2 FINITE STATE AND ACTION MDPS

Lodewijk Kallenberg

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

2.1 INTRODUCTION

2.1.1 Origin

Bellman's book [13], can be considered as the starting point of Markov decision processes (MDPs). However, already in 1953, Shapley's paper [221] on stochastic games includes as a special case the value iteration method for MDPs, but this was recognized only later on. About 1960 the basics for the other computational methods (policy iteration and linear programming) were developed in publications like Howard [121], De Ghellinck [42], D'Epenoux [55], Manne [164] and Blackwell [27]. Since the early sixties, many results on MDPs are published in numerous journals, monographs, books and proceedings. Thousands of papers were published in scientific journals. There are about fifty books on MDPs. Around 1970 a first series of books was published. These books (e.g. Derman [58], Hinderer [107], Kushner [148], Mine and Osaki [167] and Ross [198]) contain the fundamentals of the theory of finite MDPs. Since that time nearly every year one or more MDP-books appeared. These books cover special topics (e.g. Van Nunen [250], Van der Wal [246], Kallenberg [134], Federgruen [69], Vrieze [260], Hernández-Lerma [102], Altman [2] and Sennott [218]) or they deal with the basic and advanced theory of MDPs (e.g. Bertsekas [15], Whittle [289], [290], Ross [200], Dietz and Nollau [63], Bertekas [17], Denardo [50], Heyman and Sobel [106], White [285], Puterman [186], Bertsekas [18], [19], Hernández-Lerma and Lasserre [103], [104], and Filar and Vrieze [79].

2.1.2 The model

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

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

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

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

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

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

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

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

2.1.3 Optimality criteria

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a) the lower limit of the average expected reward:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.1.4 Applications

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

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

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

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

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

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

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

Inventory problems

In inventory problems an optimal balance between inventory costs and ordering costs has to be determined. We assume that the probability distribution of the demand is known. There are different variants of the inventory problem. They differ, for instance, in the following aspects:
- stationary or nonstationary costs and demands;
- a finite planning horizon or an infinite planning horizon;
- backlogging or no backlogging.

For all these variants different performance measures may be considered.
In many inventory models the optimal policy is of (s, S)-type, i.e. when the inventory is smaller than or equal to s, then replenish the stock to level S. The existence of optimal (s, S)-policies in finite horizon models with fixed cost K is based on the so-called K-convexity, introduced by Scarf [202]. The existence of an optimal (s, S)-policy in the infinite horizon model is shown by Iglehart [126]. Another related paper is Veinott [255]. For the relation between discounted and average costs we refer to Hordijk and Tijms [119]. For the computation of the values s and S we refer to papers like Federgruen and Zipkin [76], and Zheng and Federgruen [292].

Optimal control of queues

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

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

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

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

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

2.2 FINITE HORIZON

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

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

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

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

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

Algorithm I (finite horizon)

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

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