f)$. Since $[I - \alpha P(f)]^{-1} = \Sigma_{t=0}^\infty \alpha^t P(f)^t \geq 0$, we obtain $v \geq [I - \alpha P(f)]^{-1}$ $r(f) = v^\alpha(f^\infty)$. Hence, $v^\alpha = \max_f v^\alpha(f^\infty) \leq v$, i.e. v^α is the smallest α-superharmonic vector. ∎

Corollary 20 *v^α is the unique optimal solution of the LP-problem*

$$min\{\Sigma_j \beta_j v_j | \Sigma_j [\delta_{ij} - \alpha p_{ij}(a)]v_j \geq r_{ia}, (i,a) \in E \times A\} \tag{3.14}$$

where $\beta_j > 0$ for every $j \in E$.

By corollary 20, the value vector v^α can be found as optimal solution of the linear program (3.14). This program does not give an optimal policy. However, an optimal policy can be obtained from the solution of the *dual program*

$$\max\left\{\Sigma_i\Sigma_a r_i(a)x_{ia} \;\middle|\; \begin{array}{r} \Sigma_i\Sigma_a[\delta_{ij} - \alpha p_{ij}(a)]\ x_{ia} = \beta_j,\ j \in E \\ x_{ia} \geq 0,\ (i,a) \in E \times A \end{array}\right\} \tag{3.15}$$

Theorem 21 *Let x^* be an optimal solution of (3.15). Then, a policy f^∞ with $x^*_{jf(j)} > 0$ for every $j \in E$ exists and is an optimal policy.*

Proof. From the constraints of (3.15) it follows that $\Sigma_a x^*_{ja} = \beta_j + \alpha\Sigma_i\Sigma_a p_{ij}(a)x^*_{ia} \geq \beta_j > 0$, $j \in E$. Let f^∞ be such that $x^*_{jf(j)} > 0, j \in E$. The complementary slackness property of linear programming implies that $\Sigma_j[\delta_{ij} - \alpha p_{ij}(f(i)]v_j^\alpha = r_i(f), i \in E$. Hence, $[I - \alpha P(f)]^\alpha = r(f)$, implying $v^\alpha = [I - \alpha P(f)]^{-1} r(f) = v^\alpha(f^\infty)$, i.e. f^∞ is an optimal policy. ∎

Moreover, there is a one-to-one correspondence between the set of feasible solutions of (3.15) and the set of stationary policies, given by the following relations. For a stationary policy π^∞ the feasible solution $x(\pi)$ satisfies

$$x_{ia}(\pi) = [\beta^T(I - \alpha P(\pi))^{-1}]_i \cdot \pi_{ia}, (i,a) \ \in \ E \times A. \tag{3.16}$$

Conversely, for a feasible solution x of (3.15), define $\pi^\infty(x)$ by

$$\pi_{ia}(x) = x_{ia}/\Sigma_a x_{ia}(i,a) \ \in \ E \times A. \tag{3.17}$$

Theorem 22 *The mapping (3.16) is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the dual program (3.15)* *with (3.17) as the inverse mapping; furthermore, the set of extreme feasible solutions of (3.15) corresponds to the set of deterministic policies.*

Algorithm III (linear programming; discounted rewards)

1. Take any $\beta \in \mathbb{R}^N$ with $\beta_j > 0, j \in E$.
2. Compute optimal solutions v^* and x^* of the dual pair LP-problems (3.14) and (3.15).
3. Take any f_* such that $x^*_{if_*(i)} > 0$, $i \in E$.
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 23 *(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 (3.12)) are available in the simplex tableaus as the so-called *reduced costs*.
2. The variables $x_{ia}(\pi)$, defined in (3.16) 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 $\beta_j, j \in E$.
3. The linear programming method is the only method which can 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 [187], section 8.9. The constraints have to be expressed in terms of the state-action frequencies and added to the dual program (3.15). Constrained problems have a stationary, but not necessarily deterministic optimal policy. For the details we refer to Kallenberg [135], and to Hordijk and Kallenberg [116]. In Altman and

Shwartz [4] the sensitivity of constrained MDPs is investigated. Altman, Hordijk and Kallenberg [3] have analysed 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 [116]. In Sun [239] an implementation of the LP-method is described, based on the revised simplex method. Stein [234] has investigated the computational aspects of the linear programming method in comparison with other methods. 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.

1.3.4 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. Furthermore, 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 14(ii) it follows that $v^\alpha = lim_{n\to\infty} U^n x$ for every $x \in \mathbb{R}^N$.

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

$$\begin{cases} v^1 \in \mathbb{R}^N & \text{arbitrarily chosen} \\ v^{n+1} = Uv^n, & n = 1, 2, \dots \end{cases} \tag{3.18}$$

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

$$v^{n+1} = Uv^n = L_{f_n} v^n = r(f_n) + \alpha P(f_n) v^n, n \in \mathbb{N}. \tag{3.19}$$

The next lemma shows that f_n^∞ is an ϵ-optimal policy for n sufficiently large. The proof is based on contraction properties as applied in lemma 15.

Lemma 24 $\| v^\alpha(f_n^\infty) - v^\alpha \|_\infty \leq 2\alpha^n (1-\alpha)^{-1} \cdot \| v^2 - v^1 \|_\infty, n \in \mathbb{N}.$

Algorithm IV (value iteration; discounted rewards)

1. Choose $\epsilon > 0$ and $x \in \mathbb{R}^N$ arbitrarily.

2. Compute $y = Ux$ 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 goto step 2.

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

Theorem 25 *(i)* $\| v^\alpha(f_x^\infty) - v^\alpha \|_\infty \leq 2\alpha(1-\alpha)^{-1} \cdot \| Ux - x \|_\infty$; *(ii)* $\| Ux - v^\alpha \|_\infty \leq \alpha(1-\alpha)^{-1} \cdot \| Ux - x \|_\infty$.

In the next theorem we summarize some suboptimality tests.

Theorem 26 *An action $a \in A(i)$ is suboptimal if one of the following tests is satisfied:*

$$r_i(a) + \alpha\Sigma_j p_{ij}(a)x_j < (Ux)_i - 2\alpha(1-\alpha)^{-1} \cdot \| Ux - x \|_\infty \tag{3.20}$$

$$r_i(a) + \alpha\Sigma_j p_{ij}(a)(Ux)_j < (U^2x)_i - 2\alpha^2(1-\alpha)^{-1} \cdot \| Ux - x \|_\infty \tag{3.21}$$

$$r_i(a) + \alpha\Sigma_j p_{ij}(a)x_j < (Ux)_i - \alpha(1-\alpha)^{-1} span(Ux - x) \tag{3.22}$$

$$\begin{aligned} r_i(a) + \alpha\Sigma_j p_{ij}(a)(Ux)_j < \ & (Ux)_i + \alpha(1-\alpha)^{-1}\min_i(Ux-x)_i - \\ & \alpha^2(1-\alpha)^{-1}\max_i(Ux-x)_i \end{aligned} \tag{3.23}$$

Remarks
1. In the usual computation scheme of the value iteration algorithm, test (3.22) 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 $y_j = (Ux)_j$ which are already computed, i.e.

$$\begin{aligned} y_i = \max_a\{r_i(a) + \alpha\Sigma_{j=1}^{i-1} p_{ij}(a)y_j + \ & \alpha\Sigma_{j=i}^N p_{ij}(a)x_j\}, \\ & i = 1, 2, \dots, N \end{aligned} \tag{3.24}$$

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

$$\begin{aligned} y_i \ &= \max_a[1 - \alpha p_{ii}(a)]^{-1} \cdot \{r_i(a) + \alpha\Sigma_{j=1}^{i-1} p_{ij}(a)y_j \\ &+ \alpha\Sigma_{j=i+1}^N p_{ij}(a)x_j\}, \ i = 1, 2, \dots, N \end{aligned} \tag{3.25}$$

For both variants it can be shown that the corresponding operators are contraction mappings with fixed point v^α and with contraction factor at most α. Hence, they may be considered as an acceleration of the basic algorithm.

The proofs can be given by induction of the states. Also suboptimaly tests for the actions can be included.

Howard [122] was the first who studied value iteration for a Markov decision process. For a survey of the basic properties 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 [150]. An overview of these variants is presented in Porteus [183]. Other techniques, based on successive overrelaxation and stopping times, in order to accelerate the convergence can be found in Reetz [191] and [192], Schellhaas [204], Wessels [272], Van Nunen [250], Van Nunen and Wessels [253] and [254], Porteus and Totten [185], Porteus [182], Herzberg and Yechali [105], and Bertsekas [20]. Holzbaur [110] has presented a theoretically polynomial bound for the number of steps in the value iteration method.

1.3.5 Modified policy iteration

In section 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

$$y = x + A(Ux - x), \text{ where } A = [I - \alpha P(g)]^{-1} \text{ with } g \text{ such that } L_g x = Ux \tag{3.26}$$

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

$$A^{(k)} = \Sigma_{i=0}^{k-1} \alpha^i [P(g)]^i \text{ for some } 1 \leq k \leq \infty \tag{3.27}$$

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. $L_{f_n} x^n = Ux^n$.

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

Algorithm V (modified policy iteration; discounted rewards)

1. Choose $x \in \mathbb{R}^N$, $\epsilon > 0$ and a deterministic policy f^∞.
2. a. Choose k with $1 \leq k \leq \infty$;

   b. Determine g such that $L_g x = Ux$, where $g(i) = f(i)$ if possible.
3. If $\| Ux - x \|_\infty \leq (1-\alpha)\epsilon$: g^∞ is an 2ϵ-optimal policy and Ux is an $\alpha\epsilon$-approximation of v^α (STOP);

   Otherwise: $x = L_g^k x$, $f = g$ and go to step 2.

Remarks
1. Since $x^{n+1} = L_{f_n}^{k(n)}x^n$, 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. Although this operator is neither a contraction nor monotone, it can be shown that $v^\alpha = \lim_{n\to\infty} L_{f_n}^{k(n)}x^n$ 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 [189] and independently Van Nunen [250], [251] and [252] have developed the modified policy iteration method. The first authors have shown the convergence under the assumption that the starting vector x satisfies $Ux \geq x$. The convergence of the method for an arbitrary starting vector was proved by Rothblum [202]. In Van Nunen [250] 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 [190]. Puterman [186] reviews computational results for the modified policy iteration method.

1.4 AVERAGE REWARD CRITERION

1.4.1 Introduction

We start this section with some properties of the *transition matrix* P, i.e. a matrix with nonnegative elements and with every row sum equal to 1. Note that, for every stationary policy $\pi^\infty, P(\pi)$ is a transition matrix. The *stationary matrix* P^* of P is defined by the so-called *Cesaro-limit* of P^n, i.e.

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

The stationary matrix has some nice properties, summarized in the next theorem.

Theorem 27 *(i)* $P^*P = PP^* = P^*P^* = P^*$; *(ii)* $[P - P^*]^n = P^n - P^*,\ n \geq 1$; *(iii)* $\lim_{\alpha\uparrow 1}(1-\alpha)\Sigma_{n=0}^{\infty}\alpha^n(P^n - P^*) = 0$; *(iv)* $[I - P + P^*]$ *is nonsingular with inverse* $[I-P+P^*]^{-1} = \lim_{n\to\infty}\frac{1}{n}\Sigma_{k=1}^{n}\Sigma_{i=1}^{k}[P-P^*]^{i-1}$; *(v) for every stationary policy* π^∞, *the average reward* $\phi(\pi^\infty)$ *satisfies* $\phi(\pi^\infty) = P^*(\pi)r(\pi)$, *where* $P^*(\pi)$ *is the stationary matrix of the transition matrix* $P(\pi)$; *(vi)* $\phi(\pi^\infty) = \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(\pi^\infty)$.

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

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

Theorem 28 *The deviation matrix satisfies (i)* $D = \lim_{n\to\infty}\frac{1}{n}\Sigma_{k=1}^{n}\Sigma_{i=1}^{k}[P^{i-1} - P^*]$ *(ii)* $P^*D = DP^* = [I-P]D + P^* - I = D[I-P] + P^* - I = 0$.

For the proofs of the theorems 27 and 28 we refer to books on Markov chains (e.g. Kemeny and Snell [145] [1960]). A treatment, related to MDPs, of the stationary, the fundamental and the deviation matrix can be found in Veinott [259] [1974]. 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 analysed easier in case they are unichain or irreducible, which may lead to simplified algorithms for solving these MDPs. We assume the reader familiar with concepts from Markov chains as *recurrent state*, *transient state*, *recurrent class* and *irreducibility*.

An MDP is *irreducible* if the Markov chain $P(f)$ is irreducible for every policy f^∞. We say that an MDP is *unichain* if for every policy f^∞ the Markov chain $P(f)$ has exactly one recurrent class plus a (possibly empty) set of transient states. Hence, an irreducible MDP is unichained (the reverse statement is not true, in general). An MDP is *multichain* if there exists a policy f^∞ for which the Markov chain $P(f)$ has at least two ergodic classes. If the ergodic structure is unknown, one can always use the general approach for multichain MDPs. The determination whether or not an MDP is irreducible is an easy problem, i.e. polynomially solvable, which means that the number of steps is bounded by a polynomial function of the problem's data (see Kallenberg [138]).

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.

Theorem 29 *There are numbers* $\alpha_m, \alpha_{m-1}, ..., \alpha_0, \alpha_{-1}$ *and policies* f_m^∞, $f_{m-1}^\infty, ..., f_0^\infty$ *such that (i)* $0 = \alpha_m < \alpha_{m-1} < \cdots < \alpha_0 < \alpha_{-1} = 1$; *(ii)* $v^\alpha(f_j^\infty) = v^\alpha$ *for all* $\alpha \in [\alpha_j, \alpha_{j-1}), j = m, m-1, ..., 0$.
Proof. We give an outline of the proof. Since $v^\alpha(f^\infty)$ is the solution of the system $[I - \alpha P(f)]x = r(f)$, each component $v_i^\alpha(f^\infty)$ 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 $\{\alpha_k \mid k = 1, 2, \dots\}$ and $\{f_k \mid k = 1, 2, \dots\}$ such that $\alpha_1 \leq \alpha_2 \leq \cdots$ with $\lim_{k\to\infty} \alpha_k = 1$ and $v^\alpha = v^\alpha(f_k^\infty) > v^\alpha(f_{k-1}^\infty)$ for $\alpha = \alpha_k, k = 2, 3, \dots$. Because there is a finite number of deterministic policies this implies the existence of two policies, say f^∞ and g^∞, that both are in turn optimal for an infinite number of increasing α's with limit $\alpha = 1$. Let $h(\alpha) = v^\alpha(f^\infty) - v^\alpha(g^\infty)$, then for any $i \in E$, $h_i(\alpha)$ 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 treatment of the *Laurent expansion* of a stationary policy π^∞.

Theorem 30 *Let $u^k(\pi), k = -1, 0, \dots$ 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 $v^\alpha(\pi^\infty) = \alpha^{-1} \cdot \Sigma_{k=-1}^\infty [(1-\alpha)/\alpha]^k \cdot u^k(\pi)$, where $\alpha_0(\pi) = \parallel D(\pi) \parallel / [\parallel D(\pi) \parallel +1]$.*

Proof. Let $x(\pi) = \alpha^{-1} \cdot \Sigma_{k=-1}^\infty [(1-\alpha)/\alpha]^k \cdot u^k(\pi)$. Then, $x(\pi) = \frac{\phi(\pi^\infty)}{1-\alpha} + \frac{D(\pi)}{\alpha} \cdot \Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha] \cdot D(\pi)\}^k r(\pi)$ for $\parallel [(\alpha-1)/\alpha] \cdot D(\pi) \parallel < 1$, i.e. $\alpha_0(\pi) < \alpha < 1$. Since $v^\alpha(\pi^\infty)$ is the unique solution of the linear system $[I - \alpha P(\pi)]x = r(\pi)$, it is sufficient to show that $r(\pi) - [I - \alpha P(\pi)]x(\pi) = 0$. This can be done in the following steps, using the results of theorem 27(i) and theorem 28(ii).
$r(\pi) - [I - \alpha P(\pi)]x(\pi) = r(\pi) - [I - \alpha P(\pi)](1-\alpha)^{-1}P^*(\pi)r(\pi) - [I - \alpha P(\pi)]\alpha^{-1}D(\pi) \cdot \Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k r(\pi) = r(\pi) - P^*(\pi)r(\pi) - [\alpha\{I - P(\pi)\} + (1-\alpha)I]\alpha^{-1}D(\pi) \cdot \Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k r(\pi) = [I - P^*(\pi)]r(\pi) - [I - P^*(\pi)]\Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k r(\pi) + \Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^{k+1} r(\pi) = [I - P^*(\pi)]r(\pi) - [I - P^*(\pi)]\Sigma_{k=0}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k \cdot r(\pi) + \Sigma_{k=1}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k r(\pi) = [I - P^*(\pi)]r(\pi) - [I - P^*(\pi)]r(\pi) - [I - P^*(\pi) - I] \cdot \Sigma_{k=1}^\infty \{[(\alpha-1)/\alpha]D(\pi)\}^k r(\pi) = 0$. ∎

Corollary 31 *(i) $\phi(\pi) = \lim_{\alpha \uparrow 1}(1-\alpha)v^\alpha(\pi)$; (ii) $v^\alpha(\pi^\infty) = \frac{\phi(\pi^\infty)}{1-\alpha} + u^0(\pi) + \epsilon(\alpha)$, where $\lim_{\alpha \uparrow 1} \epsilon(\alpha) = 0$.*

The first part of the Laurent expansion as presented in corollary 31(ii) was derived by Blackwell [27]. The complete Laurent expansion was proposed by Miller and Veinott [167]. The vector $u^0(\pi)$ is called the *bias vector* of policy π^∞.

1.4.2 The optimality equation

The multichain case

Before we introduce the optimality equation, we first give some prerequisites.

Lemma 32 $\lim_{\alpha \uparrow 1}(1-\alpha)v^\alpha(R) \geq \phi(R)$ *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 33 *Any stationary Blackwell optimal policy is also average optimal.*

Proof. Let π^∞ be any stationary Blackwell optimal policy. Then, by theorem 27(vi) $\phi(\pi^\infty) = \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(\pi^\infty) = \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha \geq \lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(R) \geq \phi(R)$ for any policy R. ∎

In the discounting 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 34 *Consider the system*

$$\begin{cases} x_i & = \max_{a\in A(i)} \Sigma_j p_{ij}(a)x_j, & i \in E \\ x_i + y_i & = \max_{a\in A(i,x)}\{r_i(a) + \Sigma_j p_{ij}(a)y_j\}, & i \in E \end{cases} \tag{4.3}$$

where $A(i,x) = \{a \in A(i) \mid x_i = \Sigma_j p_{ij}(a)x_j\}$, $i \in E$.
This system has the following properties: (i) $x = u^{-1}(f_0)$, $y = u^0(f_0)$, where f_0^∞ is a Blackwell optimal policy, satisfies (4.3); (ii) If (x,y) is a solution of (4.3), then x equals the value vector ϕ.

Proof. Since f_0^∞ is a Blackwell optimal policy, we have for $\alpha \in [\alpha_0, 1)$ $v_i^\alpha(f_0^\infty) = v_i^\alpha = \max_{a\in A(i)}\{r_i(a) + \alpha\Sigma_j p_{ij}(a)v_j^\alpha\} \geq r_i(a) + \alpha\Sigma_j p_{ij}(a)v_j^\alpha(f_0^\infty)$, $(i,a) \in E \times A$. Combining this result with corollary 31(ii) yields
$(1-\alpha)^{-1}\phi_i(f_0^\infty) + u_i^0(f_0) + \epsilon_i(\alpha) = v_i^\alpha(f_0^\infty) \geq r_i(a) + \alpha\Sigma_j p_{ij}(a)v_j^\alpha(f_0^\infty) = r_i(a) + \Sigma_j p_{ij}(a)v_j^\alpha(f_0^\infty) - (1-\alpha)\Sigma_j p_{ij}(a)v_j^\alpha(f_0^\infty) = r_i(a) + \Sigma_j p_{ij}(a).$
$[(1-\alpha)^{-1}\phi_j(f_0^\infty) + u_j^0(f_0) + \epsilon_j(\alpha)] - (1-\alpha)\Sigma_j p_{ij}(a)[(1-\alpha)^{-1}\phi_j(f_0^\infty) + u_j^0(f_0) + \epsilon_j(\alpha)]$, $(i,a) \in E \times A$, $\alpha \in [\alpha_0, 1]$. This result holds for all $\alpha \in [\alpha_0, 1)$, so comparing the terms with $(1-\alpha)^{-1}$ gives $\phi_i(f_0^\infty) \geq \Sigma_j p_{ij}(a)\phi_j(f_0^\infty)$. Furthermore, when the terms with $(1-\alpha)^{-1}$ are equal, we compare the terms with $(1-\alpha)^0$:
$u_i^0(f_0) \geq r_i(a) + \Sigma_j p_{ij}(a)u_j^0(f_0) - \Sigma_j p_{ij}(a)\phi_j(f_0^\infty) = r_i(a) + \Sigma_j p_{ij}(a)u_j^0(f_0) - \phi_i(f_0^\infty)$ for each (i,a) with $\phi_i(f_0^\infty) = \Sigma_j p_{ij}(a)\phi_j(f_0^\infty)$. Since $\phi(f_0^\infty) = P(f_0)\phi(f_0^\infty)$ and $u^0(f_0) = D(f_0)r(f_0) = r(f_0) - \phi(f_0^\infty) + P(f_0)u^0(f_0)$, part (i) is shown.

Next, we prove part (ii). Let (x,y) be a solution (4.3). Then, for any f^∞, $x \geq P(f)x$, implying $x \geq P^*(f)x$ and consequently $p_{ij}^*(f)\{x_j - [P(f)x]_j\} = 0$ for every $i,j \in E$. Let i be a recurrent state in the Markov chain induced by $P(f)$. Since $p_{ii}^*(f) > 0$, $f(i) \in A(i,x)$ and $x_i + y_i \geq r_i(f) + \Sigma_j p_{ij}(f)y_j$. Therefore, $P^*(f)[x+y] \geq P^*(f)[r(f) + P(f)y] = \phi(f^\infty) + P^*(f)y$, i.e. $\phi(f^\infty) \leq P^*(f)x \leq x$.

On the other hand, any solution of the system gives a policy g^∞ which satisfies $x = P(g)x$ and $x + y = r(g) + P(g)$y. Hence, $x = P^*(g)x$, and

$\phi(g^\infty) = P^*(g)r(g) = P^*(g)[x + y - P(g)y] = x$. This completes the proof that $x = \max_f \phi(f^\infty) = \phi$. ∎

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, i.e. $\phi_i = \Sigma_j p_{ij}(a)\phi_j$, $i \in E$, $a \in A(i)$. 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 35 *Consider the system* $x + y_i = max_a\{r_i(a) + \Sigma_j p_{ij}(a)y_j\}, i \in E$. *This system has the following properties: (i)* $x \cdot e = u^{-1}(f_0)$, $y = u^0(f_0)$, *where* f_0^∞ *is a Blackwell optimal policy, satisfies 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 [217]. 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 [213]. 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].

1.4.3 Policy iteration

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

The multichain case

Theorem 36 *Consider the following system of linear equations*

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

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* $y(f) = u^0(f)$.

Proof. We first show that $x(f) = u^{-1}(f)$, $y(f) = u^0(f)$ and $z(f) = u^1(f)$ is a solution of (4.4). We can write,
$I - P(f)]x(f) = [I - P(f)]u^{-1}(f) = [I - P(f)]P^*(f)]r(f) = 0$. $x(f) + [I - P(f)]y(f) = P^*(f)]r(f) + [I - P(f)]D(f)r(f) = \{P^*(f)] + [I - P(f)]D(f)\}$ $r(f) = r(f)$. $y(f) + [I - P(f)]z(f) = D(f)r(f) - [I - P(f)]D(f)^2r(f) = \{I - [I - P(f)]D(f)\}D(f)r(f) = P^*(f)D(f)r(f) = 0$. Next, let (x, y, z) be any solution of (4.4). Then, $x = P(f)x$ implies $x = P^*(f)x = P^*(f)\{r(f) - [I - P(f)]y\} = P^*(f)r(f) = u^{-1}(f)$. Since $y + [I - P(f)]z = 0$, $P^*(f)y = 0$, we can write $\{[I - P(f)] + P^*(f)\}y = [I - P(f)]y = r(f) - P^*(f)r(f)$, i.e. $y = \{[I-P(f)]+P^*(f)\}^{-1}[I-P^*(f)]r(f) = [D(f)+P^*(f)][I-P^*(f)]r(f) = D(f)r(f) = u^0(f)$. ∎

For every $i \in E$ and every policy f^∞, we define the action subset $B(i, f)$ by

$$B(i,f) = \left\{ a \in A(i) \;\middle|\; \begin{array}{l} \Sigma_j p_{ij}(a)\phi_j(f^\infty) > \phi_i(f^\infty) \text{ or} \\ \Sigma_j p_{ij}(a)\phi_j(f^\infty) = \phi_i(f^\infty) \text{ and} \\ r_i(a) + \Sigma_j p_{ij}(a)u_j^0(f) > \phi(f^\infty) + u_i^0(f) \end{array} \right\} \tag{4.5}$$

Theorem 37 *(i) If $B(i, f) = \emptyset$ for every $i \in E$, then f^∞ is an average optimal policy; (ii) If $B(i, f) \neq \emptyset$ for at least one i and the policy g^∞ satisfies $g \neq f$ and $g(i) \in B(i, f)$ if $g(f) \neq f(i)$, then $\phi(g^\infty) \geq \phi(f^\infty)$ and $v^\alpha(g^\infty) > v^\alpha(f^\infty)$ for α sufficiently close to 1.*
Proof. (i) Since $B(i, f) = \emptyset$ for every $i \in E$, we have for any policy h^∞, $\Sigma_j p_{ij}(h)\phi_j(f^\infty) \leq \phi_i(f^\infty)$ and $r_i(h) + \Sigma_j p_{ij}(h)u_j^0(f) \leq \phi_i(f^\infty) + u_i^0(f)$, if $\Sigma_j p_{ij}(h)\phi_j(f^\infty) = \phi_i(f^\infty)$. Let policy $R = (h, f, f, \dots)$. Then, $v^\alpha(R) = r(h) + \alpha P(h)v^\alpha(f^\infty)$. From theorem 30 it follows that $\alpha v^\alpha(f^\infty) = \alpha \cdot \frac{\phi(f^\infty)}{1-\alpha} + u^0(f) + \epsilon(\alpha) = \frac{\phi(f^\infty)}{1-\alpha} + u^0(f) - \phi(f^\infty) + \epsilon_1(\alpha)$, implying $v^\alpha(R) = r(h) + P(h)\{\frac{\phi(f^\infty)}{1-\alpha} + u^0(f) - \phi(f^\infty) + \epsilon_1(\alpha)\} = (1-\alpha)^{-1}P(h)\phi(f^\infty) + r(h) + P(h)u^0(f) + \epsilon_1(\alpha)$. Since $v^\alpha(f^\infty) = (1 - \alpha)^{-1}\phi(f^\infty) + u^0(f) + \epsilon_2(\alpha)$, we have

$$\begin{aligned} v^\alpha(f^\infty) \; &-v^\alpha(R) = (1-\alpha)^{-1}[\phi(f^\infty) - P(h)\phi(f^\infty)] + u^0(f) - r(h) \\ &-P(h)u^0(f) + \epsilon_3(\alpha), \text{ where } \lim_{\alpha\uparrow 1} \epsilon_k(\alpha) = 0, \; k = 1,2,3. \end{aligned} \tag{4.6}$$

Hence, $v^\alpha(f^\infty) \geq v^\alpha(R) + \epsilon(\alpha) = r(h) + \alpha P(h)v^\alpha(f^\infty) + \epsilon(\alpha)$ for α close to 1 and $[I - \alpha P(h)]v^\alpha(f^\infty) \geq r(h) + \epsilon(\alpha)$. Therefore, $v^\alpha(f^\infty) \geq [I - \alpha P(h)]^{-1}[r(h) + \epsilon(\alpha)] = v^\alpha(h^\infty) + (1 - \alpha)^{-1} \cdot \epsilon(\alpha)$ for α close to 1. From the Laurent expansion it follows that $\phi(f^\infty) \geq \phi(h^\infty)$, i.e. policy f^∞ is an average optimal policy.
(ii) Let $R = (g, f, f, \dots)$. Then, if $g(i) = f(i)$, row i of $P(f)$ is identical to row i of $P(g)$ and $r_i(g) = r_i(f)$ i.e. $v_i^\alpha(R) = [r(g) + \alpha P(g)v^\alpha(f^\infty)]_i =$

$[r(f)+\alpha P(f)v^\alpha(f^\infty)]_i = v_i^\alpha(f^\infty)$. If $g(i) \neq f(i)$, then $g(i) \in B(i,f)$, and by formula (4.6) for $h = g$, $v_i^\alpha(f^\infty) - v_i^\alpha(R) = (1-\alpha)^{-1}[\phi(f^\infty) - P(g)\phi(f^\infty)]_i + [u^0(f) - r(g) - P(g)u^0(f) - P(g)\phi(f^\infty)]_i + \epsilon_3(\alpha) < 0$ for α sufficiently close to 1.

Hence, for α sufficiently close to 1, $v^\alpha(R) = r(g)+\alpha P(g)v^\alpha(f^\infty) = L_g v^\alpha(f^\infty) > v^\alpha(f^\infty)$, and, by lemma 17, $v^\alpha(g^\infty) > v^\alpha(f^\infty)$. Then, by the Laurent expansion, it follows that $\phi(g^\infty) \geq \phi(f^\infty)$. ∎

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

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

The unichain case

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

$$B(i,f) = \{a \in A(i) | r_i(a) + \Sigma_j p_{ij}(a)u_j^0 > \phi(f) + u_i^0(f)\} \tag{4.7}$$

The following result can be shown.

Theorem 38 *The linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$, has a unique solution $x = \phi(f^\infty)$ and $y = u^0(f) - u_1^0(f)$.*

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

1. Start with any deterministic policy f^∞.
2. Determine $\phi(f^\infty)$ and $u^0(f)$ as unique solution of the linear system $x \cdot e + [I - P(f)]y = r(f)$ with $y_1 = 0$.
3. Determine for every $i \in E : B(i,f)$ defined in (4.7).

4. If $B(i, f) = \emptyset$ for every $i \in E$: go to step 6.

   Otherwise: take any $g \neq f$ such that, if $g(i) \neq f(i)$, $g(i) \in B(i, f)$.

5. $f = g$ and go to step 2.

6. f^∞ is an average optimal policy.

The concept of policy iteration is originated by Howard [122] who considered the first two parts of the 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 on y the constraint $P^*(f)y = 0$. In order to compute $P^*(f)$ the chain structure of the transition matrix $P(f)$ has to be analysed. The formulation with system (4.4) was proposed by Miller and Veinott [167]. Other anti-cycling rules, which avoid the analysis of the chain structure, are introduced in Schweitzer and Federgruen [216], Federgruen and Spreen [75], and Spreen [233]. Various treatments of the policy iteration method in the unichain case (or other special cases) can be found in Schweitzer [209], Denardo [49], Haviv and Puterman [100], and Lasserre [152].

1.4.4 Linear programming

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

$$\left\{ \begin{array}{lll} v_i & \geq \Sigma_j p_{ij}(a) v_j & \text{for every } (i,a) \in E \times A \\ v_i + u_i & \geq r_i(a) + \Sigma_j p_{ij}(a) u_j & \text{for every } (i,a) \in E \times A \end{array} \right. \tag{4.8}$$

Theorem 39 *The value vector ϕ is the (componentwise) smallest average-superharmonic vector.*

Proof. Let f_0^∞ be a Blackwell optimal policy. From theorem 34 it follows that $\phi_i \geq \Sigma_j p_{ij}(a)\phi_j, (i, a) \in E \times A$, and $\phi_i + u_i^0(f_0) \geq r_i(a) + \Sigma_j p_{ij}(a) u_j^0(f_0)$ for every $i \in E$ and $a \in A(i, \phi)$, where $A(i, \phi) = \{a \in A(i) \mid \phi_i = \Sigma_j p_{ij}(a)\phi_j\}$. Let $A^*(i) = \{a \in A(i) \mid \phi_i + u_i^0(f_0) < r_i(a) + \Sigma_j p_{ij}(a) u_j^0(f_0)\}, i \in E$, and $s_{ia} = \phi_i - \Sigma_j\, p_{ij}(a)\phi_j, t_{ia} = \phi_i + u_i^0(f_0) - r_i(a) + \Sigma_j p_{ij}(a) u_j^0(f_0)$, $a \in A(i)$, $i \in E$. Define $M = \min\{s_{ia}/t_{ia} \mid a \in A^*(i), i \in E\}$ ($M = 0$ if $\cup_{i \in E} A^*(i) = \emptyset$) and $u = u^0(f_0) - M \cdot \phi$. For $a \in A(i, \phi)$, we have $\phi_i = \Sigma_j p_{iaj}\phi_j$ and $\phi_i + u_i = \phi_i + u_i^0(f_0) - M \cdot \phi_i \geq r_i(a) + \Sigma_j p_{ij}(a)[u_j^0(f_0) - M \cdot \phi_j] = r_i(a) + \Sigma_j p_{ij}(a) u_j$. For $a \in A^*(i)$ or $a \notin A(i, \phi) \cup A^*(i)$, we can derive: $\phi_i > \Sigma_j p_{ij}(a)\phi_j$ and $\phi_i + u_i \geq r_i(a) + \Sigma_j p_{ij}(a) u_j$. Hence, ϕ is average-superharmonic. Suppose that y is also average-superharmonic with corresponding x. Then, $y \geq P(f_0)y$, implying that $y \geq P^*(f_0)y \geq P^*(f_0)\{r(f_0) + [P(f_0) - I]x\} = P^*(f_0)r(f_0) = \phi(f_0^\infty) = \phi$, i.e. ϕ is the smallest average-superharmonic vector. ∎

The multichain case

Corollary 40 *Let (u, v) be an optimal solution of the linear program*

$$min\left\{\Sigma_j\beta_j v_j \left| \begin{array}{ll} \Sigma_j[\delta_{ij} - p_{ij}(a)v_j] \geq 0 & ,(i,a) \in E \times A \\ v_i + \ \Sigma_j[\delta_{ij} - p_{ij}(a)u_j] \geq r_i(a) & ,(i,a) \in E \times A \end{array}\right.\right\} \tag{4.9}$$

where $\beta_j > 0$, $j \in E$, is arbitrarily chosen, then $u = \phi$.

The dual program of (4.9) is

$$max\left\{\Sigma_{i,a}r_i(a)x_{ia} \left| \begin{array}{lll} \Sigma_{i,a}[\delta_{ij} - p_{ij}(a)]x_{ia} & = 0 & ,j \in E \\ \Sigma_a x_{ja} + \Sigma_{i,a}[\delta_{ij} - p_{ij}(a)y_{ia} & = \beta_j & ,j \in E \\ x_{ia}, y_{ia} \geq 0, \ (i,a) \in E \times A & & \end{array}\right.\right\} \tag{4.10}$$

For any feasible solution (x, y) of (4.10) we denote by E_x

$$E_x = \{j \in E \mid \Sigma_a x_{ja} > 0\} \tag{4.11}$$

Theorem 41 *Let (x, y) be an extreme optimal solution of (4.10). Then, any policy f^∞ such that* $\begin{cases} x_{if(i)} > 0 & if \quad i \in E_x \\ y_{if(i)} > 0 & if \quad i \notin E_x \end{cases}$ *is an average optimal policy.*

Proof. We will give an outline of the proof. First, we show that f^∞ is well defined, because for every $j \in E$, $\Sigma_a x_{ja} + \Sigma_a y_{ja} = \Sigma_{i,a}p_{ij}(a)y_{ia} + \beta_j > 0$. Since $x_{if(i)} > 0$, $i \in E_x$ and $y_{if(i)} > 0$, $i \notin E_x$, it follows from the complementary slackness of linear programming that $\phi_i + \Sigma_j[\delta_{ij} - p_{ij}(f(i))]u_j = r_i(f)$, $i \in E_x$ and $\Sigma_j[\delta_{ij} - p_{ij}(f(i))]\phi_j = 0$, $i \notin E_x$. Program (4.9) implies that $\Sigma_j[\delta_{ij} - p_{ij}(a)]\phi_j \geq 0$, $i \in E$, $a \in E$. Suppose that $\Sigma_j[\delta_{kj} - p_{kj}(f(k))]\phi_j > 0$ for some $k \in E_x$. Since $x_{kf(k)} > 0$, $\Sigma_j[\delta_{kj} - p_{kj}(f(k))]\phi_j \cdot x_{kf(k)} > 0$. Furthermore, $\Sigma_j[\delta_{ij} - p_{ij}(a)]\phi_j \cdot x_{ia} \geq 0$, $(i,a) \in E \times xA$. Hence, $\Sigma_{i,a}\Sigma_j[\delta_{ij} - p_{ij}(a)\phi_j \cdot x_{ia} > 0$.
On the other hand, this result is contradictory to the constraints of program (4.10) because $\Sigma_{i,a}\Sigma_j[\delta_{ij} - p_{ij}(a)]\phi_j \cdot x_{ia} = \Sigma_j\{[\Sigma_{i,a}[\delta_{ij} - p_{ij}(a)x_{ia}]\}\phi_j = 0$. Thus, we have shown that $\Sigma_j[\delta_{ij} - p_{ij}(f(i))]\phi_j = 0$ for every $i \in E$, i.e. $\phi = P(f)\phi$, and consequently $\phi = P^*(f)\phi$. Next, it can easily be shown

that E_x is closed under $P(f)$, i.e. $p_{ij}(f(i)) = 0$ for $i \in E_x$, $j \notin E_x$. Then, we can prove that the states of $E \backslash E_x$ are transient in the Markov chain induced by $P(f)$ (therefore we use the property that the columns corresponding to the basic variables of a basic solution of an LP are linear independent).
This implies that the columns of $E \backslash E_x$ in $P^*(f)$ are zero. Now, we can finish the proof as follows. For every $k \in E$, we can write,
$\phi_k(f^\infty) = [P^*(f)r(f)]_k = \Sigma_i[P^*(f)]_{ki}r_i(f) = \Sigma_{i \in E_x}[P^*(f)]_{ki}r_i(f) =$
$\Sigma_{i \in E_x}[P^*(f)]_{ki}\{\phi_i + \Sigma_j[\delta_{ij} - p_{ij}(f(i))]u_j\} = [P^*(f)\{\phi + \{(I - P(f))u\}]_k = \phi_k$.
Hence, f^∞ is an average optimal policy. ■

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

1. Take any β with $\beta_j > 0, j \in E$, and compute extreme optimal solutions (u, v) and (x, y) of the dual pair linear programs (4.9) and (4.10) respectively.

2. Choose f^∞ such that $x_{if(i)} > 0$ if $i \in E_x$ and $y_{if(i)} > 0$ if $i \notin E_x$. 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 (4.10) and the stationary policies. However, there are interesting relations. For a feasible solution (x, y) of (4.10) we define a stationary policy $\pi^\infty(x, y)$ by

$$\pi_{ia}(x, y) = \begin{cases} x_{ia}/\Sigma_a x_{ia} & a \in A(i),\ i \in E_x \\ y_{ia}/\Sigma_a y_{ia} & a \in A(i),\ i \notin E_x \end{cases} \tag{4.12}$$

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

$$\begin{cases} x_{ia}(\pi) & = [\Sigma_k \beta_k p^*_{ki}(\pi)] \cdot \pi_{ia} \quad a \in A(i), \quad i \in E \\ y_{ia}(\pi) & = [\Sigma_k \beta_k d_{ki}(\pi) + \Sigma_k \gamma_k p^*_{ki}(\pi)] \cdot \pi_{ia} \quad a \in A(i), \quad i \in E \end{cases} \tag{4.13}$$

with $\gamma_k = max_{i \in E_j}\{-\Sigma_k \beta_k d_{ki}(\pi)/\Sigma_k p^*_{ki}(\pi)\}$, $k \in E_j$, where E_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 ka183 [135]).

Theorem 42 *(i) $(x(\pi), y(\pi))$ is feasible for (4.10) and for a deterministic policy f^∞ $(x(f), y(f))$ is an extreme point of (4.10); (ii) if π^∞ is an average optimal policy, then $(x(\pi), y(\pi))$ is an optimal solution of (4.10) and vice-versa.*

Remarks

- As mentioned before, the only available methods for MDPs with constraints are based on linear programming. In the multichain case, there is no optimal stationary policy, in general. The variables x_{ia} of program (4.10) 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 [117]. An interpretation of the second type of variables, the variables y_{ia}, is not obvious. They are related to the deviation matrix (Kallenberg [135] and can be interpreted as biased deviation measures (Altman and Spieksma [8]). Other contributions in this area, based on the sample path approach, are Ross [196] and Ross and Varadarajan [197]. Beutler and Ross [25] discuss the constrained MDP by a Lagrangean approach. In Altman and Shwartz [4] the sensitivity of constrained MDPs is investigated.

- MDPs with multi-objectives can be treated as constrained MDPs. For this topic we refer to Hordijk and Kallenberg [117], and to Durinovics, Lee, Katehakis and Filar [65].

- Only expected values can be insufficient for a decision maker. It may be desirable to consider also the variability. The last fifteen years dozens of papers on this subject are published. We mention Sobel [229], [230], Kawai and Katoh [144], White [283], [284] and [287], Filar, Kallenberg and Lee [78], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Huang and Kallenberg [124].

Open problem

For MDPs with constraints, an interesting question is find the best policy in the class of stationary policies π^∞ or in the class of deterministic policies f^∞. In the multichain case, no satisfactory algorithm is known for these problems. For the problem in stationary policies, the natural candidate $\pi^\infty(x, y)$ with (x, y) the optimal solution of (4.10) with additional constraints does not satisfy (see Kallenberg [135]).

The unichain case

Since ϕ is a vector with identical components, in the unichain case the property average-superharmonic is equivalent to the existence of a scalar

v and a vector u such that $v + u_i \geq r_i(a) + \Sigma_j p_{ij}(a) u_j$ for every $(i,a) \in E \times$A. Hence, the LP-problem for the smallest average-superharmonic vector becomes

$$min\{v \mid v + \Sigma_j [\delta_{ij} - p_{ij}(a)] u_j \geq r_i(a) \text{ for every } (i,a) \in E \times A\} \qquad (4.14)$$

with dual program

$$max \left\{ \Sigma_{i,a} r_i(a) x_{ia} \;\middle|\; \begin{array}{lll} \Sigma_{i,a}[\delta_{ij} - p_{ij}(a)]\ x_{ia} & = 0, & j \in E \\ \Sigma_{i,a} x_{ia} & = 1 & \\ \qquad\qquad\qquad\qquad x_{ia} & \geq 0, & (i,a) \in E \times A \end{array} \right\} \qquad (4.15)$$

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

1. Compute extreme optimal solutions (u, v) and x of the dual pair LPs (4.14) and (4.15) respectively.

2. Choose f^∞ such that $x_{if(i)} > 0$ if $i \in E_x$ and $f(i)$ arbitrary if $i \notin E_x$. 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 (4.14) satisfies $\Sigma_a x_{ia} > 0$, $i \in E$. Furthermore, the mapping $x_{ia} \rightarrow \pi^\infty(x)$ with $\pi_{ia}(x) = x_{ia}/\Sigma_a x_{ia}$ is a one-to-one mapping of the feasible solutions of (4.14) onto the stationary policies with as inverse mapping $\pi^\infty \rightarrow x_{ia}(\pi)$, where $x_{ia}(\pi) = p_i^*(\pi) \cdot \pi_{ia}$ with $p^*(\pi)$ the equilibrium distribution. The set of deterministic policies corresponds to the set of extreme solutions of (4.14). 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 irredicible case, we refer also to Nazareth and Kulkarni [171].
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 [153]). Furthermore, in the unichain case, problems with constraints have a sulution in the set of stationary policies: if (x, y) is the optimal solution of the LP-problem with constraints, then $\pi^\infty(x, y)$ with $\pi_{ia}(x, y) = x_{ia}/\Sigma_a x_{ia}$, $a \in A(i)$, $i \in E_x$ (and arbitrary decisions in $E \backslash E_x$) is a stationary optimal policy (see Derman [58]).

The pioneering work in solving MDPs by linear programming was made by Manne [165] 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 Denardo [47], who proposed a sequential procedure. Hordijk and Kallenberg [115] have shown that also in the multichain case one linear program suffices. Many results about the linear programming method can be found in Kallenberg [135].

1.4.5 Value iteration

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

$$\begin{cases} v_i^{n+1} = \max_a\{r_i(a) + \Sigma_j p_{ij}(a) v_j^n\},\ i \in E,\ n \geq 0 \\ v_i^0 \text{ arbitrary, } i \in E \end{cases} \tag{4.16}$$

with corresponding policies $f_0^\infty, f_1^\infty, \dots$ 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 sequences $\{v^n \mid n \geq 0\}$ or $\{v^n - v^{n-1} \mid n \geq 1\}$. Since v^n corresponds to total rewards 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 $\{v^n - n \cdot \phi | \ n \geq 0\}$. The next lemma, which appeared in Brown [29], shows that this sequence is bounded. The behaviour of this sequence is also studied by Lanery [151].

Lemma 43 *The sequence* $\{v^n - n \cdot \phi \mid n \geq 0\}$ *is bounded.*
Proof. Let f_*^∞ be a Blackwell optimal policy. From the proof of theorem 39 it follows that (v, u) with $v = \phi(f_*^\infty) = \phi$ and $u = u^0(f_*) - M \cdot \phi(f_*^\infty)$, for the choice of M see the proof of theorem 39, satisfies the superharmonicity inequalities, i.e. $\phi \geq P(f)\phi$ and $\phi + u \geq P(f)u + r(f)$ with equality for $f = f_*$. It is sufficient to show that the sequence $\{e^n = v^n - n \cdot \phi - u \mid n \geq 0\}$ is bounded. For $n \geq 1$, we have,
$P(f_*)e^{n-1} = P(f_*)v^{n-1} - (n-1)\phi - P(f_*)u \leq v^n - r(f_*) - (n-1)\phi + r(f_*) - \phi - u = v^n - n \cdot \phi - u = e^n$, implying that $P^n(f_*)e^0 \leq e^n$. Furthermore, $P(f_{n-1})e^{n-1} \geq P(f_{n-1})v^{n-1} - (n-1)\phi + [r(f_{n-1}) - \phi - u] = v^n - n \cdot \phi - u = e^n$. Hence, $P^n(f_*)e^0 \leq e^n \leq P(f_{n-1})P(f_{n-2}) \cdots P(f_0)e^0$, and consequently $[\min_i e_i^0] \cdot e \leq e^n \leq [\max_i e_i^0] \cdot e, n \geq 0$. ■

Corollary 44 *(i)* $\phi = \lim_{n\to\infty} \frac{1}{n} v^n$; *(ii)* $\phi = \lim_{n\to\infty} \frac{1}{n} \Sigma_{k=1}^n (v^k - v^{k-1})$.

Although corollary 44 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 $\{v^n - n\cdot\phi\}_{n=0}^{\infty}$. 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 [210]. Schweitzer and Federgruen [215] have given necessary and sufficient conditions which guarantee the convergence of the sequence $\{v^n - n \cdot \phi\}_{n=0}^{\infty}$.

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

$$p_{ij}^{\lambda}(a) = \lambda\delta_{ij} + (1-\lambda)p_{ij}(a), i, j \in E \text{ and } a \in A(i) \tag{4.17}$$

Since $p_{ii}^{\lambda}(f) \geq \lambda > 0$, the transition matrix $P^{\lambda}(f)$ is aperiodic. Let $\phi^{\lambda}(f^{\infty})$ be the average reward of policy f^{∞} with respect to the transitions (4.17), then the next lemma shows that $\phi^{\lambda}(f^{\infty}) = \phi(f^{\infty})$. Hence, we may assume that for every f^{∞} the Markov chain with transition matrix $P(f)$ is aperiodic, in which case $P^*(f) = \lim_{n\to\infty} P^n(f)$. The next lemma can be shown by the properties of the stationary and the deviation matrix as mentioned in theorem 28.

Lemma 45 $\phi^{\lambda}(f^{\infty}) = \phi(f^{\infty})$ *for every deterministic policy* f^{∞}.

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

Theorem 46 *Let* $b_{ia} = r_i(a) - \phi_i + \Sigma_j p_{ij}(a)u_j - u_i$; $m_i = \liminf_{n\to\infty} e_i^n$, $i \in E$; $M_i = \limsup_{n\to\infty} e_i^n$, $i \in E$ *and* $A_*(i) = \{a \in A(i) | \phi_i = \Sigma_j p_{ij}(a)\phi_j\}$, $i \in E$. *Then,* $\max_{a\in A_*(i)}\{b_{ia} + \Sigma_j p_{ij}(a)m_j\} \leq m_i \leq M_i \leq \max_{a\in A_*(i)}\{b_{ia} + \Sigma_j p_{ij}(a)M_j\}$.

Let $F_0 = \{f^{\infty} | P(f)\phi = \phi \text{ and } r(f) + P(f)u = \phi + u\}$, where u is the vector mentioned in the proof of theorem 43.

Theorem 47 *Assume that the aperiodicity assumption holds. Then, the sequence* $\{e^n | n \geq 0\}$ *is convergent.*
Proof. Let α and β be two limit vectors of the sequence $\{e^n\}_{n=0}^{\infty}$, and suppose that n_k and m_k satisfy $\alpha_j = \lim_{k\to\infty} e_j^{n_k}$ and $\beta_j = \lim_{k\to\infty} e_j^{m_k}$, $j \in E$. Choose for every k an integer $h(k)$ such that $r_k = m_{h(k)} - n_k \geq k$. From theorem 43(i), we obtain $e^{m_{h(k)}} = e^{r_k+n_k} \geq [P(f)]^{r_k} e^{n_k}$, $k \in N$ Hence, for any policy $f^{\infty} \in F_0$, $\beta = \lim_{k\to\infty} e^{m_{h(k)}} \geq \lim_{k\to\infty}[P(f)]^{r_k} \cdot e^{n_k} \geq \lim_{k\to\infty}[P(f)]^k \cdot e^{n_k} = P^*(f)\alpha$
Similarly we obtain $\alpha \geq P^*(f)\beta$. Since $\alpha \geq P^*(f)\beta \geq P^*(f)\alpha$, we have for every recurrent state j, $\alpha_j = [P^*(f)\alpha]_j$ and similarly $\beta_j = [P^*(f)\beta]_j$.
Therefore, $m_j = \alpha_j \geq [P^*(f)\beta]_j = \beta_j = M_j$ for every recurrent state j, i.e. $m_j = M_j$ for every state which is recurrent under $P(f)$ for some $f^{\infty} \in F_0$.

Let f_* such that $f_*(i) \in A_*(i)$, $i \in E$, and $b(f_*) + P(f_*)M = \max_{E\times A_*}\{b + PM\}$. By theorem 46, $b(f_*) + P(f_*)M \geq M$, implying $P^*(f_*)b(f_*) \geq 0$. Since (ϕ, u) is superharmonic, $b(f_*) \leq 0$, i.e. $b_j(f_*) = 0$ for j recurrent under f_*. Take f^∞ equal to f_*^∞ on the states which are recurrent under f_*, and equal to a policy $f_0^\infty \in F_0$ on the transient states. Then, the states which are recurrent under f_* are a subset of the states which are recurrent under f, i.e. $m_j = M_j$ for the states j recurrent under f_*. By theorem 46, we obtain $b(f_*)+P(f_*)m \leq m \leq M \leq b(f_*)+P(f_*)M$, i.e. $P(f_*)(M-m) \geq M-m \geq 0$, and consequently, $P^*(f_*)(M-m) \geq M-m \geq 0$. Since $m_j = M_j$ for the states j recurrent under f_*, $P^*(f_*)(M-m) = 0$, implying that by $M = m$. ∎

Lemma 48 *Assume that the sequence $\{v^n - n\cdot\phi\}_{n=0}^\infty$ converges. Then, (i) f_n^∞ is average optimal for n sufficiently large; (ii) $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$.*
Proof. (i) Let $v^* = \lim_{n\to\infty}[v^n - n\cdot\phi]$, i.e. $v^* = v^n - n\cdot\phi + \mathcal{O}(1)$. Hence, $v^{n+1} = r(f_n)+P(f_n)[v^*+n\cdot\phi+\mathcal{O}(1)]$. On the other hand, $v^{n+1} = v^* + (n+1)\phi + \mathcal{O}(1)$. From $r(f_n)+P(f_n)v^* + n\cdot P(f_n)\phi + \mathcal{O}(1) = v^* + (n+1)\phi$ for all n, we have for n sufficiently large, $P(f_n)\phi = \phi$ and $r(f_n)+P(f_n)v^* = v^*+\phi$, which implies that $\phi = P^*(f_n)\phi = P^*(f_n)[r(f_n) + P(f_n)v^* - v^*] = \phi(f_n^\infty)$.
(ii) $\phi = (v^{n+1} - v^n) - (e^{n+1} - e^n)$. Since the sequence $\{e^n\}_{n=0}^\infty$ converges, $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$. ∎

The multichain case

Since, for large n, $\phi \approx v^{n+1} - v^n$, $\| v^{n+1} - v^n \|$ and $\text{span}(v^{n+1} - v^n)$ do not provide a valid stopping criterion. If ϕ is constant, no stopping criteria are available. Therefore, another approach is necessary. Schweitzer [211] employs a hierarchical decomposition of the MDP into a set of *communicating* MDPs. This decomposition was proposed by Bather [11]. Schweitzer and Federgruen [217] 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 [218] 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 (4.16) is, in general, no contraction or 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 [208], Denardo [49] and Bather [10]. Surveys on value iteration for undiscounted multichain MDPs can be found in Schweitzer and Federgruen [215] and in Federgruen and Schweitzer [70] and [71].

The unichain case

In this section we 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 48) that $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$. We will formulate an algorithm to compute an ϵ-optimal policy.

Theorem 49 *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)* $l_n \cdot e \leq \phi(f_{n-1}^\infty) \leq \phi_0 \cdot e \leq u_n \cdot e$, $n \geq 1$.

Proof. (i) Since $\phi = \lim_{n\to\infty}(v^{n+1} - v^n)$, it is sufficient to show that $l_{n+1} \geq l_n$ and $u_{n+1} \leq u_n$. We have, $v^{n+1} - v^n \geq [r(f_{n-1}) + P(f_{n-1})v^n] - [r(f_{n-1}) + P(f_{n-1})v^{n-1}] = P(f_{n-1})(v^n - v^{n-1}) \geq P(f_{n-1}) \cdot \min_i(v^n - v^{n-1})_i \cdot e = l_n \cdot e$, i.e. $l_{n+1} \geq l_n$. Similarly, it can be shown that $u_{n+1} \leq u_n$.
(ii) From part (i), we have $u_n \geq \phi_0$, $n \geq 1$. Furthermore, we can write
$\phi(f_{n-1}^\infty) = P^*(f_{n-1})r(f_{n-1}) = P^*(f_{n-1})[v^n - P(f_{n-1})v^{n-1}] = P^*(f_{n-1})[v^n - v^{n-1}] \geq P^*(f_{n-1}) \cdot \min_i(v^n - v^{n-1})_i \cdot e = l_n \cdot e$. ∎

By the results of theorem 49 an algorithm can be formulated. Since v^n grows linearly in n, a direct application of (4.15) may cause numerical difficulties. Therefore, we use the following transformation.
Let $w_i^n = v_i^n - v_N^n$, $i \in E$, $n \geq 0$; $g^n = v_N^n - v_N^{n-1}$, $n \geq 1$. Then, we have $w_i^n = [e_i^n + n\phi + u_i] - [e_N^n + n\phi + u_N] = [e_i^n - e_N^n] + [u_i - u_N]$, which is a bounded sequence, and $g^n = [e_N^n + n\phi + u_N] - [e_N^{n-1} + (n-1)\phi + u_N] = [e_N^n - e_N^{n-1}] + \phi$, which is also bounded. Furthermore, the relations become $g^{n+1} = v_N^{n+1} - v_N^n = \max_{a\in A(N)}\{r_N(a) + \Sigma_j p_{Nj}(a)[v_j^n - v_N^n]\} = \max_{a\in A(N)}\{r_N(a) + \Sigma_j p_{Nj}(a)w_j^n\}$, and $w_i^{n+1} = v_i^{n+1} - v_N^{n+1} = \max_{a\in A(i)}\{r_i(a) + \Sigma_j p_{ij}(a) \cdot [v_j^n - v_N^n]\} + [v_N^n - v_N^{n+1}] = \max_{a\in A(i)}\{r_i(a) + \Sigma_j p_{ij}(a)w_j^n\} - g^{n+1}$, $i \in E$.
For the bounds l_n and u_n, we have $l_n = \min_i(v_i^n - v_i^{n-1}) = \min_i(w_i^n - w_i^{n-1}) + g^n$ and $u_n = \max_i(v_i^n - v_i^{n-1}) = \max_i(w_i^n - w_i^{n-1}) + g^n$.

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

1. Choose $\epsilon > 0$, and take $v \in \mathbb{R}^N$ arbitrarily.

2. Compute:

   a. $s_{ia} = r_i(a) + \Sigma_j p_{ij}(a)v_j$, $(i, a) \in E \times xA$;

b. $g = \max_{a \in A(N)} s_{Na}$;

c. $w_i = \max_{a \in A(i)} s_{ia} - g$, $i \in E$ 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 $\frac{1}{2}\epsilon$-approxiation 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 formula (3.7) of the discounted rewards, it can be shown that an action $a \in A(i)$ is *suboptimal* if

$$\phi + u_i > r_i(a) + \Sigma_j p_{ij}(a) u_j, \tag{4.18}$$

where (ϕ, u) is a solution of the optimality equation of theorem 35. 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 49 provides bounds for ϕ; however, bounds for u are unknown. One may well apply a suboptimality test in one iteration of formula (4.16). 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 [125]).

Bounds on the value vector as formulated in theorem 49 can be found in Hastings [95], Odoni [173], Hordijk and Tijms [119], and Platzman [178]. Hordijk and Tijms [121] have proposed an approximation method with a sequence of discounted value iterations with discount factors tending to 1. Algorithm X is established by White [278]. 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.

1.4.6 Modified policy iteration

As in the discounted reward case, modified policy iteration can be applied. However, we assume in this section the *strong aperiodicity assumption*, i.e. for some $0 < \lambda < 1$, $p_{ii}(a) \geq \lambda > 0$ for all $i \in E$, $a \in A(i)$. As shown in section 4.5 by Schweitzer's aperiodicity transformation, any MDP can be transformed to an equivalent MDP with the strong aperiodicity property.

Furthermore, we assume that the value vector ϕ is a constant vector: $\phi = \phi_0 \cdot e$.

Let U and L_f be the operators as defined in (3.2) and (3.5), repectively, with $\alpha = 1$.

Lemma 50 *Let $l_n = \min_i (Ux^n - x^n)_i$, $n \in \mathbb{N}$. Then, the sequence $\{l_n\}_{n=1}^{\infty}$ is monotonically nondecreasing.*
Proof. $Ux^n - x^n \geq L_{f_{n-1}} x^n - x^n = L^{k+1}_{f_{n-1}} x^{n-1} - L^k_{f_{n-1}} x^{n-1} = P^k(f_{n-1})$ $[L_{f_{n-1}} x^{n-1} - x^{n-1}] = P^k(f_{n-1})[Ux^{n-1} - x^{n-1}] \geq P^k(f_{n-1}) l_{n-1} \cdot e = l_{n-1} \cdot e$. Hence, $l_n \geq l_{n-1}$. ∎

Remark
Let $u_n = \max_i (Ux^n - x^n)_i$, $n \in \mathbb{N}$. Then the sequence $\{u_n \mid n \in \mathbb{N}\}$ is in general not monotonically nonincreasing, unless $k = 1$ (for $k = 1$, see theorem 49).

Without proof we present the following result.

Theorem 51 *(i) The sequences $\{l_n \mid n \in \mathbb{N}\}$ and $\{u_n \mid n \in \mathbb{N}\}$ both 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_n and $\frac{1}{2}[u_n + l_n]$ is an $\frac{1}{2}\epsilon$-approximation of ϕ_0.*

Algorithm XI (modified policy iteration; average rewards; 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 $L_f x = Ux$;

   c. Let $l = \min_i (Ux - x)_i$ and $u = \max_i (Ux - x)_i$.
3. If $u - l \leq \epsilon :$ f^{∞} is an ϵ-optimal policy and $(u+l)/2$ is a $\frac{1}{2}\epsilon$-approxiation of the value ϕ_0 (STOP);

   otherwise: $x = L^k_f x$ and go to step 2.

Remark
If $k = 1$ the method becomes the standard value iteration (without White's relative values). We will also make it plausible that, in the unichain case, policy iteration corresponds to $k = \infty$. By theorem 28, we have $\phi(f^{\infty}) + [I - P(f)]u(f) = r(f)$, $L^k_{f_n} x^n = L^k_{f_n}[u(f_n)] + P^k(f_n)[x^n - u(f_n)] = u(f_n) + k \cdot \phi(f_n^{\infty}) + P^k(f_n)[x^n - u(f_{n+1})]$. 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, $\phi(f_n^\infty)$ is also a constant vector, the difference between $L_{f_n}^k x^n$ 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 $r_i(a) + \Sigma_j p_{ij}(a) u_j(f_n)$, which is the same as $U[L_{f_n}^k x^n]$. Hence, both methods are very similar.

The modified policy iteration method was first mentioned by Morton [170]. Vander Wal [246] and [247] has analysed this method extensively under various chain structure assumptions (irreducible case, unichain case, communicating case and simply connected case).

1.5 MORE SENSITIVE OPTIMALITY CRITERIA

1.5.1 Introduction

In section 1.3 the concepts of an *n-discount optimal* and an *n-average optimal policy* R_*, for $n = -1, 0, 1, \ldots$, are introduced, respectively by

$$\lim_{\alpha \uparrow 1} (1-\alpha)^{-n} [v^\alpha(R_*) - v^\alpha] = 0 \tag{5.1}$$

and

$$\liminf_{T \to \infty} \frac{1}{T} [v^{n,T}(R_*) - v^{n,T}(R)] \geq 0 \text{ for every policy } R, \tag{5.2}$$

where

$$v^{n,t}(R) = \begin{cases} v^t(R) & \text{for } n = -1 \\ \Sigma_{s=1}^t v^{n-1,s}(R) & \text{for } n = 0, 1, \ldots \end{cases}$$

Without proof we state that for both criteria optimal deterministic policies exist.

Theorem 52 *The criteria (-1)-discount optimal and (-1)-average optimal are equivalent, and also equivalent to average optimal.*

Proof. From corollary 31 it follows that $\lim_{\alpha \uparrow 1} (1-\alpha) v^\alpha(f^\infty) = \phi(f^\infty)$ for every deterministic policy. Hence, using a Blackwell optimal policy, we obtain $\lim_{\alpha \uparrow 1} (1-\alpha) v^\alpha = \phi$. Therefore, if f_*^∞ is (-1)-discount optimal, then we have $0 = \lim_{\alpha \uparrow 1} (1-\alpha)[v^\alpha(f_*^\infty) - v^\alpha] = \phi(f_*^\infty) - \phi$, i.e. f_*^∞ is average optimal.

(-1)-average optimality means that $\lim_{T \to \infty} \frac{1}{T}[v^T(f_*) - v^T(f)] \geq 0$, i.e. $\lim_{T \to \infty} \frac{1}{T}[\Sigma_{t=1}^T P^{t-1}(f_*) r(f_*) - \Sigma_{t=1}^T P^{t-1}(f) r(f)] \geq 0$ for every policy f^∞. Since $\lim_{T \to \infty} \frac{1}{T} \Sigma_{t=1}^T P^{t-1}(f) r(f) = \phi(f^\infty)$ for every policy f^∞, an average optimal policy is also (-1)-average optimal. Finally, let f_*^∞ be (-1)-average optimal and let f_0^∞ be a Blackwell optimal policy. Then, by the Laurent expansion, $0 \geq \lim_{\alpha\uparrow 1}(1-\alpha)[v^\alpha(f_*^\infty)-v^\alpha] = \lim_{\alpha\uparrow 1}(1-\alpha)[v^\alpha(f_*^\infty)-v^\alpha(f_0^\infty)] = \phi(f_*^\infty)-\phi(f_0^\infty) = \lim_{T\to\infty}\frac{1}{T}[\Sigma_{t=1}^T P^{t-1}(f_*)r(f_*) - \Sigma_{t=1}^T P^{t-1}(f_0)r(f_0)] \geq 0$, i.e. f_*^∞ (-1)-discount optimal. ■

Theorem 53 *The criteria 0-discount optimal and 0-average optimal are equivalent (such policy is called a bias optimal policy).*

Proof. Suppose that f_*^∞ is a 0-optimal policy, and let f_0^∞ be Blackwell optimal. Using the properties of the stationary and the deviation matrix, it can be shown that $\Sigma_{s=1}^t P^{s-1}(f)r(f) = t\phi(f^\infty)+u^0(f)-P^t(f)D(f)r(f)$. Therefore, $0 \leq \liminf_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T\Sigma_{s=1}^t\{P^{s-1}(f_*)r(f_*) - P^{s-1}(f_0)r(f_0)\} = \liminf_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T\{[t\phi(f_*^\infty)+u^0(f_*)-P^t(f_*)D(f_*)r(f_*)]-[t\phi(f_0^\infty)+u^0(f_0)-P^t(f_0)D(f_0)r(f_0)]\} = \liminf_{T\to\infty}[\frac{1}{2}(T+1)\{\phi(f_*^\infty)-\phi\}+\{u^0(f_*)-u^0(f_0)\} - \frac{1}{T}\Sigma_{t=1}^T\{P^t(f_*)D(f_*)r(f_*) - P^t(f_0)D(f_0)r(f_0)\}]$. Since $\lim_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T P^t(f)$ $D(f) = P^*(f)D(f) = 0$ and $\phi(f_*^\infty)-\phi \leq 0$, we have $\phi(f_*^\infty) = \phi = \phi(f_0^\infty)$ and $u^0(f_*)-u^0(f_0) \geq 0$.
Hence, we obtain, by the Laurent expansion, $0 \geq \lim_{\alpha\uparrow 1}[v^\alpha(f_*^\infty)-v^\alpha] = \lim_{\alpha\uparrow 1}[v^\alpha(f_*^\infty)-v^\alpha(f_0^\infty)] = \lim_{\alpha\uparrow 1}\{(1-\alpha)^{-1}[\phi(f_*^\infty)-\phi(f_0^\infty)]+[u^0(f_*)-u^0(f_0)]\} = u^0(f_*)-u^0(f_0) \geq 0$. On the other hand, suppose that $\lim_{\alpha\uparrow 1}[v^\alpha$ $(f_*^\infty)-v^\alpha] = 0$. For any f^∞, also by the Laurent expansion, $\phi(f_*^\infty) \geq \phi(f^\infty)$ and if $\phi_i(f_*^\infty) = \phi_i(f^\infty)$, then $u_i^0(f_*) \geq u_i^0(f)$. Hence, we obtain $\liminf_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T\Sigma_{s=1}^t\{P^{s-1}(f_*)r(f_*)-P^{s-1}(f)r(f)\} = \liminf_{T\to\infty}[\frac{1}{2}(T+1)\{\phi(f_*^\infty)-\phi(f^\infty)\}+\{u^0(f_*)-u^0(f)\}+\frac{1}{T}\Sigma_{t=1}^T\{P^t(f_*)D(f_*)r(f_*)-P^t(f_0)$ $D(f_0)r(f_0)\}] \geq 0$. ■

In Blackwell [27] the concept of bias optimality was introduced. Veinott [257] presented a policy iteration algorithm for finding a bias optimal policy. In Veinott [257] is also shown that an *average overtaking* deterministic optimal policy is bias optimal, and conjectured that the reverse statement is also true. A policy R_* is average overtaking optimal if $\liminf_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T[v^t(R_*)-v^t(R)] \geq 0$ for every policy R. The conjecture was proved by Denardo and Miller [52]. Lippman [155] 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 [134]. We refer also to Rieder's chapter in this book on bias optimality. Denardo and Rothblum [54] have studied the stronger criterion of *overtaking optimality.* A policy R_* is overtaking optimal if $\liminf_{T\to\infty}\ \Sigma_{t=1}^T[v^t(R_*)-v^t(R)] \geq 0$ for every policy R. This criterion may have no optimal policy, 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 vei69 [258]. Sladky sla [224] has introduced the concept of n-average optimality; furthermore, he showed the equivalence between this criterion and the n-discount optimality.

1.5.2 Lexicographic ordering of Laurent series

Instead of the discount factor α we can also use the interest rate ρ, where the relations between α and ρ are given by $\rho = (1-\alpha)/\alpha$. From theorem 30 it follows that the total discounted reward $v^\alpha(f^\infty)$ can be written as the Laurent series $v^\rho(f^\infty) = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k u^k(f)$ for $0 < \rho < \| D(f) \|^{-1}$. Let

$$H^\rho(f) = (1+\rho)\{P^*(f) + \Sigma_{k=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f)\} \text{ for } 0 < \rho < \| D(f) \|^{-1}. \tag{5.3}$$

Theorem 54 *(i)* $H^\rho(f) = \rho[I-(1+\rho)^{-1}P(f)]^{-1}$; *(ii)* $H^\rho(f)r(f) = \rho v^\rho(f^\infty)$; *(iii)* $\rho[v^\rho(f^\infty) - x] = H^\rho(f)[r(f) + (1+\rho)^{-1}P(f)x - x]$.
Proof. The proof is based on the properties of the stationary and deviation matrix as stated in theorem 28. ∎

Define the sets LS_1 and LS_2 of Laurent series by

$$LS_1 = \{u(\rho) | u(\rho) = \Sigma_{k=-1}^{\infty}\rho^k u^k;\ u^k \in \mathbb{R}^N, k \geq -1; \\ \limsup_{k\to\infty} \| u^k \|^{1/k} < \infty\} \tag{5.4}$$

and

$$LS_2 = \left\{ x(\rho) \;\middle|\; \begin{array}{l} x(\rho) = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k x^k;\ x^k \in \mathbb{R}^N,\ k \geq 1 \\ \limsup_{k\to\infty} \| x^k \|^{1/k} < \infty \end{array} \right\} \tag{5.5}$$

Lemma 55 $LS_1 = LS_2$.
Proof. The proof can be given by showing that any $u(\rho) \in LS_1$ satisfies $u(\rho) = \Sigma_{k=-1}^{\infty}\rho^k u^k = (1+\rho)[\Sigma_{j=0}^{\infty}(-\rho)^j]\Sigma_{k=-1}^{\infty}\rho^k u^k = (1+\rho)\Sigma_{k=-1}^{\infty}\rho^k x^k$, where $x^k = \Sigma_{j=-1}^{k}(-1)^{k-j}u^j$. Hence, $u(\rho)$ is also an element of LS_2. Similarly, the reverse statement can be shown. ∎

Corollary 56. *The sets LS_1 and LS_2 are the same, and will be denoted by LS.*

Notice that LS is a linear vector space. We define a *lexicographic ordering* on LS by: $u(\rho)$ is *nonnegative (positive)* if the first nonzero vector of $(u^{-1}, u^0, \dots)$ is nonnegative (positive), i.e.

$$\begin{cases} u(\rho) \geq_\ell 0 & \text{if } \liminf_{\rho\downarrow 0}\rho^{-k}u(\rho) \geq 0 \text{ for } k = -1, 0, 1, \dots \\ u(\rho) >_\ell 0 & \text{if } u(\rho) \geq_\ell 0 \text{ and } u(\rho) \neq 0. \end{cases} \tag{5.6}$$

Let $L_f^\rho : LS \to LS$ be defined by

$$L_f^\rho x(\rho) = r(f) + (1+\rho)^{-1} P(f) x(\rho). \tag{5.7}$$

The Laurent expansion of $L_f^\rho x(\rho) - x(\rho)$ for $x(\rho) = (1+\rho)\Sigma_{k=-1}^\infty \rho^k x^k$ satisfies $L_f^\rho x(\rho) - x(\rho) = r(f) + (1+\rho)^{-1} P(f)(1+\rho)\Sigma_{k=-1}^\infty \rho^k x^k - (1+\rho)\Sigma_{k=-1}^\infty \rho^k x^k = r(f) + \Sigma_{k=-1}^\infty \rho^k P(f) x^k - \Sigma_{k=-1}^\infty \rho^k x^k - \Sigma_{k=0}^\infty \rho^k x^{k-1} = \rho^{-1}[P(f)x^{-1} - x^{-1}] + [r(f) + P(f)x^0 - x^0 - x^{-1}] + \Sigma_{k=1}^\infty \rho^k [P(f)x^k - x^k - x^{k-1}]$, which implies $L_f^\rho x(\rho) \in LS$. From theorem 11, we obtain

$$L_f^\rho v^\rho(f^\infty) - v^\rho(f^\infty) = 0. \tag{5.8}$$

For $x \in \mathbb{R}^N$, let $y = \max_{E\times A}[r + Px]$, i.e. $y_i = \max_{a\in A(i)}[r_i(a) + \Sigma_j p_{ij}(a) x_j]$, $i \in E$, and $\text{argmax}_{E\times A}[r+Px] = \{f \mid \max_{E\times A}[r+Px] = r(f) + P(f)x\}$.

Consider the mapping $B : LS \to LS$ where, for $x(\rho) = (1+\rho)\Sigma_{k=-1}^\infty \rho^k x^k$, $Bx(\rho)$ is given by

$$Bx(\rho) = \Sigma_{k=-1}^\infty \rho^k B^{(k)}(x^{-1}, x^0, \dots, x^k) \tag{5.9a}$$

with

$$\begin{cases} B^{(-1)}(x^{-1}) &= \max\limits_{E\times A}[Px^{-1} - x^{-1}] \\ A^{(-1)}(x^{-1}) &= \operatorname*{argmax}\limits_{E\times A}[Px^{-1} - x^{-1}] \end{cases} \tag{5.9b}$$

$$\begin{cases} B^{(0)}(x^{-1}, x^0) &= \max\limits_{E\times A^{-1}(x^{-1})}[r + Px^0 - x^0 - x^{-1}] \\ A^{(0)}(x^{-1}, x^0) &= \operatorname*{argmax}\limits_{E\times A^{-1}(x^{-1})}[r + Px^0 - x^0 - x^{-1}] \end{cases} \tag{5.9c}$$

and for $k \geq 1$ we define $B^{(k)}(x^{-1}, x^0, \dots, x^k)$ and $A^{(k)}(x^{-1}, x^0, \dots, x^k)$ by

$$\begin{cases} B^{(k)}(x^{-1}, x^0, \dots, k) &= \max\limits_{E\times A^{k-1}(x^{-1}, x^0, \dots, x^k)}[Px^k - x^k - x^{k-1}] \\ A^{(k)}(x^{-1}, x^0, \dots, k) &= \operatorname*{argmax}\limits_{E\times A^{k-1}(x^{-1}, x^0, \dots, k)}[Px^k - x^k - x^{k-1}] \end{cases} \tag{5.9d}$$

Because
$\rho^{-1}[P(f)x^{-1} - x^{-1}] + [r(f) + P(f)x^0 - x^0 - x^{-1}] + \Sigma_{k=1}^\infty \rho^k [P(f)x^k - x^k - x^{k-1}] = r(f) + \Sigma_{k=-1}^\infty \rho^k P(f) x^k - (1+\rho)\Sigma_{k=-1}^\infty \rho^k x^k = r(f) + (1+\rho)^{-1} P(f) x(\rho) - x(\rho) = L_f^\rho x(\rho) - x(\rho), Bx(\rho) \in LS$, which is the result of lexicographically maximizing $r(f) + (1+\rho)^{-1}P(f)x(\rho) - x(\rho)$, i.e. $Bx(\rho) = \text{lexmax}_f[r(f) + (1+\rho)^{-1}P(f)x(\rho) - x(\rho)] = lexmax_f[L_f^\rho x(\rho) - x(\rho)]$
Since $L_g^\rho[v^\rho(g^\infty)] - v^\rho(g^\infty) = 0$, we have for every g^∞

$$Bv^\rho(g^\infty) = \underset{f}{\text{lexmax}}[L_f^\rho v^\rho(g^\infty) - v^\rho(g^\infty)] \geq_\ell L_g^\rho v^\rho(g^\infty) - v^\rho(g^\infty) = 0. \quad (5.10)$$

$$H^\rho(f) \text{ is a } \textit{positive operator} \text{ if } \begin{cases} H^\rho(f)[u(\rho)] \geq_\ell 0 & \text{when } u(\rho)_\ell \geq 0 \\ H^\rho(f)[u(\rho)] >_\ell 0 & \text{when } u(\rho)_\ell > 0 \end{cases} \quad (5.11)$$

Then, the following results can be proven.

Theorem 57 *(i)* $H^\rho(g)[r(g) + (1+\rho)^{-1}P(g)v^\rho(f^\infty) - v^\rho(f^\infty)] = \rho[v^\rho(g^\infty) - v^\rho(f^\infty)]$; *(ii)* $H^\rho(f)$ *is a positive operator.*

Theorem 58 *(i)* $Bx = 0$ *has in* LS *a unique solution* $x = v^\rho(f_0^\infty)$*, where* f_0^∞ *is Blackwell optimal; (ii) If* g *satisfies* $Bx = r(g) + (1+\rho)^{-1}P(g)x - x = 0$*, then* g^∞ *is a Blackwell optimal policy.*

The representation of the expected discounted rewards as a Laurent series is provided by Miller and Veinott [167]. They also introduced the lexicographic ordering. The approach of the mapping B on the set LS is developed by Hordijk and Dekker [113].

1.5.3 n-Discount optimality and policy iteration

In this section we derive a policy iteration algorithm to compute a policy that lexicographically maximizes the vector $(u^{-1}(f), u^0(f), \ldots, u^n(f))$ over all deterministic policies, i.e. an n-discount optimal policy, for $n = -1, 0, 1, \ldots$. For $n = -1$ an average optimal policy and for $n = 0$ a bias optimal policy is obtained. Furthermore, we will show that for all $n \geq N - 1$ an n-discount optimal policy is a Blackwell optimal policy. The algorithm is as follows.

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 $\max_{E \times A}[Pu^{-1}(f) - u^{-1}(f)] > 0$, then $A^{(-1)} = \text{argmax}_{E \times A}[Pu^{-1}(f) - u^{-1}(f)]$, choose g from $A^{(-1)}$ and go to step 5.

b If $\max_{E\times A^{(-1)}}[r + Pu^0(f) - u^0(f) - u^{-1}(f)] > 0$, then
$A^{(0)} = \operatorname{argmax}_{E\times A^{(-1)}}[r + Pu^0(f) - u^0(f) - u^{-1}(f)]$, choose g from $A^{(0)}$ and go to step 5.

c For $k = 0$ until n do:
If $\max_{E\times A^{(k)}}[Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)] > 0$, then
$A^{(k+1)} = \operatorname{argmax}_{E\times A^{(k)}}[Pu^{k+1}(f) - u^{k+1}(f) - u^k(f)]$, choose g from $A^{(k+1)}$ and go to step 5.

4. f^∞ is n-discount optimal (STOP);
5. $f(i) = g(i)$, $i \in E$, and go to step 2.

Theorem 59 *The linear system*

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

has the unique solution $(u^{-1}(f), u^0(f), \dots, u^{n+1}(f))$.
Proof. Suppose that $(x^{-1}, x^0, \dots, x^{n+1})$ is a solution. Then, $x^{-1} = P(f)x^{-1}$ implies that $x^{-1} = P^*(f)x^{-1}$. From $x^0 + [I - P(f)]x^1 = 0$ we obtain $P^*(f)x^0 = 0$.
Hence, $x^{-1} = P^*(f)x^{-1} = P^*(f)\{r(f) - [I - P(f)]x^0\} = P^*(f)r(f) = u^{-1}(f)$ and $r(f) = x^{-1}+[I-P(f)]x^0 = P^*(f)r(f)+[I-P(f)+P^*(f)]x^0$, i.e. $x^0 = [I-P(f)+P^*(f)]^{-1}[I-P^*(f)]r(f) = [D(f)+P^*(f)][I-P^*(f)]r(f) = u^0(f)$. By induction on k, we obtain $x^{k-1} = u^{k-1}(f)$, $P^*(f)x^k = 0$ and $0 = x^{k-1}+[I-P(f)]x^k = u^{k-1}(f)+[I-P(f)+P^*(f)]x^k = (-1)^{k-1}D^k(f)r(f)+[I - P(f) + P^*(f)]x^k$ i.e.
$x^k = [D(f) + P^*(f)](-1)^k D^k(f)r(f) = (-1)^k D^{k+1}(f)r(f) = u^k(f)$, $k \ge 1$.∎

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.

In order to show that algorithm XII determines an n-discount optimal policy, we use the following notation and lemma (which is stated without proof).

$$\left\{ \begin{array}{ll} \psi^{-1}(f,g) & = P(g)u^{-1}(f) - u^{-1}(f) \\ \psi^0(f,g) & = r(g) + P(g)u^0(f) - u^0(f) - u^{-1}(f) \\ \psi^k(f,g) & = P(g)u^k(f) - u^k(f) - u^{k-1}(f), \;\; k \ge 1 \end{array} \right. \tag{5.12}$$

Lemma 60 *For every f^∞, g^∞ and every $m \in \mathbb{N}$, we have $\alpha v^\alpha(g^\infty) = \Sigma_{k=-1}^{m-1}\rho^k\{u^k(f) + \Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\} + \rho^m\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)u^{m-1}(f)$.*

Theorem 61 *Let f^∞ and g^∞ be subsequent policies in algorithm XII, then $v^\rho(g^\infty) > v^\rho(f^\infty)$ for ρ sufficiently small.*

Proof. From theorem 59 it follows that $\psi^k(f,f) = 0$ for $k = -1, 0, \dots$. Hence, by lemma 60 with $n = m+2$, $v^\rho(g^\infty) - v^\rho(f^\infty) = (1+\rho)\{\Sigma_{k=-1}^{n+1}\rho^k [\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)] + \rho^{n+2}\Sigma_{t=1}^{\infty}\alpha^t[P^{t-1}(f) - P^{t-1}(g)]u^{n+1}(f)\}$. Since $\| \rho^{n+2}\Sigma_{t=1}^{\infty}\alpha^t[P^{t-1}(f) - P^{t-1}(g)]u^{n+1}(f) \| \leq \rho^{n+2}(1-\alpha)^{-1} \| P^{t-1}(f) - P^{t-1}(g) \| \cdot \| u^{n+1}(f) \| \leq 2\rho^{n+1}(1+\rho) \| u^{n+1}(f) \|$, $\lim_{\rho\downarrow 0}(1+\rho)\rho^{n+2} \Sigma_{t=1}^{\infty}\alpha^t[P^{t-1}(f) - P^{t-1}(g)]u^{n+1}(f) = 0$ for $n \geq 0$. On the other hand, since f^∞ and g^∞ are subsequent policies, $\psi^k(f,g) = 0, -1 \leq k \leq m-1$ and $\psi^m(f,g) > 0$ for some $-1 \leq m \leq n+1$. Hence, if we define $\psi^k(f,g) = 0$ for $k \geq n+2$, we have $\Sigma_{k=-1}^{\infty}\rho^k\psi^k(f,g) \in LS$ and $\Sigma_{k=-1}^{\infty}\rho^k\psi^k(f,g) >_\ell 0$. Since, by theorem 57 (ii), $H^\rho(g) = \rho[I-(1+\rho)^{-1}P(g)]^{-1}$ is a positive operator, $[I-(1+\rho)^{-1}P(g)]^{-1}$ is also positive, i.e. $[I-(1+\rho)^{-1}P(g)]^{-1}[\Sigma_{k=-1}^{\infty}\rho^k\psi^k(f,g)] >_\ell 0$. Since $[I-(1+\rho)^{-1}P(g)]^{-1}[\Sigma_{k=-1}^{\infty}\rho^k\psi^k(f,g)] = \Sigma_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(g)[\Sigma_{k=-1}^{\infty}\rho^k \psi^k(f,g)] = \Sigma_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(g)\Sigma_{k=-1}^{n+1}\rho^k\psi^k(f,g) = \alpha^{-1}\{\Sigma_{k=-1}^{n+1}\rho^k[\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)]\} = (1+\rho)\{\Sigma_{k=-1}^{n+1}\rho^k[\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)]\}$, it follows that $v^\rho(g^\infty) - v^\rho(f^\infty) = (1+\rho)\{\Sigma_{k=-1}^{n+1}\rho^k[\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)]\} + (1+\rho)\rho^{n+2}\Sigma_{t=1}^{\infty}\alpha^t \cdot [P^{t-1}(f) - P^{t-1}(g)]u^{n+1}(f)\} >_\ell 0$: $v^\rho(g^\infty) > v^\rho(f^\infty)$ for ρ sufficiently small. ∎

Theorem 62 *Algorithm XII terminates in a finite number of iterations with an n-discount optimal policy.*

Proof. From theorem 61 it follows that algorithm XII produces a sequence of different deterministic policies. Since the set of deterministic policies is finite, the algorithm terminates after a finite number of iterations. Let f^∞ be the last policy. We obtain by applying lemma 60, $v^\rho(g^\infty) - v^\rho(f^\infty) = (1+\rho)\{\Sigma_{k=-1}^{n+1}\rho^k[\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)] + \rho^{n+2}\Sigma_{t=1}^{\infty}\alpha^t[P^{t-1}(f) - P^{t-1}(g)] u^{n+1}(f)\}$. Since the algorithm terminates, $\Sigma_{k=-1}^{\infty}\ \rho^k\psi^k(f,g)\ \leq_\ell\ 0$, where $\psi^k(f,g) = 0$ for $k \geq n+2$. This implies that $(1+\rho)\{\Sigma_{k=-1}^{n+1}\rho^k[\Sigma_{t=1}^{\infty}\alpha^t P^{t-1}(g) \psi^k(f,g)] \leq_\ell 0$ and $\| \rho^{n+2}\Sigma_{t=1}^{\infty}\alpha^t[P^{t-1}(f) - P^{t-1}(g)]u^{n+1}(f) \| \leq 2\rho^{n+1}(1+\rho) \| u^{n+1}(f) \|$. Hence, $\lim_{\rho\downarrow 0}\rho^{-n}[v^\rho(f^\infty) - v^\rho(g^\infty)] \geq 0$ for every policy g^∞, i.e. f^∞ is an n-discount optimal policy. ∎

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

Lemma 63 *If $\psi^k(f,g) = 0$ for $1 \leq k \leq N$, then also $\psi^k(f,g) = 0$ for all $k \geq N+1$.*

Proof. Let $L = \{x \in \mathbb{R}^N | [P(f) - P(g)]x = 0\}$. Since, for $k \geq 1$, $\psi^k(f,g) = P(g)u^k(f) - u^k(f) - u^{k-1}(f) = P(g)u^k(f) - (-1)^k[D(f) - I]D^k(f)r(f) = P(g)u^k(f) - (-1)^k[P(f)D(f) - P^*(f)]D^k(f)r(f) = P(g)u^k(f) - P(f)u^k(f)$,i.e. $u^k(f) \in L$ for $k = 1, 2, \ldots, N$. Since L is a linear vector space in $\mathbb{R}^N$, the $N+1$ vectors $u^k(f)$, $1 \leq k \leq N+1$, are linear dependent. Because $u^k(f) = B^k x_0$ for $x_0 = u^1(f)$ and $B = -D(f)$, the set $\{x_0, Bx_0, \ldots, B^N x_0\}$ is dependent, i.e. for some $1 \leq k \leq N$, we have $B^k x_0 = \Sigma_{j=0}^{k-1} \lambda_j B^j x_0$ for some scalars λ_j, $1 \leq j \leq k-1$, i.e. $B^k x_0 \in L$. Hence we have $B^{k+1}x_0 = \Sigma_{j=0}^{k-1} \lambda_j B^{j+1} x_0$, which is a linear combination of the elements $Bx_0, \ldots, B^k x_0$ from L, so $B^{k+1}x_0 \in L$. Similarly, by induction, it can be shown that $u^k(f) = B^k x_0 \in L$ for every $k \geq 1$, implying that $\psi^k(f,g) = 0$ for $k \geq 1$. ∎

Theorem 64 *If algorithm XII is used to determine an $(N-1)$-discount optimal policy f^∞ then f^∞ is also a Blackwell optimal policy.*

Proof. If the algorithm terminates with f^∞, we have $\Sigma_{k=-1}^{N} \rho^k \psi^k(f,g) \leq_\ell 0$ for every policy g^∞, i.e. either $\Sigma_{k=-1}^{N} \rho^k \psi^k(f,g) <_\ell 0$ or $\Sigma_{k=-1}^{N} \rho^k \psi^k(f,g) = 0$. In the first case, $v^\rho(g^\infty) < v^\rho(f^\infty)$ for ρ sufficiently small; in the second case $\psi^k(f,g) = 0$, $1 \leq k \leq N$. From lemma 63 it follows that $\psi^k(f,g) = 0$, $k \geq 1$. Then, by lemma 60 with $m \to \infty$, $\alpha v^\alpha(g^\infty) = \Sigma_{k=-1}^{\infty} \rho^k \{u^k(f) + \Sigma_{t=1}^{\infty} \alpha^t P^{t-1}(g) \psi^k(f,g)\} = \Sigma_{k=-1}^{\infty} \rho^k u^k(f) = \alpha v^\alpha(f^\infty)$ for all $\alpha \in [0,1)$. Hence, f^∞ is Blackwell optimal. ∎

The policy iteration method of this section was proposed in Veinott [258] and in Miller and Veinott [167]. They have also shown that Blackwell optimality is the same as n-discount optimality for $n \geq N-1$. In Veinott [259] 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.

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

$$\mathrm{P}(\mathbb{R}) = \{p(x)|p(x) = a_0 + a_1x + \cdots + a_nx^n, a_i \in \mathbb{R}, 1 \leq i \leq n\}. \quad (5.13)$$

By p_0 and p_1 we denote the polynomials $p_0(x) = 0$ and $p_1(x) = 1$ for every x. The *field* $F(\mathbb{R})$ of rational functions with real coefficients consists of the elements $\frac{p(x)}{q(x)}$, where p and q are from $P(\mathbb{R})$ 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 ordereing in $F(\mathbb{R})$ is obtained by $\frac{p}{q} >_\ell p_0$ if and only if $d(p)d(q) > 0$. 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 $\frac{p}{q} >_\ell p_0$. $F(\mathbb{R})$ 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 65 *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 $x \in (0, x_0]$.*

The total expected discounted reward $v^\rho(f)$ for a deterministic policy 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 $v_i^\rho(f^\infty)$, $i \in E$, is an element of $F(\mathbb{R})$, say $\frac{p}{q}$, where the degree of the polynomials p and q is at most N. It is well known (theorem 29) 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 $F(\mathbb{R})$.

Furthermore, the optimality equation (3.2) implies that $(1+\rho)v_i^\rho \geq (1+\rho)r_i(a) + \Sigma_j p_{ij}(a)v_j^\rho$, $(i,a) \in E \times A$, $\rho > 0$. Therefore, in the ordered field $F(\mathbb{R})$, we have $(1+\rho)v_i^\rho \geq_\ell (1+\rho)r_i(a) + \Sigma_j p_{ij}(a)v_j^\rho$, $(i,a) \in E \times A$. 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 $F(\mathbb{R})$ is called *superharmonic* if $(1+\rho)w_i(\rho) \geq_\ell (1+\rho)r_i(a) + \Sigma_j p_{ij}(a)w_j(\rho)$, $(i,a) \in E \times A$. Hence, v^ρ is superharmonic. The concept of superharmonicity is useful to derive linear programs for MDP's.

Lemma 66 *v^ρ is the smallest superharmonic vector with components in $F(\mathbb{R})$, i.e. for any superharmonic vector $w(\rho)$, $w_i(\rho) \geq_\ell v_i^\rho$, $i \in E$.*

Proof. Let $w(\rho)$ be a superharmonic vector. Because there are only a finite number of states and actions, there exists a $\rho_0 > 0$ such that $(1+\rho)w_i(\rho) \geq (1+\rho)r_i(a) + \Sigma_j p_{ij}(a)w_j(\rho)$, $(i,a) \in E \times A$, $\rho \in (0, \rho_0]$. Therefore, by the results of discounted rewards, $w_i(\rho) \geq v_i^\rho$, $i \in E$, for every $\rho \in (0, \rho_0]$, i.e. $w_i(\rho) \geq_\ell v_i^\rho, i \in E$. ∎

Lemma 66 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 $F(\mathbb{R})$:

$$\min\{\Sigma_j w_j(\rho) | \Sigma_j [(1+\rho)\delta_{ij} - p_{iaj}]w_j(\rho) \geq_\ell (1+\rho)r_{ia},\ (i,a) \in E \times A\}. \quad (5.14)$$

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

$$\max\left\{\Sigma_{i,a}(1+\rho)r_i(a) \cdot x_{ia}(\rho) \;\middle|\; \begin{array}{ll} \Sigma_{i,a}[(1+\rho)\ \delta_{ij} - p_{ij}(a)]\cdot & \\ \qquad\qquad x_{ia}(\rho) =_\ell p_1,\ j \in E & \\ x_{ia} \geq_\ell p_0, \quad (i,a) \in E \times A & \end{array}\right\} \quad (5.15)$$

For a fixed real value of ρ, the linear programs (5.14) and (5.15) are the linear programs (3.14) and (3.15) respectively. Also from section 3.3 4 it is known that there is a one-to-one correspondence between the extreme points of (5.15) and the set of deterministic policies.

As in the simplex method, we will rewrite the equalities of (5.15) such that at each iteration there is precisely one positive $x(\rho)$ component in each state. The main difference with the usual simplex method for a fixed value of ρ is that, instead of real numbers, the elements are rational functions. During any iteration, the set of constraints is written in the special form

$$x_B = B^{-1}e - B^{-1}Ax_N \quad (5.16)$$

where e is the vector with the right-hand-side of (5.15) 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 (5.15).

We solve the dual program (5.15) 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 (5.15) and a corresponding "trial solution" $w(\rho)$ of (5.14), i.e. $w(\rho)$ satisfies the complementary slackness conditions

$$x_{ia}(\rho) \cdot \{\Sigma_j [(1+\rho)\delta_{ij} - p_{ij}(a)]w_j(\rho) - (1+\rho)r_i(a)\} = 0,\ (i,a) \in E \times A \quad (5.17)$$

for all ρ in the interval which 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 (5.17),

$$[(1+\rho)I - P(f)w(\rho) = (1+\rho)r(f), \text{ i.e. } w(\rho) = v^\rho(f). \tag{5.18}$$

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

Theorem 67 *(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 $r_i(f(i)) = \max_a r_i(a)$, $i \in E$. (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 $z_j(\rho)$, $j \in E$, 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 $z_2, \dots, x_{Nf(N)}$ with z_N, where $f(i)$ is such that $r_i(f(i)) = \max_a r_i(a)$, $1 \le i \le N$. This tableau is optimal for $\rho \ge \rho_1$, 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 tranformation 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 as follows: multiplication and division of the polynomials in the tableaux is $\mathcal{O}(N^2)$. Hence, the computation of a new element is of order N^2, i.e. the computation of a new column is of order N^3. Since there are $\Sigma_{i=1}^N \# A(i)$ columns, the complexity of one transformation is of order $N^3[\Sigma_{i=1}^N \# A(i)]$.

Hordijk, Dekker and Kallenberg [114] 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 [225], Jeroslow [129] and Holzbaur [108], [109].

1.6 A VARIETY OF OTHER TOPICS

1.6.1 Mean-variance trade-offs

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 simulaneously is a too strong requirement. Therefore, we will consider the criterion for a fixed initial distribution β. Then, as mean of the rewards for a given policy R, we use

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

If a policy R satisfies $\phi(\beta, R) = \Sigma_i \beta_i \phi_i$, where ϕ is the value vector, then R is called a β-average-optimal policy.

As variance of the rewards for a given policy R, we use the definition

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

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.

$$x_{ja}^T(R) = \frac{1}{T} \Sigma_{t=1}^{T} \mathbb{P}_{\beta,R}[X_t = j,\ Y_t = a],\ (i, a) \in E \times A \tag{6.3}$$

For a stationary policy π^∞, we have

$$x_{ja}^T(\pi^\infty) = \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_i \beta_i \cdot P^{t-1}(\pi)_{ij} \pi_{ja},\ (i, a) \in E \times A \tag{6.4}$$

By $X(R)$ we denote the limit points of the vectors $\{x^T(R),\ T = 1, 2, \dots\}$. Since $0 \leq x^T(R) \leq e$ for every $T \in \mathbb{N}$, $X(R) \neq \emptyset$. Any $x^T(R)$ satisfies $\Sigma_{j,a} x_{ja}^T(R) = 1$ and therefore $\Sigma_{j,a} x_{ja}(R) = 1$ for every $x(R) \in X(R)$. For a stationary policy π^∞, we can write

$$\lim_{T\to\infty} x_{ja}^T(\pi^\infty) = \Sigma_i \beta_i \cdot [P^*(\pi)]_{ij} \pi_{ja},\ (i, a) \in E \times A \tag{6.5}$$

i.e. $X(\pi^\infty)$ consists of the unique element $x(\pi^\infty)$. Furthermore, $\Sigma_{j,a} r_{ja} x_{ja}$ $(\pi^\infty) = \Sigma_i \beta_i \cdot \phi_i(\pi^\infty)_i$. We introduce the notation L, $L(M)$, $L(S)$ and $L(D)$ for the elements of $X(R)$ corresponding to general, Markov, stationary and deterministic policies, respectively. For policies R with $\#X(R) = 1$, e.g. stationisty policies, we denote the unique element of $X(R)$ by $x(R)$. For such policies the variance satisfies

$$v(\beta, R) = \Sigma_{j,a} x_{ja}(R) r_j^2(a) - [\Sigma_{j,a} x_{ja}(R) r_j(a)]^2 \tag{6.6}$$

Let X be the projection of the feasible solutions (x, y) of the linear program (4.10) on the x-space, i.e.

$$X = \left\{ x \;\middle|\; \begin{array}{lll} \Sigma_{i,a}[\delta_{ij} - p_{ij}] x_{ia} & = 0, & j \in E \\ \Sigma_a x_{ja} + \Sigma_{i,a}[\delta_{ij} - p_{ij}(a)] y_{ia} & = \beta_j, & j \in E \text{ for some } y \\ x_{ia}, y_{ia} \geq 0, & & (i, a) \in E \times A \end{array} \right\} \tag{6.7}$$

Then, the next theorem can be shown.

Theorem 68 *$\overline{L(D)} = \overline{L(S)} = L(M) = L = X$, where $\overline{S}$ is the closed convex hull of a set S.*

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

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

$$\max\left\{\frac{[\phi(\beta, R]^2}{v(\beta, R)}\,|\phi(\beta, R) \geq L\right\} \tag{6.8}$$

This problem can equivalently be written as

$$\max\left\{\frac{-\Sigma_{j,a}x_{ja}(R)r_j^2(a)}{[\Sigma_{j,a}x_{ja}(R)r_j(a)]^2}\,\middle|\;\Sigma_{j,a}x_{ja}(R)r_j(a) \geq L\right\} \tag{6.9}$$

For this problem we consider the mathematical programming problem

$$\max\left\{\frac{-\Sigma_{j,a}x_{ja}r_j^2(a)}{[\Sigma_{j,a}x_{ja}r_j^2(a)]}\,\middle|\;\begin{array}{l} x \in X, \\ \Sigma_{j,a}x_{ja}r_j(a) \geq L \end{array}\right\} \tag{6.10}$$

(2) Minimal variance with lower bound on the mean

$$\min\{v(\beta, R)|\phi(\beta, R) \geq L\} \tag{6.11}$$

This problem can be rewritten as

$$\max\left\{-\Sigma_{j,a}x_{ja}(R)r_j^2(a)+[\Sigma_{j,a}x_{ja}(R)r_j(a)]^2 \mid \Sigma_{j,a}x_{ja}(R)r_j(a)\geq L\right\} \tag{6.12}$$

with as mathematical programming formulation

$$\max\left\{-\Sigma_{j,a}x_{ja}r_j^2(a) + [\Sigma_{j,a}x_{ja}r_j(a)]^2\,\middle|\;\begin{array}{l} x \in X \\ \Sigma_{j,a}x_{ja}r_j(a) \geq L \end{array}\right\} \tag{6.13}$$

(3) Variance-penalized formulation

$$\max\{\phi(\beta, R) - \lambda \cdot v(\beta, R)\} \text{ for some penalty } \lambda > 0 \tag{6.14}$$

This problem is equivalent to

$$\max\{\Sigma_{j,a}x_{ja}(R)r_j(a) - \lambda \cdot [\Sigma_{j,a}x_{ja}(R)r_j^2(a) - (\Sigma_{j,a}x_{ja}(R)r_j(a))^2]\} \tag{6.15}$$

which yields the mathematical program

$$\max \left\{ \Sigma_{j,a} x_{ja} r_j(a) - \lambda \cdot \{ \Sigma_{j,a} x_{ja} r_j^2(a) - [\Sigma_{j,a} x_{ja} r_j(a)]^2 \} \, | x \in X \right\} \quad (6.16)$$

The mathematical programs (6.10), (6.13) and (6.16) are special cases of the following *unifying program*

$$\max \left\{ \frac{\Sigma_{j,a} B_{ja} x_{ja}}{D(\Sigma_{j,a} R_{ja} x_{ja}]} + C(\Sigma_{j,a} R_{ja} x_{ja}) \,\middle|\, \begin{array}{l} x \in X \\ L \le \Sigma_{j,a} R_{ja} x_{ja} \le U \end{array} \right\} \quad (6.17)$$

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 and (iii) $\Sigma_{j,a} B_{ja} x_{ja} \le 0$ for every $x \in X$.

In order to solve (6.17), we consider a parametric version of (4.10) with $B_{ia} + \vartheta R_{ia}$ instead of $r_i(a)$, $(i,a) \in E \times xA$, i.e.

$$\max \left\{ \Sigma_{i,a} x_{ia} [B_{ia} + \vartheta R_{ia}] \,\middle|\, \begin{array}{ll} \Sigma_{i,a} [\delta_{ij} - p_{ij}] x_{ia} & = 0, \quad j \in E \\ \Sigma_a x_{ja} + \Sigma_{i,a} [\delta_{ij} - p_{ij}(a)] y_{ia} & = \beta_j, \ j \in E \\ x_{ia}, y_{ia} \ge 0, \ (i,a) \in E \times A & \end{array} \right\} \quad (6.18)$$

with $\vartheta \in (-\infty, +\infty)$ as the parameter. It is well known that 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 $\vartheta \in [\vartheta_{n-1}, \vartheta_n]$, $1 \le n \le m$, 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 $\Sigma_{i,a} R_{ia} x_{ia}^{k+1} > U$ and $\Sigma_{i,a} R_{ia} x_{ia}^{j+1} \ge L$. 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 $\Sigma_{i,a} R_{ia} x_{ia}^U = U$ and $\Sigma_{i,a} R_{ia} x_{ia}^L = L$.
Let $G(x) = \Sigma_{j,a} B_{ja} x_{ja}$, $g(x) = \Sigma_{j,a} R_{ja} x_{ja}$ 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 $V^n = V(x^n)$, $1 \le n \le m$. Furthermore, define $V_{\text{opt}} = \max\{\max_{j+1 \le n \le k} V^n, V(x^L), V(x^U)\}$.

Theorem 69 *(i) Program (6.17) is feasible if and only if $g(x^m) \ge L$ and $g(x^1) \le U$; (ii) If program (6.17) is feasible, then V_{opt} is the optimal value of (6.17), and the maximizing x is the optimal solution x_{opt}.*

If there are no bounds L and U for $\Sigma_{j,a} R_{ja} x_{ja}$, then $V_{\text{opt}} = V(x^n)$ for some extreme point x^n of X. Theorem 69 provides a way to find an optimal solution for the program (6.17). But it does not provide a procedure to construct an optimal policy for the corresponding Markov decision chain. The next two theorems give the answer.

Theorem 70 *If (x, y) is an extreme optimal solution for (6.18) for all ϑ in an open interval, then any deterministic policy f^∞ with $x_{if(i)} > 0$ if $\Sigma_a x_{ia} > 0$ and $y_{if(i)} > 0$ if $\Sigma_a x_{ia} = 0 \wedge \Sigma_a y_{ia} > 0$ has a state-action frequency vector $x(f^\infty)$ satisfying $\Sigma_{i,a} B_{ia} x_{ia}(f^\infty) = \Sigma_{i,a} B_{ia} x_{ia}$, $\Sigma_{i,a} R_{ia} x_{ia}(f^\infty) = \Sigma_{i,a} R_{ia} x_{ia}$ and $V(x(f^\infty)) = V(x)$.*

Theorem 71 *If program (6.17) is feasible, then either $x_{opt} = x^n$ for some $j+1 \leq n \leq k$ and there exists an optimal deterministic policy, or $x_{opt} = x^L$ (or x^U) and an initial randomization of two deterministic policies is optimal for the mean-variance trade-off problem. These policies can be determined analogously to the policy in theorem 70.*

Corollary 72 *For an unconstrained mean-variance trade-off problem, i.e. without the constraint $L \leq \Sigma_{i,a} R_{ia} x_{ia} = U$, there exists a deterministic optimal policy.*

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. In the discounted case, there is a one-to-one correspondence between the deterministic policies and the extreme solutions. In the unichain case the initial distribution has no influence and the parametric linear program is more simple.
2. The optimal policy R_* is also *Pareto-optimal* with respect to the pair $(\phi(R), -V(R))$, i.e. there doesn't exist a policy R such that $\phi(R) \geq \phi(R_*)$ and $-V(R) \geq -V(R_*)$, with a strict inequality holding for at least one of the two inequalities.

State-action frequencies for the unichain case are discussed in Derman [58]. The multichain case is analysed in Kallenberg [135] and Hordijk en Kallenberg [117], who have shown theorem 68. 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 [241], Ross [196], Ross and Varadarajan [197], Altman and Shwartz [7] and by Liu and Ohno [157]. The mean-variance formulations (6.8), (6.11) and (6.14) were proposed by Sobel [229], Kawai [143] and Filar, Kallenberg and Lee [78], respectively. Other contributions in this area are Kawai and Katoh [144], White [283], [284] and [287], Chung [37], [38] and [39], Bayal-Gursoy and Ross [12], and Sobel [230]. The unifying framework is proposed by Huang and Kallenberg [124], who also have shown the theorems 69, 70 and 71. 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].

1.6.2 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 with 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:

$$E = \{1, 2, \dots, N\};\ A(i) = \{1, 2\},\ i \in E;\ r_i(1) = r_i,\ i \in E;$$
$$r_i(2) = -c_i,\ i \in E;\ p_{ij}(1) = 0,\ i, j \in E;\ p_{ij}(2) = p_{ij},\ i, j \in E.$$

We are interested in finding an optimal stopping rule, i.e. we consider only transient policies. A policy R is called *transient* if $\Sigma_{t=1}^{\infty} \mathbb{P}_{i,R}[X_t \in X] < \infty$ for all $i \in E$, 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.

$$v_i(R) = \Sigma_{t=1}^{\infty} \Sigma_{j,a} \mathbb{P}_{i,R}[X_t = j,\ Y = a] \cdot r_j(a) \tag{6.19}$$

For the computation of an optimal transient policy, the usual properties of discounted MDPs hold (cf. Kallenberg [135] chapter 3). Let v be the value vector, i.e. $v = \sup\{v(R) | R \text{ is transient}\}$. 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

$$\begin{cases} v_i \geq r_i & ,\ i \in E \\ v_i \geq -c_i + \Sigma_j p_{ij} v_j & ,\ i \in E. \end{cases} \tag{6.20}$$

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

$$\min \left\{ \Sigma_j v_j \middle| \begin{array}{ll} v_i \geq r_i & ,\ i \in E \\ v_i \geq -c_i + \Sigma_j p_{ij} v_j & ,\ i \in E \end{array} \right\} \tag{6.21}$$

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
$$\max\left\{\Sigma_i r_i x_i - \Sigma_i c_i y_i \middle| \begin{array}{rl} x_j + y_j - \Sigma_i\ p_{ij} y_i & = 1, \quad j \in E \\ x_j, y_j & \geq 0, \quad j \in E \end{array}\right\}$$

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

Let $S = \{i \in E | r_i \geq -c_i + \Sigma_j p_{ij} r_j\}$, 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 73 *In a monotone optimal stopping problem the policy f^∞, where $f(i) = 1$ if and only if $i \in S$, is optimal.*

Proof. Let w be such that $w_i = \begin{cases} r_i, & i \in S \\ v_i, & i \notin S. \end{cases}$ Then, w is superharmonic, namely

$$\begin{array}{ll} i \in S : w_i & = r_i \geq -c_i + \Sigma_j p_{ij} r_j = -c_i + \Sigma_{j \in S} p_{ij} r_j = -c_i + \Sigma_j p_{ij} w_j. \\ i \notin S : w_i & = v_i \geq -c_i + \Sigma_j p_{ij} v_j \geq -c_i + \Sigma_{j \in S} p_{ij} r_j + \Sigma_{j \notin S} p_{ij} v_j \\ & = -c_i + \Sigma_j p_{ij} w_j \text{ and } w_i = v_i \geq r_i. \end{array}$$

Since v is the smallest superharmonic vector, it follows that $w \geq v$. Hence,
$i \in S:\ r_i = w_i \geq v_i \geq r_i$, i.e. $r_i = w_i = v_i$ and $f(i) = 1$ is optimal;
$i \notin S:\ r_i < -c_i + \Sigma_j p_{ij} r_j \leq -c_i + \Sigma_j p_{ij} v_j \leq v_i$, i.e. $f(i) = 2$ is optimal. ∎

For monotone stopping problems it is sufficient to determine the set S. The determination of S has complexity of order $\mathcal{O}(N)$. For instance, the house selling problem of section 1.4 is monotone with $S = \{i \in E | c \geq \Sigma_{j>i} (j-i) p_j\} = \{i \in E | i \geq i_*\}$ where $i_* = \min\{i | c \geq \Sigma_{j>i}\ (j-i) p_j\}$, i.e. S consists of the states i for which the expected income above i in the next week is at most the costs in that week. Furthermore, an optimal policy is to accept an offer as soon as it is at least i_*. Such a policy is said to be a *control-limit* policy.

A classical paper on optimal stopping problems is Breiman [28]. Other papers in this area are Chen [34], Ross [198], Yasuda [292] and Sonin [232]. We refer also to the chapter "Optimal stopping: old and new results and methods" in this book, written by Sonin.

1.6.3 Multi-armed bandit problems

We have introduced this model in section 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 E_j, $1 \leq j \leq n$. Hence, the state space is the Cartesian product: $E = E_1 \times E_2 \times \cdots \times E_n$. Each state has the same action set $\{1, 2, \dots, n\}$, where action k means that project k is chosen, $1 \leq k \leq n$. When project k is chosen, i.e. project k is the *active project*, the immediate reward and the transition probabilities only depend on project k and the state $i \in E_k$. Let $r_i(k)$ and $p_{ij}(k)$, $j \in E_k$, 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

$$\min \left\{ \Sigma_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i + \alpha \Sigma_j p_{ij} v_j, & i \in E \\ v_i \geq M, & i \in E \end{array} \right\} \tag{6.22}$$

and its dual

$$\max \left\{ \Sigma_i r_i x_i + M \cdot \Sigma_i y_i \;\middle|\; \begin{array}{rl} \Sigma_i (\delta_{ij} - \alpha p_{ij})\; x_i + y_j & = 1,\; j \in E \\ x_i, y_i & \geq 0,\; i \in E \end{array} \right\} \tag{6.23}$$

Lemma 74 *For all $i \in E$, $v_i^\alpha(M) - M$ is a nonnegative continuous nonincreasing function in M.*

Proof. The nonnegativity of $v_i^\alpha(M) - M$ is obvious. By the method of value iteration $v^\alpha(M)$ can be approximated by
$v_i^1(M) = M$, $i \in E$; $v_i^{n+1}(M) = \max\{M, r_i + \alpha \Sigma_j p_{ij} v_j^n(M)\}$, $i \in E$, $n \geq 1$.
Substituting $w_i^n = v_i^n - M$ for all i and n, we obtain
$w_i^1(M) = 0$, $i \in E$; $w_i^{n+1}(M) = \max\{0, r_i + \alpha \Sigma_j p_{ij} w_j^n(M) - (1 - \alpha)M\}$, $i \in E$, $n \geq 1$.
By induction on n, it can be shown that $w_i^n(M)$ is continuous and nonincreasing in M. Hence, the limit $v_i^\alpha(M) - M$ is also a continuous nonincreasing function in M for all $i \in E$. ∎

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

Theorem 75 *The policy f^∞, which chooses the stopping action in state i if and only if $M_i^\alpha \leq M$, is optimal.*

Proof. Let (x, y) be an extreme optimal solution of linear program (6.23). Then in each state i only one of the variables x_i or y_i is basic and positive.

When $x_i > 0$ then continuing is optimal; otherwise it is optimal to stop. Suppose that, $M_i^\alpha > M$, i.e. $v_i^\alpha(M) > M$. Then, by the complementary slackness, $y_i = 0$ and it is optimal to continue. Otherwise, suppose that when the stopping action is chosen, i.e. $M_i^\alpha \leq M$, this action is not optimal. This implies, $v_i^\alpha(M) = r_i + \alpha\Sigma_j p_{ij} v_j^\alpha(M) > M$ and consequently $M < M_i^\alpha$, wich yields a contradiction. ∎

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 k, $1 \leq k \leq n$, means continue with project k, 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 $C = (1-\alpha)^{-1} \cdot \max_{i,k} |r_i(k)|$, 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 results are 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 76 *For all $i \in E$, (i) $v_i^\alpha(M)$ is a nondecreasing, convex function in M; (ii) $\frac{\partial}{\partial M} v_i^\alpha(M) = \mathbb{E}_{i,f_M^\infty}[\alpha^{T(M)}]$.*

The next theorem is the key theorem for the multi-armed bandit problem. It says that an optimal action in a state $i = (i_1, i_2, \dots, i_n)$ is to choose the 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 77 *In state $i = (i_1, i_2, \dots, i_n)$ the optimal policy chooses action k, where k is such that $M_{i_k}^\alpha = \max_j M_{i_j}^\alpha$.*

Alternative interpretation of the Gittins index

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

$$\begin{aligned}(1-\alpha)M_i^\alpha &= \max_{T\geq 1}\mathbb{E}[\text{discounted reward before } T]/[\{1-\mathbb{E}(\alpha^T)\}/(1-\alpha)]\\ &= \max_{T\geq 1}\mathbb{E}[\text{discounted reward before } T]/\mathbb{E}[\text{discounted time before } T],\end{aligned}$$

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 (6.22) and (6.23). For M big enough, an optimal solution (x, y) of (6.23) will satisfy $y_i > 0,\ i \in E$. 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. In the first tableau the y-variables are the basic variables. Setting up this tableau is of $\mathcal{O}(N^2)$. Then, there are N transformations, by each transformation one Gittins index, is computed and each transformation is of $\mathcal{O}(N^2)$. Therefore, the overall complexity is $\mathcal{O}(N^3)$.

Interpretation as restart-in-k problem
We will derive another interpretation for the Gittins index M_k^α in a fixed state k. For any terminal value M, we have

$$v_i^\alpha(M) = \max\{M, r_i + \alpha\Sigma_j p_{ij} v_j^\alpha(M)\},\ i \in E \tag{6.24}$$

and in state k, for $M = M_k^\alpha$,

$$v_k^\alpha(M) = M_k^\alpha = r_k + \alpha\Sigma_j p_{kj} v_j^\alpha(M_k^\alpha) \tag{6.25}$$

Substituting (6.25) in (6.24) gives

$$v_i^\alpha(M_k^\alpha) = \max\{r_k + \alpha\Sigma_j p_{kj} v_j^\alpha(M_k^\alpha),\ r_i + \alpha\Sigma_j p_{ij} v_j^\alpha(M_k^\alpha)\},\ i \in E \tag{6.26}$$

Hence, M_k^α 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_k^α 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

$$\max\left\{\Sigma_j v_j \;\middle|\; \begin{array}{l}\Sigma_j(\delta_{ij}-\alpha p_{ij})v_j \geq r_i, \;\; i \neq k \\ \Sigma_j(\delta_{ij}-\alpha p_{kj})v_j \geq r_k, \; i \in E\end{array}\right\} \tag{6.27}$$

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

Theorem 78 *Let $C_k = \{i|$ to continue is optimal for the restart-in-k problem$\}$. Then, $C_k = \{i \in E | M_i^\alpha \geq M_k^\alpha\}$.*

Largest remaining index
As in parametric linear programming, also by the result on the largest remaining index (see the theorem below) the indices can be computed in a sequence, starting with the largest index.

Theorem 79 *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 $[P^k]_{ij} = \begin{cases} p_{ij}, & j \leq k \\ 0, & j > k. \end{cases}$

Proof. $v_i^\alpha(M_{l_k}^\alpha) \geq r_i + \alpha\Sigma_j p_{ij} v_j^\alpha(M_{l_k}^\alpha)$, $i \in E$, with equality for the state l_k. Since $v_i^\alpha(M) = M$ for $M \geq M_i^\alpha$, $v_i^\alpha(M_{l_k}^\alpha) \geq r_i + \alpha\Sigma_{j\leq k} p_{ij} v_j^\alpha(M_{l_k}^\alpha) + \alpha M_{l_k}^\alpha[1 - \Sigma_{j\leq k} p_{ij}]$, $i \in E$. In vector notation this relation can be written as
$v \geq r + \alpha P^k v + \alpha M_{l_k}^\alpha e - \alpha M_{l_k}^\alpha P^k e = r + \alpha P^k v - (1-\alpha)M_{l_k}^\alpha e + M_{l_k}^\alpha [I - \alpha P^k]e$,
i.e. $[I - \alpha P^k]v \geq r - (1-\alpha)M_{l_k}^\alpha e + M_{l_k}^\alpha[I - \alpha P^k]e$, and consequently,
$v \geq [I - \alpha P^k]r - (1-\alpha)M_{l_k}^\alpha[I - \alpha P^k]e + M_{l_k}^\alpha e$. Componentwise for $i > k$, we have $M_{l_k}^\alpha = v_i^\alpha(M_{l_k}^\alpha) \geq [(I - \alpha P^k)r]_i - (1-\alpha)M_{l_k}^\alpha[(I - \alpha P^k)e]_i + M_{l_k}^\alpha$, with equality for $i = l_k$. Hence, we obtain $(1-\alpha)M_{l_k}^\alpha = \max_{i>k} \frac{[(I-\alpha P^k)^{-1}r]_i}{[(I-\alpha P^k)^{-1}e]_i}$. ∎

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 $\Sigma_{k=1}^N \mathcal{O}(k^2) = \mathcal{O}(N^3)$.

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 [289] and [290], Ross [201], Tsitsiklis [243] and [244], Katehakis and Veinott [142], Weber [268] and Ishikida and Varaiya [128]. In honour of Gittins, Whittle has introduced the term Gittins indices. A first linear programming method of $\mathcal{O}(N^4)$ is proposed by Chen and Kathehakis [35]. Kallenberg [136] has improved this method to $\mathcal{O}(N^3)$. The interpretation as restart-in-k problem is made by Kathehakis and Veinott [142]. The method of the largest remaining index rule is due to Varaiya, Walrand and Buyokkoc [255]. 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 [271]; 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].

1.6.4 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. $r_i(a) = s_i + t(a)$, (ii) the transition probabilities depend only on the action and not on the state from which the transition occurs, i.e. $p_{ij}(a) = p_j(a)$.

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 $E_1 = \{1, 2, \dots, m)$, there are subsets of the action sets, say subset $A_1(i)$ in state $i \in E_1$, such that:

(i) $r_i(a) = s_i + t(a)$, $i \in E_1$, $a \in A_1(i)$;

(ii) $p_{ij}(a)$ is independent of $i : p_{ij}(a) = p_j(a)$, $i \in E_1$, $a \in A_1(i)$, $j \in E$.

(2) The action subsets are nested: $A_1(1) \supseteq A_1(2) \supseteq \cdots \supseteq A_1(m) \neq \emptyset$.

Let $E_2 = E \backslash E_1$, $A_2(i) = A(i) \backslash A_1(i)$, $1 \leq i \leq m$, $A_2(i) = A(i)$, $m + 1 \leq i \leq N$, and $B(i) = A_1(i) - A_1(i + 1)$, $1 \leq i \leq m - 1$, $B(m) = A_1(m)$. Then $A_1(i) = \cup_{j=i}^{m} B(j)$, and the sets $B(j)$ are disjoint. E_1, E_2, $A_2(i)$ or $B(i)$ may be empty.

If the system is observed in a state $i \in E_1$, and the decision maker will choose an action from $A_1(i)$, the decision process can be considered as follows. First a reward s_i is earned and the systems makes a zero-time transitio n 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 $A_1(i) - B(i) = A_1(i+1)$. In the first case reward $t(a)$ is earned and the process moves to state j with probability $p_j(a)$, $j \in E$; 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 6.5). It can be shown that a linear program, which directly provides a deterministic average optimal policy, can be formulated. This linear program is based on linear programming for semi-Markov decision problems (cf. Kallenberg [135], chapter 7). It is not a straightforward application, because there are zero-time transitions which are not allowed in general.

Consider the linear program

$$\begin{array}{l}
\textit{minimize } \Sigma_{j=1}^N g_j + \Sigma_{j=1}^m h_j \hfill (6.28)\\
\Sigma_{j=1}^N [\delta_{ij} - p_{ij}(a)] g_j \geq 0,\ 1 \leq i \leq N,\ a \in A_2(i);\ g_i - h_i \geq 0,\ 1 \leq i \leq m\\
-\Sigma_{j=1}^N p_j(a) g_j + h_i \geq 0,\ 1 \leq i \leq m,\ a \in B(i);\ h_i - h_{i+1} \geq 0,\ 1 \leq i \leq m-1\\
g_i + \Sigma_{j=1}^N [\delta_{ij} - p_{ij}(a)] u_j \geq r_i(a),\ 1 \leq i \leq N,\ a \in A_2(i);\ u_i - v_i \geq s_i,\ 1 \leq i \leq m\\
h_i - \Sigma_{j=1}^N p_j(a) u_j + v_i \geq t(a),\ 1 \leq i \leq m,\ a \in B(i);\ v_i - v_{i+1} \geq 0,\ 1 \leq i \leq m-1
\end{array}$$

The corresponding dual program is (the dual variables corresponding to the constraints of (6.28) are y_{ia}, μ_i, z_{ia}, σ_i, x_{ia}, λ_i, w_{ia} and ρ_i):

$$\begin{array}{l}
\text{maximize } \Sigma_{i=1}^N \Sigma_{a \in A(i)} r_i(a) x_{ia} + \Sigma_{i=1}^m s_i \lambda_i + \Sigma_{i=1}^m \Sigma_{a \in B(i)} t(a) w_{ia} \hfill (6.29)\\
\Sigma_{i=1}^N \Sigma_{a \in A(i)} [\delta_{ij} - p_{ij}(a)] y_{ia} + \Sigma_{i=1}^m \delta_{ij} \mu_i \ \ -\Sigma_{i=1}^m \Sigma_{a \in B(i)} p_j)a) z_{ia}\\
\hfill + \Sigma_{a \in A(j)} x_{ja} = 1,\ \ 1 \leq j \leq N\\
\sigma_j - \sigma_{j-1} + \Sigma_{a \in B(j)} w_{ja} - \mu_j + \Sigma_{a \in B(j)} z_{ja} = 1,\ 1 \leq j \leq m\\
\Sigma_{i=1}^N \Sigma_{a \in A_2(i)} [\delta_{ij} - p_{ij}(a)] x_{ia} + \Sigma_{i=1}^m \delta_{ij} \lambda_i - \Sigma_{i=1}^m \Sigma_{a \in B(i)} p_j(a) w_{ia} = 0,\\
\hfill 1 \leq j \leq N\\
\rho_j - \rho_{j-1} + \Sigma_{a \in B(j)} w_{ja} - \lambda_j = 0, 1 \leq j \leq m;\ \rho_0 = \rho_m = \sigma_0 = \sigma_m = 0;\\
x_{ia},\ y_{ia},\ z_{ia},\ w_{ia},\ \lambda_i,\ \mu_i,\ \rho_i,\ \sigma_i \geq 0 \text{ for all } i \text{ and } a.
\end{array}$$

A proof for the next result can be found in Kallenberg [137].

Theorem 80 *(i) The linear programs (6.28) and (6.29) have finite optimal solutions; (ii) If (g,h,u,v) is an optimal solution of program (6.28), then g is the value vector; (iii) Let $(y,\mu,z,\sigma,x,\lambda,w,\rho)$ be an extreme optimal solution of program (6.29). Define m_i and n_i by:*
$m_i = \min\{j \geq i \mid \Sigma_a w_{ja} > 0\}$ and $n_i = \min\{j \geq i \mid \Sigma_a (w_{ja} + z_{ja}) > 0\}$, $i \in E$.
Take a deterministic policy f^∞ such that in state $i : x_{if(i)} > 0$ if $\Sigma_a x_{ia} > 0$;

$w_{m_i f(i)} > 0$ *if* $\Sigma_a x_{ia} = 0 \wedge \lambda_i > 0$; $y_{if(i)} > 0$ *if* $\Sigma_a x_{ia} = 0 \wedge \lambda_i = 0 \wedge \Sigma_a y_{ia} > 0$; $w_{n_i f(i)} > 0$ *if* $\Sigma_a x_{ia} = 0 \wedge \lambda_i = 0 \wedge \Sigma_a y_{ia} = 0 \wedge \Sigma_a w_{n_i a} > 0$; $z_{n_i f(i)} > 0$ *if* $\Sigma_a x_{ia} = 0 \wedge \lambda_i = 0 \wedge \Sigma_a y_{ia} = 0 \wedge \Sigma_a w_{n_i a} = 0$.
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 [122] 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 general 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 $0 \leq i \leq N$. The possible actions are: either to order nothing or to order $a - i$ items, where $i + 1 \leq a \leq N$, 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 general Markov decision problem has $(N+1)(N+2)$ variables and $2(N+1)$ constraints. In the reduced formulation of this 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:
$E = \{1, 2, \dots, N\}$; $A(i) = \{1, 2, \dots, M\}$, $i \in E$; $r_i(a) = s_i + t(a)$, $(i, a) \in E \times A$; $p_{ij}(a) = p_j(a)$, $(i, a) \in E \times A$ and $j \in E$.
Examples of this model can be found in Sobel [228]. 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:

$$t(a_*) + \Sigma_{j=1}^N p_j(a) s_j = \max_{1 \leq a \leq M} \{t(a) + \Sigma_{j=1}^N p_j(a) s_j\} \qquad (6.30)$$

This result is a special case of the stochastic game studied in Sobel [228] and Parthasarathy, Tijs and Vrieze [177].

1.6.5 Further subjects

In this chapter the main topics, but not all, of finite MDPs were 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 $F^a_{ij}(t)$ 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 [130], [131], Howard [123], De Cani [41] and Schweitzer [207]. Contributions for *discounted rewards* are e.g. Denardo [45] (contraction property), De Ghellinck and Eppen [43] and Kallenberg [135] (linear programming), Wessels and Van Nunen [273] (linear programming and policy iteration), Ohno [175] and Schweitzer [214] (value iteration).

In the *average reward case*, there is a very elegant data transformation, proposed by Schweitzer [210], 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 $0 < \tau \leq \min_{i,a} \tau_i(a)$, let

$$\begin{cases} \overline{r}_1(a) = r_i/\tau_i(a) & , \ i \in E, \ a \in A(i) \\ \overline{p}_{ij}(a) = \delta_{ij} - [\delta_{ij} - p_{ij}(a)] \cdot \tau/\tau_i(a) & , \ i,j \in E, \ a \in A(i) \end{cases} \tag{6.31}$$

Then, $\phi(\pi^\infty) = \overline{\phi}(\pi^\infty)$, where $\phi(\pi^\infty)$ is the average reward per unit time of the semi-MDP and $\overline{\phi}(\pi^\infty)$ the average reward of the discrete-time MDP with rewards $\overline{r}_i(a)$ and transition probabilities $\overline{p}_{ij}(a)$ as defined in (6.31).

Other papers on average reward MDPs are Schweitzer and Federgruen [217] (optimality equation), Federgruen and Spreen [75] (policy iteration), Denardo and Fox [51] (linear programming and policy iteration), Osaki and Mine [176] and Kallenberg [135] (linear programming), Schweitzer and Federgruen [218], and Schweitzer [211] (value iteration). In Denardo [48] more sensitive optimality criteria for semi-MDPs are considered.

MDPs with partial information, partial observation and adaptive control

In an MDP with *partial information* the exact state of the process can not 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 $E_1, E_2, \dots, E_m$ such that at each decision epoch the only available information is the subset E_k to which the state belongs. In the partial information case not all decision rules are feasible: in all states of a subset $E_k (1 \leq k \leq m)$ 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 [226], Hastings and Sadjani [98], Hordijk and Loeve [118], and Loeve [159].

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 [231], Albright [1], Monahan [169], Altman and Shwartz [5] and [6], Lovejoy [160], [161] and [162], Rieder [193], Sernik and Markus [220], White III [274] and [275], White III and Scherer [276] and [277], and White [288].

In *adaptive control* models the transition probabilities $p_{ij}(a)$ 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. For a survey on adaptive control we refer to Katehakis' chapter in this book. Further literature about this topic is e.g. Kurano [148], Hübner [126], 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 $r_i^k(a)$, $1 \leq k \leq m$, of rewards. For this model, the concept of optimality is not unambigious. Given an initial distribution β, a policy R and a utility function u (e.g. discounted or average expected rewards), there is an m-vector $u(\beta, R)$ of returns, where the k-th component $u_k(\beta, R)$ corresponds to the rewards $r_i^k(a)$. 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 not a 'better' policy, i.e. there is not a 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 [280], Henig [101], Kallenberg [135], Durinovic, Lee, Katehakis and Filar [65], Ghosh [90], Liu, Ohno and Nakayama [158], and Wakuta [262], [263], [264] and [265].

Acknowledgement

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.

Bibliography

albr79 [1] Albright,S.C. [1979]: "Structural results for partially observable Markov decision processes", Operations Research *27*, 1041-1053.

alt99 [2] Altman,E. [1999]: "Constrained Markov decision processes", Chapman & Hall/CRC, Boca Raton, Florida.

ahk96 [3] Altman,E., 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.

alsh91a [4] Altman,E. and A.Shwartz [1991a]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1-22.

alsh91b [5] Altman,E. and A.Shwartz [1991b]: "Adaptive control of constrained Markov chains", IEEE-Transactions on Automatic Control *36*, 454-462.

alsh91c [6] Altman,E. and A.Shwartz [1991c]: "Adaptive control of constrained Markov decision chains: criteria and policies", Annals of Operations Research *28*, 101-134.

alsh91 [7] Altman,E. and A.Shwartz [1991]: "Sensitivity of constrained Markov decision processes", Annals of Operations Research *33*, 1-22.

alsp95 [8] Altman,E. and F.M.Spieksma [1995]: "The linear program approach in Markov decision processes", Mathematical Methods of Operations Research *42*, 169-188.

bmm85 [9] Baras,J.S., 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.

bath73a [10] Bather,J. [1973a]: "Optimal decision procedures for finite Markov chains. Part II: Communicating systems", Advances in Applied Probability *5*, 521-540.

bath73b [11] Bather,J. [1973b]: "Optimal decision procedures for finite Markov chains. Part III: General convex systems", Advances in Applied Probability *5*, 541-553.

bayros92 [12] Bayal-Gursoy,M. and K.W.Ross [1992]: "Variability-sensitivity Markov decision processes", Mathematics of Operations Research *17*, 558-571.

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

benis90 [14] Ben-Israel,A. and S.D.Flam [1990]: "A bisection/successive approximation method for computing Gittins indices", Zeitschrift für Operations Research *34*, 411-422.

bertse76a [15] Bertsekas,D.P. [1976]: "Dynamic programming and stochastic control", Academic Press, New York.

bertse76b [16] Bertsekas,D.P. [1976b]: "On error bounds for successive approximation methods", IEEE Transactions on Automatic Control *21*, 394-396.

bertse87 [17] Bertsekas,D.P. [1987]: "Dynamic programming: deterministic and stochastic models", Prentice-Hall, Englewood Cliff.

bertse95a [18] Bertsekas,D.P. [1995]: "Dynamic programming and optimal control I", Athena Scientific, Belmont, Massachusetts.

bertse95b [19] Bertsekas,D.P. [1995]: "Dynamic programming and optimal control II", Athena Scientific, Belmont, Massachusetts+.

bertse95c [20] Bertsekas,D.P. [1995c]: "Generic rank-one corrections for value iteration in Markovian decision problems", OR Letters *17*, 111-119.

bertse98 [21] Bertsekas,D.P. [1998]: "A new value iteration method for the average cost dynamic programming problem", SIAM Journal on Control and Optimization *36*, 742-759.

bertse78 [22] Bertsekas,D.P. and S.E.Shreve [1978]: "Stochastic optimal control: the discrete time case", Academic Press, New York.

bertse91 [23] Bertsekas,D.P. and J.N.Tsitsiklis [1991]: "An analysis of stochastic shortest path problems", Mathematics of Operations Research *16*, 580-595.

bertsim96 [24] Bertsimas,D. 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.

beut85 [25] Beutler,F.J. and K.W.Ross [1985]: "Optimal policies for controlled Markov chains with a constraint", Journal of Mathematical Analysis and Applications *112*, 236-252.

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

bl62 [27] Blackwell,D. [1962]: "Discrete dynamic programming", Annals of Mathematical Statistics, 719-726.

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

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

bruno [30] Bruno,J., 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.

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

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

chang [33] Chang,C.-S., 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.

chen73 [34] Chen,M.C.Jr. [1973]: "Optimal stopping in a discrete search problem", Operations Research *21*, 741-747.

chen86 [35] Chen,Y.-R. and M.N.Katehakis [1986] : "Linear programming for finite state bandit problems", Mathematics of Operations Research *11*, 180-183.

chow [36] Chow,Y.S. 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.

chung89 [37] Chung, K.-J. [1989]: "A note on maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *8*, 201-204.

chung92 [38] Chung, K.-J. [1992]: "Remarks on maximal mean/standard deviation ratio in an undiscounted MDPs", Opimization *26*, 385-392.

chung94 [39] Chung, K.-J. [1994]: "Mean-variance trade-offs in an undiscounted MDP: the unichain case", Operations Research *42*, 184-188.

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

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

degh60 [42] De Ghellinck,G.T. [1960]: "Les problèmes de décisions séquentielles", Cahiers du Centre de Recherche Opérationelle, 161-179.

degh67 [43] De Ghellinck,G.T. and G.D.Eppen [1967]: "Linear programming solutions for separable Markovian decision problems", Management Science *13*, 371-394.

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

den67 [45] Denardo,E.V. [1967]: " Contraction mappings in the theory underlying dynamic programming", SIAM Review *9*, 165-167.

den68 [46] Denardo,E.V. [1968]: "Separable Markovian decision problems", Management Science *14*, 451-462.

den70 [47] Denardo,E.V. [1970]: "Computing a bias-optimal policy in a discrete-time Markov decision problem", Operations Research *18*, 279-289.

den71 [48] Denardo,E.V. [1971]: "Markov renewal programs with small interest rates", Annals of Mathematical Statistics *42*, 477-496.

den73 [49] Denardo,E.V. [1973]: "A Markov decision problem", in: T.C.Hu and S.M.Robinson (eds.), "Mathematical Programming", Academic Press, 33-68.

den82 [50] Denardo,E.V. [1982]: "Dynamic programming: models and applications", Prentice-Hall, Englewood Cliff.

denfox68 [51] Denardo,E.V. and B.L.Fox [1968]: "Multichain Markov renewal programs", SIAM Journal on Applied Mathematics *16*, 468-487.

denmil68 [52] Denardo,E.V. and B.L.Miller [1968]: "An optimality condition for discrete dynamic programming with no discounting", Annals of Mathematical Statistics *39*, 1220-1227.

den79a [53] Denardo,E.V. and U.G.Rothblum [1979a]: "Optimal stopping, exponential utility and linear programming", Mathematical Programming *16*, 228-244.

den79b [54] Denardo,E.V. and U.G.Rothblum [1979b]: "Overtaking optimality for Markov decision chains", Mathematics of Operations Research *4*, 144-152.

depn60 [55] D'Epenoux,F. [1960]: "Sur un problème de production et de stockage dans l'aléatoire", Revue Française de Recherche Opérationelle, 3-16.

der62 [56] Derman,C. [1962]: "On sequential decisions and Markov chains", Management Science *9*, 16-24.

der63 [57] Derman,C. [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.

der70 [58] Derman,C. [1970]: "Finite state Markovian decision processes", Academic Press, New York.

der65 [59] Derman,C. and M.Klein [1965]: "Some remarks on finite horizon Markovian decision models", Operations Research *13*, 272-278.

der60 [60] Derman,C. and J.Sacks [1960]: "Replacement of periodically inspected equipment (an optimal stopping rule)", Naval Research Logistics Quarterly *7*, 597-607.

der66 [61] Derman,C. and R.Strauch [1966]: "A note on memoryless rules for controlling sequential control problems", Annals of Mathematical Statistics *37*, 276-278.

der72 [62] Derman.C. and A.F.Veinott Jr. [1972]: "Constrained Markov decision chains", Management Science *19*, 389-390.

dietz83 [63] Dietz,H.M. and V. Nollau [1983]: "Markov decision problems with countable state space", Akademie-Verlag, Berlin.

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

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

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

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

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

fed84 [69] Federgruen,A. [1984]: "Markovian control problems: functional equations and algorithms", Mathematical Centre Tract 97, Mathematical Centre, Amsterdam.

fed78 [70] Federgruen,A. 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.

fed80 [71] Federgruen,A. 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.

fed84a [72] Federgruen,A. and P.J.Schweitzer [1984a]: "A fixed-point approach to undiscounted Markov renewal programs", SIAM Journal on Algebraic Discrete Methods *5*, 539-550.

fed84b [73] Federgruen,A. and P.J.Schweitzer [1984b]: "Successive approximation methods for solving nested functional equations in Markov decision problems", Mathematics of Operations Research *9*, 319-344.

fst78 [74] Federgruen,A. P.J.Schweitzer and H.C.Tijms [1978]: "Contraction mappings underlying undiscounted Markov decision problems", Journal of Mathematical Analysis and Applications *65*, 711-730.

fedspre80 [75] Federgruen,A. and D.Spreen [1980]: "A new specification of the multichain policy iteration algorithm in undiscounted Markov renewal programs", Management Science *26*, 1211-1217.

fedzip84 [76] Federgruen,A. and P.Zipkin [1984]: "An efficient algorithm for computing optimal (s,S) policies", Operations Research *34*, 1268-1285.

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

filar89 [78] Filar,J.A., L.C.M.Kallenberg and H.M.Lee [1989]: "Variance-penalized Markov decision processes", Mathematics of Operations Research *14*, 147-161.

filar97 [79] Filar,J.A. and O. J. Vrieze [1997]: "Competitive Markov decision processes", Springer-Verlag, New York.

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

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

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

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

garbe [84] Garbe,R. 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.

gitt79 [85] Gittins,J.C. [1979]: "Bandit processes and dynamic allocation indices", Journal of the Royal Statistic Society Series B *14*, 148-177.

gitt74 [86] Gittins,J.C. 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.

glaze96 [87] Glazebrook,K.D. and R.Garbe [1996]: "Reflections on a new approach to Gittins indexation", Journal of the Operational Research Society *47*, 1301-1309.

glaze95 [88] Glazebrook,K.D. and S. Greatrix [1995]: "On transforming an index for generalized bandit problems", Journal of Applied Probability *32*, 168-182.

glaze91 [89] Glazebrook,K.D. and R.W.Owen [1991]: "New results for generalized bandit problems", International Journal of System Sciences *22*, 479-494.

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

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

hart [92] Hartley,R., 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.

hast68 [93] Hastings,N.A.J. [1968]: "Some notes on dynamic programming and replacement", Operational Research Quarterly *19*, 453-464.

hast69 [94] Hastings,N.A.J. [1969]: "Optimization of discounted Markov decision problems", Operations Research Quarterly *20*, 499-500.

hast71 [95] Hastings,N.A.J. [1971]: "Bounds on the gain of a Markov decision process", Operatios Research *19*, 240-243.

hast76 [96] Hastings,N.A.J. [1976]: "A test for nonoptimal actions in undiscounted finite Markov decision chains", Management Science *23*, 87-92.

hast73 [97] Hastings,N.A.J. and J.M.C.Mello [1973]: "Tests for nonoptimal actions in discounted Markov decision problems", Management Science *19*, 1019-1022.

hast79 [98] Hastings,N.A.J. and D.Sadjani [1979]: "Markov programming with policy constraints", European Journal of Operations Research *3*, 253-255.

hast77 [99] Hastings,N.A.J. 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.

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

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

her89 [102] Hernández-Lerma,O. 1987]: "Adaptive Markov control processes", Springer-Verlag, New York.

her96 [103] Hernández-Lerma,O. and J. B. Lasserre [1996]: "Discrete-time Markov control processes: Basic optimality criteria", Springer-Verlag, New York.

her99 [104] Hernández-Lerma,O. and J. B. Lasserre [1999]: "Further topics on discrete-time Markov control processes", Springer-Verlag, New York.

herz [105] Herzberg,M. 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.

hey84 [106] Heyman,D.P. and M. J. Sobel [1984]: "Stochastic models in Operations Research, Volume II, MacGraw-Hill, New York.

hin70 [107] Hinderer,K. [1970]: "Foundations of non-stationary dynamic programming with discrete time parameter", Springer-Verlag, New York.

hol86a [108] Holzbaur,U.D. [1986a]: "Entscheidungsmodelle über angeordneten Körpern", Optimization *17*, 515-524.

hol86b [109] Holzbaur,U.D. [1986b]: "Sensitivitätsanalysen in Entscheidungsmodellen", Optimization *17*, 525-533.

hol94 [110] Holzbaur,U.D. [1994]: "Bounds for the quality and the number of steps in Bellman's value iteration algorithm", OR Spektrum *15*, 231-234.

hor71 [111] Hordijk,A. [1971]: "A sufficient condition for the existence of an optimal policy with respect to the average cost criterion in Markovian decision processes", Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Prague, 263-274.

hor74 [112] Hordijk,A. [1974]: "Dynamic programming and Markov potential theory", Mathematical Centre Tract 51, Amsterdam.

hor88 [113] Hordijk,A. and R.Dekker [1988]: "Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards", Mathematics of Operations Research *13*, 395-420.

hor85 [114] Hordijk,A., R.Dekker and L.C.M.Kallenberg [1985]: "Sensitivity-analysis in discounted Markovian decision problems", OR Spektrum *7*, 143-151.

hor79 [115] Hordijk,A. and L.C.M.Kallenberg [1979]: "Linear programming and Markov decision chains", Management Science *25*, 352-362.

hor84a [116] Hordijk,A. 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.

hor84b [117] Hordijk,A. and L.C.M.Kallenberg [1984b]: "Constrained undiscounted stochastic dynamic programming", Mathematics of Operations Research *9*, 276-289.

hor94 [118] Hordijk,A. 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.

hortym74 [119] Hordijk,A. and H.C.Tijms [1974]: "The method of successive approximations and Markovian decision problems", Operations Research *22*, 519-521.

hor75 [120] Hordijk,A. and H.C.Tijms [1975]: "A modified form of the iterative method of dynamic programming", Annals of Statictics *3*, 203-208.

hortym75 [121] Hordijk,A. and H.C.Tijms [1975]: "On a conjecture of Iglehart", Management Science *11*, 1342-1345.

ho60 [122] Howard,R.A. [1960]: "Dynamic programming and Markov processes", MIT Press, Cambridge.

ho63 [123] Howard,R.A. [1963]: "Semi-Markovian decision processes", Proceedings International Statistical Institute, Ottawa, Canada.

huang [124] Huang,Y. and L.C.M.Kallenberg [1994]: On finding optimal policies for Markov decision chains: a unifying framework for mean-variance trade-offs", Mathematics of Operations Research *19*, 434-448.

hub77 [125] Hübner,G. [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.

hub88 [126] Hübner,G. [1988]: "A unified approach to adaptive control of average reward Markov decision processes", OR Spektrum *10*, 161-166.

igle [127] Iglehart,D. [1963]: "Optimality of (s, S)-policies in the infinite horizon dynamic inventory problem", Management Science *9*, 259-267.

ishi [128] Ishikida,T. and P.Varaiya [1994]: "Multi-armed bandit problem revisited", Journal of Optimization Theory and Applications *83*, 113-154.

jero [129] Jeroslow,R.G. [1972]: "An algorithm for discrete dynamic programming with interest rates near zero", Management Science Research Report no. 300, Carnegie-Mellon University, Pittsburgh.

jew63a [130] Jewell,W.S. [1963a]: "Markov renewal programming. I: Formulation, finite return models", Operations Research *11*, 938-948.

jew63b [131] Jewell,W.S. [1963b]: "Markov renewal programming. II: Infinite return models, example", Operations Research *11*, 949-971.

kal81a [132] Kallenberg,L.C.M. [1981a]: "Finite horizon dynamic programming and linear programming", Methods of Operations Research *43*, 105-112.

kal81b [133] Kallenberg,L.C.M. [1981b]: "Unconstrained and constrained dynamic programming over a finite horizon", Report, University of Leiden, The Netherlands.

kal81c [134] Kallenberg,L.C.M. [1981c]: "Linear programming to compute a bias-optimal policy", in: B.Fleischmann et al. (eds.) "Operations Research Proceedings", 433-440.

kal83 [135] Kallenberg,L.C.M. [1983]: "Linear programming and finite Markovian control problems", Mathematical Centre Tract 148, Mathematical Centre, Amsterdam.

kal86 [136] Kallenberg,L.C.M. [1986]: "Note on M.N.Katehakis and Y.-R.Chen's computation of the Gittins index", Mathematics of Operations Research *11*, 184-186.

kal92 [137] Kallenberg, L.C.M. [1992]: " Separable Markovian decision problem: the linear programming method in the multichain case", OR Spektrum *14*,43-52.

kal99 [138] 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).

kao [139] Kao.P.C. [1973]: "Optimal replacement rules when the changes of states are semi-Markovian", Operations Research *21*, 1231-1249.

kate84 [140] Katehakis,M.N. and C.Derman [1984]: "Optimal repair allocation in a series system", Mathematics of Operations Research *9*, 615-623.

kate89 [141] Katehakis,M.N. and C.Derman [1989]: "On the maintenance of systems composed of highly reliable components", Management Science *35*, 551-560.

kate87 [142] Katehakis,M.N. and A.F.Veinott Jr. [1987]: "The multi-armed bandit problem: decomposition and computation", Mathematics of Operations Research *12*, 262-268.

kawai87 [143] Kawai,H. [1987]: "A variance minimization problem for a Markov decision process", European Journal of Operational Research *31*, 140-145.

kaka87 [144] Kawai,H. and N.Katoh [1987]: "Variance constrained Markov decision process", Journal of the Operations Research Society of Japan *30*, 88-100.

kem [145] Kemeny,J.G. and J.L.Snell [1960]: "Finite Markov chains", Van Nostrand, Princeton.

klein [146] Klein,M. [1962]: "Inspection-maintenance-replacement schedules under Markovian deterioration", Management Science *9*, 25-32.

kol [147] Kolesar,P. [1966]: "Minimum-cost replacement under Markovian deterioration", Management Science *12*, 694-706.

kur [148] Kurano,M. [1983]: "Adaptive policies in Markov decision processes with uncertain transition matrices", Journal of Information and Optimization Sciences *4*, 21-40.

kush71 [149] Kushner,H. [1971]: "Introduction to stochastic control", Holt, Rineholt and Winston, New York.

kuke71 [150] Kushner,H. and A.J.Keinmann [1971]: "Accelerated procedures for the solution of discrete Markov control problems", IEEE Transactions on Automatic Control *16*, 147-152.

lan [151] Lanery,E. [1967]: "Etude asymptotique des systèmes Markovien à commande", Revue d'Informatique et Recherche Operationelle *1*, 3-56.

las94a [152] Lasserre,J.B. [1994a]: "A new policy iteraton scheme for Markov decision processes using Schweitzer's formula", Journal of Applied Probability *31*, 268-273.

las94b [153] Lasserre,J.B. [1994b]: "Detecting optimal and non-optimal actions in average- cost Markov decision processes", Journal of Applied Probability *31*, 979-990.

lin [154] Lin,W. and P.R.Kumar [1984]: "Optimal control of a queueing system with two heterogeneous servers", IEEE Tansactions on Automatic Control AC-29, 696-705.

lipp [155] Lippman,S.A. [1969]: "Criterion equivalence in discrete dynamic programming", Operations Research *17*, 920-923.

liu94 [156] Liu,J.Y. and K.Liu [1994]: "An algorithm on the Gittins index", Systems Science and Mathematical Science *7*, 106-114.

liu92 [157] Liu,Q.-S. 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.

liuoh [158] Liu,Q.-S., K.Ohno and H.Nakayama [1992]: "Multi-objective discounted Markov processes with expectation and variance criteria", International Journal of System Science *23*, 903-914.

loe [159] Loeve,J.A. [1995]: "Markov decision chains with partial information", PhD dissertation, University of Leiden, The Netherlands.

lov87 [160] Lovejoy,W.S. [1987]: "Some monotonicity results for partially observed Markov processes", Operations Research *35*, 736-743.

lov91a [161] Lovejoy,W.S. [1991a]: "Computationally feasible bounds for partially observed Markov decision processes", Operations Research *39*, 162-175.

lov91b [162] Lovejoy,W.S. [1991b] "A survey of algorithmic methods for partially observed Markov decision processes", Annals of Operations Research *28*, 47-66.

mac66 [163] Macqueen,J. [1966]: "A modified programming method for Markovian decision problems", Journal of Mathematical Analysis and Applications *14*, 38-43.

mac67 [164] Macqueen,J. [1967]: "A test for suboptimal actions in Markov decision problems", Operations Research *15*, 559-561.

manne60 [165] Manne,A.S. [1960]: "Linear programming and sequential decisions", Management Science, 259-267.

meis [166] Meister,U. and U.Holzbaur [1986]: "A polynomial time bound for Howard's policy improvement algorithm", OR Spektrum *8*, 37-40.

mil [167] Miller,B.L. and A.F.Veinott Jr. [1969]: "Discrete dynamic programming with a small interest rate", Annals of Mathematical Statistics *40*, 366-370.

mine70 [168] Mine,H. and S. Osaki [1970]: "Markov decision processes", American Elsevier, New York.

mon [169] Monahan,G.E. [1982]: "A survey of partially observable Markov decision processes: theory, models and algorithms", Management Science *28*, 1-16.

mor [170] Morton,T. [1971]: "Undiscounted Markov renewal programming via modified successive approximations", Operations Research *19*, 1081-1089.

naz [171] Nazareth,J.L. and R.B.Kulkarni [1986]: "Linear programming formulations of Markov decision processes", OR Letters *5*, 13-16.

ng [172] Ng,M.K. [1999]: "A note on policy iteration algorithms for discounted Markov decision problems", OR Letters *25*, 195-197.

od [173] Odoni,A. [1969]: "On finding the maximal gain for Markov decision processes", Operations Research *17*, 857-860.

oez [174] Oezekici,S. [1988]: "Optimal periodic replacement of multicomponent reliability systems", Operations Research *36*, 542-552.

ohno [175] Ohno,K. [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.

osa [176] Osaki,S. and H.Mine [1968]: "Linear programming algorithms for semi-Markovian decision processes", Journal of Mathematical Analysis and Applications *22*, 356-381.

part [177] Parthasarathy,T., S.H.Tijs and O.J.Vrieze [1984], "Stochastic games with state independent transitions and reparable rewards. 262-271 in: G.Hammer and D.Pallaschke (eds), Selected Topics in Operations Research and Mathematical Economics.

pla [178] Platzman,L.K. [1977]: "Improved conditions for convergence in undiscounted Markov renewal programming", Operations Research *25*, 529-533.

pol [179] Pollatschek,M.A. and B.Avi-Itzhak [1969]: "Algorithms for stochastic games with geometric interpretation", Management Science *15*, 399-415.

por71 [180] Porteus,E.L. [1971]: "Some bounds for discounted sequential decision processes", Management Science *18*, 7-11.

por75 [181] Porteus,E.L. [1975]: "Bounds and transformations for discounted finite Markov decision chains", Operations Research *23*, 761-784.

por80a [182] Porteus,E.L. [1980a]: "Improved iterative computation of the expected return in Markov and semi-Markov chains", Zeitschrift für Operations Research *24*, 155-170.

por80b [183] Porteus,E.L. [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.

por81 [184] Porteus [1981]: "Computing the discounted return in Markov and semi- Markov chains", Naval Research Logistics Quarterly *28*, 567-577.

por78 [185] Porteus,E.L. and J.C.Totten [1978]: "Accelerated computation of the expected discounted return in a Markov chain", Operations Research *26*, 350-358.

put81 [186] Puterman,M.L. [1981]: "Computational methods for Markov decision methods", Proceedings of 1981 Joint Automatic Control Conference.

put94 [187] Puterman,M.L. [1994]: "Markov decision processes", Wiley, New York.

put79 [188] Puterman,M.L. and S.L.Brumelle [1979]: "On the convergence of policy iteration in stationary dynamic programming", Mathematics of Operations Research *4*, 60-69.

put78 [189] Puterman,M.L. and M.C.Shin [1978]: "Modified policy iteration algorithms for discounted Markov decision chains", Management Science *24*, 1127-1137.

put82 [190] Puterman,M.L. and M.C.Shin [1982]: "Action elimination procedures for modified policy iteration algorithms" Operations Research *30*, 301-318.

ree73 [191] Reetz,D. [1973]: "Solution of a Markovian decision problem by successive overrelaxation", Zeitschrift für Operations Research *17*, 29-32.

ree76 [192] Reetz,D. [1976]: "A decision exclusion algorithm for a class of Markovian decision processes" , Zeitschrift für Operations Research *20*, 125-131.

rie [193] Rieder,U. [1991]: "Structural results for partially observed control problems", Zeitschrift für Operations Research *35*, 473-490.

rig [194] Righter,R. [1994]: "Scheduling", in: Shaked,M. and J.G.Shantikumar (eds.), "Stochastic orders and their applications", Academic Press, 381-432.

roo [195] Roosta,M. [1982]: "Routing through a network with maximum reliability", Journal of Mathematical Analysis and Applications *88*, 341-347.

ross89 [196] Ross,K.W. [1989]: "Randomized and past-dependent policies for Markov decision processes with multiple constraints", Operations Research *37*, 474-477.

ross91 [197] Ross,K.W. and R.Varadarajan [1991]: "Multichain Markov decision processes with a sample path constraint: a decomposition approach", Mathematics of Operations Research *16*, 195-207.

ross69 [198] Ross,S.M. [1969]: "A problem in optimal search and stop", Operations Research *17*, 984-992.

ross70 [199] Ross,S.M. [1970]: "Applied probability models with optimization applications", Holden-Day, San Francisco.

ross74 [200] Ross,S.M. [1974]: "Dynamic programming and gambling models", Advances in Applied Probability *6*, 593-606.

ross83 [201] Ross,S.M. [1983]: "Introduction to stochastic dynamic programming", Academic Press, New York.

rot [202] Rothblum,U.G. [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.

sca [203] Scarf,H. [1960]: "The optimality of (s, S) polcies 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.

sche [204] Schellhaas,H. [1974]: "Zur extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung", Zeitschrift für Operations Research *18*, 91-104.

schm [205] Schmitz,N. [1985]: "How good is Howard's policy improvement algorithm?", Zeitschrift fur Operations Research *29*, 315-316.

schr [206] Schrage,L. [1968]: "A proof of the optimality of the shortest remaining processing time discipline", Operations Research *16*, 687-690.

schw65 [207] Schweitzer,P.J. [1965]: "Perturbation theory and Markovian decision processes", Ph.D. dissertation, M.I.T., Operations Research Center Report 15.

schw68 [208] Schweitzer,P.J. [1968]: "Perturbation theory and finite Markov chains" Journal of Applied Probability *5*, 401-413.

schw71a [209] Schweitzer,P.J. [1971a]: "Multiple policy improvements in undiscounted Markov renewal programming", Operations Research *19*. 784-793.

schw71b [210] Schweitzer,P.J. [1971b]: "Iterative solution of the functional equations of undiscounted Markov renewal programming", Journal of Mathematical Analysis and Applications *34*, 495-501.

schw84 [211] Schweitzer,P.J. [1984]: "A value-iteration scheme for undiscounted multichain Markov renewal programs", ZOR - Zeitschrift für Operations Research *28*, 143-152.

schw85 [212] Schweitzer,P.J. [1985]: "The variational calculus and approximations in policy space for Markov decision processes", Journal of Mathematical Analysis and Applications *110*, 568-582.

schw87 [213] Schweitzer,P.J. [1987]: "A Brouwer fixed-point mapping approach to communicating Markov decision processes", Journal of Mathematical Analysis and Applications *123*, 117-130.

schw91 [214] Schweitzer,P.J. [1991]: Block-scaling of value-iteration for discounted Markov renewal programming", Annals of Operations Research *29*, 603-630.

schw77 [215] Schweitzer,P.J. and A.Federgruen [1977]: "The asymptotic behavior of value iteration in Markov decision problems", Mathematics of Operations Research *2*, 360-381.

schw78a [216] Schweitzer,P.J. and A.Federgruen [1978a]: "Foolproof convergence in multichain policy iteration", Journal of Mathematical Analysis and Applications *64*, 360-368.

schw78b [217] Schweitzer,P.J. and A.Federgruen [1978b]: "The functional equations of undiscounted Markov renewal programming", Mathematics of Operations Research *3*, 308-321.

schw79 [218] Schweitzer,P.J. and A.Federgruen [1979]: "Geometric convergence of value iteration in multichain Markov decision problems", Advances of Applied Probabilty *11*, 188-217.

sen99 [219] Sennott,L.I. [1999]: "Stochastic dynamic programming and the control of queueing systems", Wiley, New York.

ser [220] Sernik,E.L. 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.

sha [221] Shapiro,J.F. [1975]: "Brouwer's fixed point theorem and finite state space Markovian decision theory", Journal of Mathematical Analysis and Applications *49*, 710-712.

sh [222] Shapley,L.S. [1953]: "Stochastic games", Proceedings of the National Academy of Sciences, 1095-1100.

she [223] Sherif,Y.S. and M.L.Smith [1981]: "Optimal maintenance policies for systems subject to failure - A review", Naval Research Logistics Quarterly *28*, 47-74.

sla [224] Sladky,K. [1974]: "On the set of optimal controls for Markov chains with rewards", Kybernatika *10*, 350-367.

sma66 [225] Smallwood,R.D. [1966]: "Optimum policy regions for Markov processes with discounting", Operations Research *14*, 658-669.

sma73 [226] Smallwood,R.D. and E.Sondik [1973]: "The optimal control of partially observable Markov processes over a finite horizon", Operations Research *21*, 1071-1088.

smi [227] Smith,D.R. [1978]: "Optimal repairman allocation - asymptotic results", Management Science *24*, 665-674.

so81 [228] Sobel,M.J. [1981], "Myopic solutions of Markov decision processes and stochastic games", Operations Research *29*, 995-1009.

so85 [229] Sobel,M.J. [1985]: "Maximal mean/standard deviation ratio in an undiscounted MDP", OR Letters *4*, 157-159.

so94 [230] Sobel,M.J. [1994]: "Mean-variance trade-offs in an undiscounted MDP", Operations Research *42*, 175-183.

sondik [231] Sondik,E. [1978]: "The optimal control of partially observable Markov processes over the infinite horizon: discounted costs", Operations Research *26*, 282-304.

sonin [232] Sonin, I.M. [1999]: The elimination algorithm for the problem of optimal stopping", Mathematical Methods of Operations Research *49*, 111-124.

spr [233] Spreen,D. [1981]: "A further anti-cycling rule in multi-chain policy iteration for undiscounted Markov renewal programs", Zeitschrift für Operations Research *25*, 225-234.

st [234] Stein,J. [1988]: "On efficiency of linear programming applied to discounted Markovian decision problems", OR Spektrum *10*, 153-160.

stid85 [235] Stidham,S.S. Jr. [1985]: "Optimal control of admission to a queueing system", IEEE Transactions on Automatic Control *AC-30*, 705-713.

stid93 [236] Stidham,S.S. Jr. and R.R.Weber [1993]: "A survey of Markov decision models for control of networks of queues", Queueing Systems *13*, 291-314.

stoer [237] Stoer,J. and R.Bulirsch [1980]: "Introduction to numerical analysis", Springer-Verlag, New York.

str [238] Strauch,R. and A.F.Veinott Jr. [1966]: "A property of sequential control processes", Report, Rand McNally, Chicago.

sun [239] Sun,M. [1993]: "Revised simplex algorithm for finite Markov decision processes", Journal of Optimization Theory and Applications *79*, 405-413.

th81 [240] Thomas,L.C. [1981]: "Second order bounds for Markov decision processes", Journal of Mathematical Analysis and Applications *80*, 294-297.

th83 [241] Thomas,L.C. [1983]: "Constrained Markov decision processes as multi-objective problems", in: "Multi-objective decision making", Academic Press, 77-94.

ty [242] Tijms,H.C. [1986]: "Stochastic modelling and analysis: a computational approach", Wiley, Chichester.

ts86 [243] Tsitsiklis,J.N. [1986]: "A lemma on the multi-armed bandit problem", IEEE Transactions on Automatic Control *31*, 576-577.

ts93 [244] Tsitsiklis,J.N. [1993]: "A short proof of the Gittins index theorem", Annals of Applied Probability *4*, 194-199.

duyn [245] Van der Duyn Schouten,F.A. 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.

wal80 [246] Van der Wal,J. [1980]: "The method of value oriented successive approximations for the average reward Markov decision processes", OR Spektrum *1*, 233-242.

wal81 [247] Van der Wal,J. [1981]: "Stochastic dynamic programming", Mathematical Centre Tract 139, Mathematical Centre, Amsterdam.

hee78 [248] Van Hee, K.M. [1978]: "Markov strategies in dynamic programming", Mathematics of Operations Research *3*, 191-201.

hee [249] Van Hee,K.M., 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.

nunen76a [250] Van Nunen,J.A.E.E. [1976a]: "A set of successive approximation method for discounted Markovian decision problems", Zeitschrift für Operations Research *20*, 203-208.

nunen76b [251] Van Nunen,J.A.E.E. [1976b]: "Contracting Markov decision processes", Mathematical Centre Tract 71, Mathematical Centre, Amsterdam.

nunen76c [252] Van Nunen,J.A.E.E. [1976c]: "Improved successive approximation methods for discounted Markovian decision processes", in: A.Prekopa (ed.), "Progress in Operations Research", North Holland, Amsterdam, 667-682.

nunen76 [253] Van Nunen,J.A.E.E. 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.

nunen77 [254] Van Nunen,J.A.E.E. 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.

var [255] Varaiya,P.P., 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.

vei66a [256] Veinott,A.F.Jr. [1966a]: "On the optimality of (s, S) inventory policies: new condition and a new proof", SIAM Journal on Applied Mathematics *14*, 1067-1083.

vei66b [257] Veinott,A.F. Jr. [1966b]: "On finding optimal policies in discrete dynamic programming with no discounting", Annals of Mathematical Statistics *37*, 1284-1294.

vei69 [258] Veinott,A.F. Jr. [1969]: "Discrete dynamic programming with sensitive discount optimality criteria", Annals of Mathematical Statistics *40*, 1635-1660.

vei74 [259] Veinott,A.F. 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.

ver [260] Vergin,R.C. and M.Scribian [1977]: "Maintenance scheduling for multicomponent equipment", AIIE Transactions *9*, 297-305.

vrieze87 [261] Vrieze,O.J. [1987]: "Stochastic games with finite state and action spaces", CWI Tract 33, Centre for Mathematics and Computer Science, Amsterdam.

wak92 [262] Wakuta,K. [1992]: "Optimal stationary policies in the vector-valued Markov decision process", Stochastic Processes and its Applications *42*, 149-156.

wak95 [263] Wakuta,K. [1995]: "Vector-valued Markov decision processes and the systems of linear inequalities ", Stochastic Processes and its Applications *56*, 159-169.

wak96 [264] Wakuta,K. [1996]: "A new class of policies in vector-valued Markov decision processes", Journal of Mathematical Analysis and Applications *202*, 623-628.

wak99 [265] Wakuta,K. [1999]: "A note on the structure of value spaces in vector-valued Markov decision processes", Mathematical Methods of Operations Research *49*, 77-86.

wal [266] Walrand,J. [1988]: "An introduction to queueing networks", Prentice-Hall, Englewood Cliffs, New Jersey.

web82 [267] Weber,R.R. [1982]: "Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime.

web92 [268] Weber,R.R. [1992]: "On the Gittins index for multi-armed bandits", Annals of Applied Probabitily *2*, 1024-1033.

web87 [269] Weber,R.R. and S.S.Stidham Jr. [1987]: "Optimal control of services rates in networks of queues", Advances in Applied Probability *19*, 202-218.

wei82 [270] Weiss,G. [1982]: "Multiserver stochastic scheduling", in: Dempster,M.A.H,, J.K.Lenstra and A.H.G. Rinnooy Kan (eds.), "Deterministic and stochastic scheduling", Reidel, Dordrecht, Holland, 157-179.

wei88 [271] Weiss,G. [1988]: "Branching bandit processes", Probability in the Engineering and Information Sciences *2*, 269-278.

wes77 [272] Wessels,J. [1977]: "Stopping times and Markov programming", in: "Transactions of the 7th Prague conference on information theory, statistical decision functions and random processes, Academia, Prague, 575-585.

wes75 [273] Wessels,J. and J.A.E.E. Van Nunen [1975]: "Discounted semi-Markov decision processes: linear programming and policy iteration", Statistica Neerlandica *29*, 1-7.

whi76 [274] White,C.C. III [1976]: "Procedures for the solution of a finite-horizon, partially observed, semi-Markov optimization problem", Operations Research *24*, 348-358.

whi91 [275] White,C.C. III [1991]: "A survey of solution techniques for the partially observed Markov decision process", Annals of Operations Research *33*, 215-230.

whi89 [276] White,C.C. III and W.T.Scherer [1989]: "Solution procedures for partially observed Markov decision processes", Operations Research *37*, 791-797.

whi94 [277] White,C.C. III and W.T.Scherer [1994]: "Finite-memory suboptimal design for partially observed Markov decision processes", Operations Research *42*, 439-455.

white63 [278] White,D.J. [1963]: "Dynamic programming, Markov chains, and the method of successive approximations", Journal of Mathematical Analysis and Applications *6*, 373-376.

white78 [279] White,D.J. [1978]: "Elimination of non-optimal actions in Markov decision processes", in: M.L.Puterman (ed.) Dynanmic programming and its applications, Academic Press, New York, 131-160.

white82 [280] White,D.J. [1982]: "Multi-objective infinite-horizon discounted Markov decision processes", Journal of Mathematical Analysis and Applications *89*, 639-647.

white85 [281] White,D.J. [1985]: "Real applications of Markov decision theory", Interfaces *15:6*, 73-83.

white881 [282] White,D.J. [1988]: "Further real applications of Markov decision theory", Interfaces *18:5*, 55-61.

white882 [283] White,D.J. [1988]: "Mean, variance and probabilistic criteria in finite Markov decision processes: a review", Journal of Optimization Theory and Applications *56*, 1-30.

white92 [284] White,D.J. [1992]: "Computational approaches to variance-penalized Markov decision processes", OR Spektrum *14*, 79-83.

white931 [285] White,D.J. [1993]: "A survey of applications of Markov decision processes", Journal of the Operational Research Society *44*, 1073-1096.

white932 [286] White,D.J. [1993]: "Markov decision processes", Wiley, Chichester.

white94 [287] White,D.J. [1994]: "A mathematical programming approach to a problem in variance penalised Markov decision processes", OR Spektrum *15*, 225-230.

white95 [288] White,D.J. [1995]: "A superharmonic approach to solving infinite horizon partially observable Markov decision problems", Mathematical Methods of Operations Research *41*, 71-88.

whittle80 [289] Whittle,P. [1980]: "Multi-armed bandits and the Gittins index", Journal of the Royal Statistical Society, Series B *42*, 143-149.

whittle82a [290] Whittle,P. [1982]: "Optimization over time; dynamic programming and stochastic control", Volume I, Wiley, New York.

whittle82b [291] Whittle,P. [1982]: "Optimization over time; dynamic programming and stochastic control", Volume II, Wiley, New York.

yas [292] Yasuda,M. [1988]: "The optimal value of Markov stopping problems with one-step look ahead policy", Journal of Applied Probability *25*, 544-552.

zhe [293] Zheng,Y.-S. and A.Federgruen [1991]: "Finding optimal (s, S)-policies is about as simple as evaluating a single policy", Operations Research *39*, 654-665.