Chapter 2
Theory of Finite Horizon Markov Decision Processes

In this chapter we will establish the theory of Markov Decision Processes with a finite time horizon and with general state and action spaces. Optimization problems of this kind can be solved by a backward induction algorithm. Since state and action space are arbitrary, we will impose a structure assumption on the problem in order to prove the validity of the backward induction and the existence of optimal policies. The chapter is organized as follows.
Section 2.1 provides the basic model data and the definition of policies. The precise mathematical model is then presented in Section 2.2 along with a sufficient integrability assumption which implies a well-defined problem. The solution technique for these problems is explained in Section 2.3. Under structure assumptions on the model it will be shown that Markov Decision Problems can be solved recursively by the so-called *Bellman equation*. The next section summarizes a number of important special cases in which the structure assumption is satisfied. Conditions on the model data are given such that the value functions are upper semicontinuous, continuous, measurable, increasing, concave or convex respectively. Also the monotonicity of the optimal policy under some conditions is established. This is an essential property for computations. Finally the important concept of upper bounding functions is introduced in this section. Whenever an upper bounding function for a Markov Decision Model exists, the integrability assumption is satisfied. This concept will be very fruitful when dealing with infinite horizon Markov Decision Problems in Chapter 7. In Section 2.5 the important case of stationary Markov Decision Models is investigated. The notion 'stationary' indicates that the model data does not depend on the time index. The relevant theory is here adopted from the non-stationary case. Finally Section 2.6 highlights the application of the developed theory by investigating three simple examples. The first example is a special card game, the second one a cash balance problem and the last one deals with the classical stochastic LQ-problems. The last section contains some notes and references.

N. Bäuerle and U. Rieder, *Markov Decision Processes with Applications to Finance*, Universitext, DOI 10.1007/978-3-642-18324-9_2,

2.1 Markov Decision Models

After having discussed the scope of Markov Decision Models informally in Chapter 1 we will now give a precise definition of a Markov Decision Model. This can be done by defining the ingredients or input data of the model in mathematical terms.

Definition 2.1.1. A (non-stationary) *Markov Decision Model* with planning horizon $N \in \mathbb{N}$ consists of a set of data $(E, A, D_n, Q_n, r_n, g_N)$ with the following meaning for $n = 0, 1, \ldots, N-1$:

- E is the *state space*, endowed with a σ-algebra $\mathfrak{E}$. The elements (states) are denoted by $x \in E$.
- A is the *action space*, endowed with a σ-algebra $\mathfrak{A}$. The elements (actions) are denoted by $a \in A$.
- $D_n \subset E \times A$ is a measurable subset of $E \times A$ and denotes the set of possible state-action combinations at time n. We assume that D_n contains the graph of a measurable mapping $f_n : E \to A$, i.e. $(x, f_n(x)) \in D_n$ for all $x \in E$. For $x \in E$, the set $D_n(x) = \{a \in A \mid (x, a) \in D_n\}$ is the set of *admissible actions* in state x at time n.
- Q_n is a stochastic transition kernel from D_n to E, i.e. for any fixed pair $(x, a) \in D_n$, the mapping $B \mapsto Q_n(B|x, a)$ is a probability measure on $\mathfrak{E}$ and $(x, a) \mapsto Q_n(B|x, a)$ is measurable for all $B \in \mathfrak{E}$. The quantity $Q_n(B|x, a)$ gives the probability that the next state at time $n+1$ is in B if the current state is x and action a is taken at time n. Q_n describes the *transition law*.
- $r_n : D_n \to \mathbb{R}$ is a measurable function. $r_n(x, a)$ gives the (discounted) *one-stage reward* of the system at time n if the current state is x and action a is taken.
- $g_N : E \to \mathbb{R}$ is a measurable mapping. $g_N(x)$ gives the (discounted) *terminal reward* of the system at time N if the state is x.

Remark 2.1.2. a) In many applications the state and action spaces are Borel subsets of Polish spaces (i.e. complete, separable, metric spaces) or finite or countable sets. The σ-algebras $\mathfrak{E}$ and $\mathfrak{A}$ are then given by the σ-algebras $\mathcal{B}(E)$ and $\mathcal{B}(A)$ of all Borel subsets of E and A respectively. Often in applications E and A are subsets of $\mathbb{R}^d$ or $\mathbb{R}^d_+$.

b) If the one-stage reward function r'_n also depends on the next state, i.e. $r'_n = r'_n(x, a, x')$, then define

$$r_n(x, a) := \int r'_n(x, a, x') Q_n(dx'|x, a).$$

c) Often D_n and Q_n are independent of n and $r_n(x, a) := \beta^n r(x, a)$ and $g_N(x) := \beta^N g(x)$ for a (discount) factor $\beta \in (0, 1]$. In this case the Markov Decision Model is called *stationary* (see Section 2.5). ◊

The stochastic transition law of a Markov Decision Model is often given by a *transition* or *system function*. To make this more precise, suppose that $Z_1, Z_2, \ldots, Z_N$ are random variables with values in a measurable space $(\mathcal{Z}, \mathfrak{Z})$. These random variables are called *disturbances*. Z_{n+1} influences the transition from the state at time n of the system to the state at time $n+1$. The distribution Q_n^Z of Z_{n+1} may depend on the current state and action at time n such that $Q_n^Z(\cdot|x,a)$ is a stochastic kernel for $(x,a) \in D_n$. The new state of the system at time $n+1$ can now be described by a *transition* or *system function* $T_n : D_n \times \mathcal{Z} \to E$ such that

$$x_{n+1} = T_n(x_n, a_n, z_{n+1}).$$

Thus, the transition law of the Markov Decision Model is here determined by T_n and Q_n^Z.

Theorem 2.1.3. *A* Markov Decision Model *is equivalently described by the set of data* $(E, A, D_n, \mathcal{Z}, T_n, Q_n^Z, r_n, g_N)$ *with the following meaning:*

- E, A, D_n, r_n, g_N *are as in Definition 2.1.1.*
- $\mathcal{Z}$ *is the* disturbance space*, endowed with a* σ*-algebra* $\mathfrak{Z}$*.*
- $Q_n^Z(B|x,a)$ *is a stochastic transition kernel for* $B \in \mathfrak{Z}$ *and* $(x,a) \in D_n$ *and* $Q_n^Z(B|x,a)$ *denotes the probability that* Z_{n+1} *is in* B *if the current state is* x *and action* a *is taken.*
- $T_n : D_n \times \mathcal{Z} \to E$ *is a measurable function and is called* transition *or* system function*.* $T_n(x,a,z)$ *gives the next state of the system at time* $n+1$ *if at time* n *the system is in state* x*, action* a *is taken and the disturbance* z *occurs at time* $n+1$*.*

Proof. Suppose first a Markov Decision Model as in Definition 2.1.1 is given. Obviously we can choose $\mathcal{Z} := E$, $T_n(x,a,z) := z$ and $Q_n^Z(B|x,a) := Q_n(B|x,a)$ for $B \in \mathfrak{E}$. Conversely, if the data $(E, A, D_n, \mathcal{Z}, T_n, Q_n^Z, r_n, g_N)$ is given then by setting

$$Q_n(B|x,a) := Q_n^Z\Big(\{z \in \mathcal{Z} \mid T_n(x,a,z) \in B\} \;\Big|\; x,a\Big), \quad B \in \mathfrak{E},$$

we obtain the stochastic kernel of the Markov Decision Model. □

Let us next consider the consumption problem of Example 1.1.1 and set it up as a Markov Decision Model.

Example 2.1.4 (Consumption Problem; continued). Let us consider the consumption problem of Example 1.1.1. We denote by Z_{n+1} the random return of our risky asset over period $[n, n+1)$. Further we suppose that $Z_1, \ldots, Z_N$ are non-negative, independent random variables and we assume that the consumption is evaluated by utility functions $U_n : \mathbb{R}_+ \to \mathbb{R}$. The final capital is also evaluated by a utility function U_N. Thus we choose the following data:

- $E := \mathbb{R}_+$ where $x_n \in E$ denotes the wealth of the investor at time n,
- $A := \mathbb{R}_+$ where $a_n \in A$ denotes the wealth which is consumed at time n,
- $D_n(x) := [0, x]$ for all $x \in E$, i.e. we are not allowed to borrow money.
- $\mathcal{Z} := \mathbb{R}_+$ where z denotes the random return of the asset,
- $T_n(x_n, a_n, z_{n+1}) := (x_n - a_n)z_{n+1}$ is the transition function,
- $Q_n^Z(\cdot|x, a) :=$ distribution of Z_{n+1} (independent of (x, a)),
- $r_n(x, a) := U_n(a)$ is the one-stage reward,
- $g_N(x) := U_N(x)$. ♦

In what follows we assume that there is a fixed planning horizon $N \in \mathbb{N}$, i.e. N denotes the number of stages. Of course when we want to control a Markov Decision Process, due to its stochastic transitions, it is not reasonable to determine all actions at all time points at the beginning. Instead we have to react to random changes. Thus we have to choose a control at the beginning which takes into account the future time points and states.

Definition 2.1.5. a) A measurable mapping $f_n : E \to A$ with the property $f_n(x) \in D_n(x)$ for all $x \in E$, is called a *decision rule* at time n. We denote by F_n the set of all decision rules at time n.

b) A sequence of decision rules $\pi = (f_0, f_1, \ldots, f_{N-1})$ with $f_n \in F_n$ is called an *N-stage policy* or *N-stage strategy*.

Note that $F_n \neq \emptyset$ since by Definition 2.1.1 D_n contains the graph of a measurable mapping $f_n : E \to A$.

Remark 2.1.6 (Randomized Policies). It is sometimes reasonable to allow for *randomized policies* or decision rules respectively. A *randomized Markov policy* $\pi = (f_0, f_1, \ldots, f_{N-1})$ is given if $f_n(B|x)$ is a stochastic kernel with $f_n(D_n(x)|x) = 1$ for all $x \in E$. In order to apply such a policy we have to do a random experiment to determine the action. Randomized decision rules are related to *relaxed controls* or *Young measures* and are sometimes necessary to guarantee the existence of optimal policies (cf. Section 8.2). ◊

We consider a Markov Decision Model as an N-stage random experiment. Thus, in order to be mathematically precise we have to define the underlying probability space. The *canonical construction* is as follows. Define a measurable space $(\Omega, \mathcal{F})$ by

$$\Omega = E^{N+1}, \quad \mathcal{F} = \mathfrak{E} \otimes \ldots \otimes \mathfrak{E}.$$

We denote $\omega = (x_0, x_1, \ldots, x_N) \in \Omega$. The random variables $X_0, X_1, \ldots, X_N$ are defined on the measurable space $(\Omega, \mathcal{F})$ by

$$X_n(\omega) = X_n((x_0, x_1, \ldots, x_N)) = x_n,$$

being the n-th projection of ω. The random variable X_n represents the state of the system at time n and (X_n) is called *Markov Decision Process*. Suppose now that $\pi = (f_0, f_1, \dots, f_{N-1})$ is a fixed policy and $x \in E$ is a fixed initial state. According to the Theorem of Ionescu-Tulcea (see Appendix B) there exists a unique probability measure $\mathbb{P}_x^\pi$ on $(\Omega, \mathcal{F})$ with

(i) $\mathbb{P}_x^\pi(X_0 \in B) = \delta_x(B)$ for all $B \in \mathfrak{E}$.
(ii) $\mathbb{P}_x^\pi(X_{n+1} \in B | X_1, \dots, X_n) = \mathbb{P}_x^\pi(X_{n+1} \in B | X_n) = Q_n\big(B | X_n, f_n(X_n)\big)$.

Equation (ii) is the so-called *Markov property*, i.e. the sequence of random variables $X_0, X_1, \dots, X_n$ is a non-stationary Markov process with respect to $\mathbb{P}_x^\pi$. By $\mathbb{E}_x^\pi$ we denote the expectation with respect to $\mathbb{P}_x^\pi$. Moreover we denote by $\mathbb{P}_{nx}^\pi$ the conditional probability $\mathbb{P}_{nx}^\pi(\cdot) := \mathbb{P}^\pi(\cdot \mid X_n = x)$. $\mathbb{E}_{nx}^\pi$ is the corresponding expectation operator.

2.2 Finite Horizon Markov Decision Models

Now we have to impose an assumption which guarantees that all appearing expectations are well-defined. By $x^+ = \max\{0, x\}$ we denote the positive part of x.

Integrability Assumption ($\mathbf{A}_N$): For $n = 0, 1, \dots, N$

$$\delta_n^N(x) := \sup_\pi \mathbb{E}_{nx}^\pi \left[\sum_{k=n}^{N-1} r_k^+(X_k, f_k(X_k)) + g_N^+(X_N) \right] < \infty, \quad x \in E.$$

We assume that (A_N) holds for the N-stage Markov Decision Problems throughout the following chapters. Obviously Assumption (A_N) is satisfied if all r_n and g_N are bounded from above. The main results are even true under a weaker assumption than (A_N) (see Remark 2.3.14).

Example 2.2.1 (Consumption Problem; continued). In the consumption problem Assumption (A_N) is satisfied if we assume that the utility functions are increasing and concave and $\mathbb{E}\, Z_n < \infty$ for all n, because then r_n and g_N can be bounded by an affine-linear function $c_1 + c_2 x$ with $c_1, c_2 \geq 0$ and since $X_n \leq x Z_1 \dots Z_n$ a.s. under every policy, the function δ_n^N satisfies

$$\begin{aligned} \delta_n^N(x) &= \sup_\pi \mathbb{E}_{nx}^\pi \left[\sum_{k=n}^{N-1} U_k^+(f_k(X_k)) + U_N^+(X_N) \right] \\ &\leq N c_1 + x c_2 \sum_{k=n}^{N} \mathbb{E}\, Z_1 \dots \mathbb{E}\, Z_k < \infty, \quad x > 0 \end{aligned}$$

which implies the statement. ♦

We can now introduce the expected discounted reward of a policy and the N-stage optimization problem we are interested in. For $n = 0, 1, \ldots, N$ and a policy $\pi = (f_0, \ldots, f_{N-1})$ let $V_{n\pi}(x)$ be defined by

$$V_{n\pi}(x) := \mathbb{E}_{nx}^{\pi}\left[\sum_{k=n}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\right], \quad x \in E.$$

$V_{n\pi}(x)$ is the *expected total reward at time n over the remaining stages n to N* if we use policy π and start in state $x \in E$ at time n. The *value function* V_n is defined by

$$V_n(x) := \sup_{\pi} V_{n\pi}(x), \quad x \in E.$$

$V_n(x)$ is the *maximal expected total reward at time n over the remaining stages n to N if we start in state $x \in E$ at time n.* The functions $V_{n\pi}$ and V_n are well-defined since

$$V_{n\pi}(x) \le V_n(x) \le \delta_n^N(x) < \infty, \quad x \in E.$$

Note that $V_{N\pi}(x) = V_N(x) = g_N(x)$ and that $V_{n\pi}$ depends only on $(f_n, \ldots, f_{N-1})$. Moreover, it is in general not true that V_n is measurable. This causes theoretical inconveniences. Some further assumptions are needed to imply this (see Section 2.4).

A policy $\pi \in F_0 \times \ldots \times F_{N-1}$ is called *optimal* for the N-stage Markov Decision Model if $V_{0\pi}(x) = V_0(x)$ for all $x \in E$.

Until now we have considered *Markov policies*. One could ask why the decision rules are only functions of the current state and do not depend on the complete history? Let us now introduce the *sets of histories* which are denoted by

$$H_0 := E,$$
$$H_n := H_{n-1} \times A \times E.$$

An element $h_n = (x_0, a_0, x_1, \ldots, x_n) \in H_n$ is called *history up to time n*.

Definition 2.2.2. a) A measurable mapping $f_n : H_n \to A$ with the property $f_n(h_n) \in D_n(x_n)$ for all $h_n \in H_n$ is called a *history-dependent decision rule* at stage n.

b) A sequence $\pi = (f_0, f_1, \ldots, f_{N-1})$ where f_n is a history-dependent decision rule at stage n, is called a *history-dependent N-stage policy*. We denote by Π_N the set of all history-dependent N-stage policies.

Let $\pi \in \Pi_N$ be a history-dependent policy. Then $V_{n\pi}(h_n)$ is defined as the conditional expectation of the total reward in $[n, N]$, given the history $h_n \in H_n$. The following theorem states that history-dependent policies do

not improve the maximal expected rewards. For a proof see Hinderer (1970), Theorem 18.4.

Theorem 2.2.3. *For $n = 0, \ldots, N$ it holds:*

$$V_n(x_n) = \sup_{\pi \in \Pi_N} V_{n\pi}(h_n), \quad h_n = (x_0, a_0, x_1, \ldots, x_n).$$

Though we are in general satisfied with the value function $V_0(x)$, it turns out that on the way computing $V_0(x)$ we also have to determine the value function $V_n(x)$. This is a standard feature of many multistage optimization techniques and explained in the next section.

2.3 The Bellman Equation

For a fixed policy $\pi \in F_0 \times \ldots \times F_{N-1}$ we can compute the expected discounted rewards recursively by the so-called *reward iteration*. But first we introduce some important operators which simplify the notation. In what follows let us denote by

$$\mathbb{M}(E) := \{v : E \to [-\infty, \infty) \mid v \text{ is measurable}\}.$$

Due to our assumptions we have $V_{n\pi} \in \mathbb{M}(E)$ for all π and n.

Definition 2.3.1. We define the following operators for $n = 0, 1, \ldots, N-1$.

a) For $v \in \mathbb{M}(E)$ define

$$(L_n v)(x, a) := r_n(x, a) + \int v(x') Q_n(dx'|x, a), \quad (x, a) \in D_n$$

whenever the integral exists.

b) For $v \in \mathbb{M}(E)$ and $f \in F_n$ define

$$(\mathcal{T}_{nf} v)(x) := (L_n v)(x, f(x)), \quad x \in E.$$

c) For $v \in \mathbb{M}(E)$ define

$$(\mathcal{T}_n v)(x) := \sup_{a \in D_n(x)} (L_n v)(x, a), \quad x \in E.$$

$\mathcal{T}_n$ is called the *maximal reward operator at time n*.

Remark 2.3.2. a) We have the following relation between the operators

$$\mathcal{T}_n v = \sup_{f \in F_n} \mathcal{T}_{nf} v.$$

b) It holds that $\mathcal{T}_{nf} v \in \mathbb{M}(E)$ for all $v \in \mathbb{M}(E)$, but $\mathcal{T}_n v$ does not belong to $\mathbb{M}(E)$ in general.
c) If a Markov Decision Model with disturbances (Z_n) is given as in Theorem 2.1.3, then $L_n v$ can be written as

$$(L_n v)(x,a) = r_n(x,a) + \mathbb{E}\left[v\big(T_n(x,a,Z_{n+1})\big)\right].$$

This representation is often more convenient. ◊

Notation: In what follows we will skip the brackets $(\mathcal{T}_n v)(x)$ around the operators and simply write $\mathcal{T}_n v(x)$ in order to ease notation. When we have a sequence of operators like $\mathcal{T}_n \mathcal{T}_{n+1} v$ then the order of application is given by $(\mathcal{T}_n(\mathcal{T}_{n+1} v))$, i.e. the inner operator is applied first. The same convention applies to the other operators.

The operators have the following important properties.

Lemma 2.3.3. *All three operators are monotone, i.e. for $v, w \in \mathbb{M}(E)$ with $v(x) \leq w(x)$ for all $x \in E$ it holds:*

a) $L_n v(x,a) \leq L_n w(x,a)$ for all $(x,a) \in D_n$,
b) $\mathcal{T}_{nf} v(x) \leq \mathcal{T}_{nf} w(x)$ for all $x \in E, f \in F_n$,
c) $\mathcal{T}_n v(x) \leq \mathcal{T}_n w(x)$ for all $x \in E$.

Proof. Let $v(x) \leq w(x)$ for all $x \in E$. Then

$$\int v(x') Q_n(dx'|x,a) \leq \int w(x') Q_n(dx'|x,a).$$

Thus, $L_n v(x,a) \leq L_n w(x,a)$ which implies the first and second statement. Taking the supremum over all $a \in D_n(x)$ implies the third statement. □

The operators $\mathcal{T}_{nf}$ can now be used to compute the value of a policy recursively.

Theorem 2.3.4 (Reward Iteration). *Let $\pi = (f_0, \ldots, f_{N-1})$ be an N-stage policy. For $n = 0, 1, \ldots, N-1$ it holds:*

a) $V_{N\pi} = g_N$ and $V_{n\pi} = T_{nf_n} V_{n+1,\pi}$,
b) $V_{n\pi} = T_{nf_n} \ldots T_{N-1 f_{N-1}} g_N$.

Proof. a) For $x \in E$ we have

$$\begin{aligned}
V_{n\pi}(x) &= \mathbb{E}^{\pi}_{nx}\left[\sum_{k=n}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\right] \\
&= \mathbb{E}^{\pi}_{nx}[r_n\big(x, f_n(x)\big)] + \mathbb{E}^{\pi}_{nx}\left[\sum_{k=n+1}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\right] \\
&= r_n\big(x, f_n(x)\big) \\
&\quad + \mathbb{E}^{\pi}_{nx}\left[\mathbb{E}^{\pi}_{nx}\left[\sum_{k=n+1}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N) \;\middle|\; X_{n+1}\right]\right] \\
&= r_n\big(x, f_n(x)\big) \\
&\quad + \int \mathbb{E}^{\pi}_{n+1,x'}\left[\sum_{k=n+1}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\right] Q_n\big(dx'|x, f_n(x)\big) \\
&= r_n\big(x, f_n(x)\big) + \int V_{n+1,\pi}(x') Q_n\big(dx'|x, f_n(x)\big)
\end{aligned}$$

where we have used the properties of $\mathbb{P}^{\pi}_{xn}$ in the fourth equality.
b) Follows from a) by induction. □

Example 2.3.5 (Consumption Problem; continued). We revisit again Example 2.1.4. First note that for $f_n \in F_n$ the $\mathcal{T}_{nf_n}$ operator in this example reads

$$\mathcal{T}_{nf_n} v(x) = U_n\big(f_n(x)\big) + \mathbb{E}\, v\big((x - f_n(x))Z_{n+1}\big).$$

Now let us assume that $U_n(x) := \log x$ for all n and $g_N(x) := \log x$. Moreover, we assume that the return distribution is independent of n and has finite expectation $\mathbb{E}\, Z$. Then (A_N) is satisfied as we have shown in Example 2.2.1. If we choose the N-stage policy $\pi = (f_0, \ldots, f_{N-1})$ with $f_n(x) = cx$ and $c \in [0, 1]$, i.e. we always consume a constant fraction of the wealth, then the Reward Iteration in Theorem 2.3.4 implies by induction on N that

$$V_{0\pi}(x) = (N+1)\log x + N \log c + \frac{(N+1)N}{2}\Big(\log(1-c) + \mathbb{E}\log Z\Big).$$

Hence $\pi^* = (f_0^*, \ldots, f_{N-1}^*)$ with $f_n^*(x) = c^* x$ and $c^* = \frac{2}{N+3}$ maximizes the expected log-utility (among all linear consumption policies). ♦

The next definition will be crucial for the solution of Markov Decision Problems.

Definition 2.3.6. Let $v \in \mathbb{M}(E)$. A decision rule $f \in F_n$ is called a *maximizer of* v at time n if $\mathcal{T}_{nf} v = \mathcal{T}_n v$, i.e. for all $x \in E$, $f(x)$ is a maximum point of the mapping $a \mapsto L_n v(x, a)$, $a \in D_n(x)$.

Theorem 2.3.8 below gives the key solution method for Markov Decision Problems. They can be solved by successive application of the $\mathcal{T}_n$-operators. As mentioned earlier it is in general not true that $\mathcal{T}_n v \in \mathbb{M}(E)$ for $v \in \mathbb{M}(E)$. However, it can be shown that V_n is analytically measurable and the sequence (V_n) satisfies the so-called *Bellman equation*

$$\begin{aligned} V_N &= g_N, \\ V_n &= \mathcal{T}_n V_{n+1}, \quad n = 0, 1, \ldots, N-1, \end{aligned}$$

see e.g. Bertsekas and Shreve (1978). In the next theorem we show that whenever a solution of the Bellman equation exists together with a sequence of maximizers, then this yields the solution of our optimization problem.

Theorem 2.3.7 (Verification Theorem). *Let $(v_n) \subset \mathbb{M}(E)$ be a solution of the Bellman equation. Then it holds:*

a) $v_n \geq V_n$ for $n = 0, 1, \ldots, N$.
b) If f_n^* is a maximizer of v_{n+1} for $n = 0, 1, \ldots, N-1$, then $v_n = V_n$ and the policy $\pi^* = (f_0^*, f_1^*, \ldots, f_{N-1}^*)$ is optimal for the N-stage Markov Decision Problem.

Proof. a) For $n = N$ we have $v_N = g_N = V_N$. Suppose $v_{n+1} \geq V_{n+1}$, then for all $\pi = (f_0, \ldots, f_{N-1})$

$$v_n = \mathcal{T}_n v_{n+1} \geq \mathcal{T}_n V_{n+1} \geq \mathcal{T}_{n,f_n} V_{n+1,\pi} = V_{n\pi}.$$

Taking the supremum over all policies π yields $v_n \geq V_n$.

b) We show recursively that $v_n = V_n = V_{n\pi^*}$. For $n = N$ this is obvious. Suppose the statement is true for $n+1$, then

$$V_n \leq v_n = \mathcal{T}_{n f_n^*} v_{n+1} = \mathcal{T}_{n f_n^*} V_{n+1} = V_{n\pi^*} \leq V_n$$

and the theorem holds. □

The Verification Theorem is similar to statements which are usually delivered for stochastic control problems in continuous time. It is sufficient for applications where a solution of the Bellman equation is obvious and the existence of maximizers easy (e.g. if state and action spaces are finite). In general the existence of an optimal policy is not guaranteed. We have to make further assumptions about the structure of the problem to ensure this. In what follows we will first make a structure assumption to state our main theorem. Important cases where this assumption is satisfied are then discussed in Section 2.4. Also note that the value of an optimization problem is always unique whereas an optimal policy may not be unique.

Structure Assumption ($\mathbf{SA}_N$): *There exist sets $\mathbb{M}_n \subset \mathbb{M}(E)$ and $\Delta_n \subset F_n$ such that for all $n = 0, 1, \ldots, N-1$:*

(i) $g_N \in \mathbb{M}_N$.
(ii) *If $v \in \mathbb{M}_{n+1}$ then $\mathcal{T}_n v$ is well-defined and $\mathcal{T}_n v \in \mathbb{M}_n$.*
(iii) *For all $v \in \mathbb{M}_{n+1}$ there exists a maximizer f_n of v with $f_n \in \Delta_n$.*

Often $\mathbb{M}_n$ is independent of n and it is possible to choose $\Delta_n = F_n \cap \Delta$ for a set $\Delta \subset \{f : E \to A \text{ measurable}\}$, i.e all value functions and all maximizers have the same structural properties.

The next theorem is the main result of this section. It shows how Markov Decision Problems can be solved recursively by solving N (one-stage) optimization problems.

Theorem 2.3.8 (Structure Theorem). *Let (SA_N) be satisfied. Then it holds:*

a) $V_n \in \mathbb{M}_n$ *and the sequence* (V_n) *satisfies the* Bellman equation, *i.e. for* $n = 0, 1, \ldots, N-1$

$$V_N(x) = g_N(x),$$
$$V_n(x) = \sup_{a \in D_n(x)} \left\{ r_n(x,a) + \int V_{n+1}(x') Q_n(dx'|x,a) \right\}, \quad x \in E.$$

b) $V_n = \mathcal{T}_n \mathcal{T}_{n+1} \ldots \mathcal{T}_{N-1} g_N$.
c) For $n = 0, 1, \ldots, N-1$ there exist maximizers f_n of V_{n+1} with $f_n \in \Delta_n$, and every sequence of maximizers f_n^* of V_{n+1} defines an optimal policy $(f_0^*, f_1^*, \ldots, f_{N-1}^*)$ for the N-stage Markov Decision Problem.

Proof. Since b) follows directly from a) it suffices to prove a) and c). We show by induction on $n = N-1, \ldots, 0$ that $V_n \in \mathbb{M}_n$ and that

$$V_{n\pi^*} = \mathcal{T}_n V_{n+1} = V_n$$

where $\pi^* = (f_0^*, \ldots, f_{N-1}^*)$ is the policy generated by the maximizers of $V_1, \ldots, V_N$ and $f_n^* \in \Delta_n$. We know $V_N = g_N \in \mathbb{M}_N$ by (SA_N) (i). Now suppose that the statement is true for $N-1, \ldots, n+1$. Since $V_k \in \mathbb{M}_k$ for $k = N, \ldots, n+1$, the maximizers $f_n^*, \ldots, f_{N-1}^*$ exist and we obtain with the reward iteration and the induction hypothesis (note that $f_0^*, \ldots, f_{n-1}^*$ are not relevant for the following equation)

$$V_{n\pi^*} = \mathcal{T}_{n f_n^*} V_{n+1,\pi^*} = \mathcal{T}_{n f_n^*} V_{n+1} = \mathcal{T}_n V_{n+1}.$$

Hence $V_n \geq \mathcal{T}_n V_{n+1}$. On the other hand for an arbitrary policy π

$$V_{n\pi} = \mathcal{T}_{nf_n} V_{n+1,\pi} \leq \mathcal{T}_{nf_n} V_{n+1} \leq \mathcal{T}_n V_{n+1}$$

where we have used the order preserving property of $\mathcal{T}_{nf_n}$. Taking the supremum over all policies yields $V_n \leq \mathcal{T}_n V_{n+1}$. Altogether it follows that

$$V_{n\pi^*} = \mathcal{T}_n V_{n+1} = V_n$$

and in view of (SA_N), $V_n \in I\!M_n$. □

From this result we conclude directly the following corollary.

Corollary 2.3.9. *Let (SA_N) be satisfied. If $n \leq m \leq N$ then it holds:*

$$V_n(x) = \sup_\pi \mathbb{E}^\pi_{nx} \Big[\sum_{k=n}^{m-1} r_k\big(X_k, f_k(X_k)\big) + V_m(X_m) \Big], \quad x \in E.$$

Theorem 2.3.8 implies the following recursive algorithm to solve Markov Decision Problems:

Backward Induction Algorithm.

1. Set $n := N$ and for $x \in E$:
$$V_N(x) := g_N(x).$$
2. Set $n := n - 1$ and compute for all $x \in E$
$$V_n(x) = \sup_{a \in D_n(x)} \Big\{ r_n(x,a) + \int V_{n+1}(x') Q_n(dx'|x,a) \Big\}.$$
Compute a maximizer f_n^* of V_{n+1}.
3. If $n = 0$, then the value function V_0 is computed and the optimal policy π^* is given by $\pi^* = (f_0^*, \ldots, f_{N-1}^*)$. Otherwise, go to step 2.

Theorem 2.3.8 tells us that the maximizers yield an optimal strategy. However the reverse statement is not true: optimal strategies do not necessarily contain only maximizers. This is shown by the following example.

Example 2.3.10. Let $N = 2$ be the planning horizon and state and action spaces be given by $S = \{0,1\} = A = D_n(x)$ for all $x \in E$. The transition probabilities are given by $Q_n(\{x'\}|x,a) = 1$ if $a = x'$ and zero otherwise (see Figure 2.1). The reward functions are given by $r_n(x,a) = a$ for $(x,a) \in D_n$ and $g_2(x) = x$. The optimal policy is easily computed to be $\pi^* = (f_0^*, f_1^*)$ with $f_0^*(x) = 1$ and $f_1^*(x) = 1$ for all $x \in E$. However, it is easy to see that $\pi = (f_0, f_1)$ with $f_0(x) \equiv 1$, and $f_1(0) = 0$, $f_1(1) = 1$ yields the same expected total reward and is thus optimal, too. Obviously the reason is that under an optimal policy state 0 will not be visited after time 1. ◊

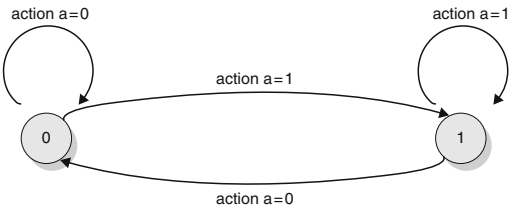

Fig. 2.1 Transition probabilities of Example 2.3.10.

Example 2.3.11 (Consumption Problem; continued). Let us now solve the consumption problem of Example 2.3.5. First suppose that the Structure Assumption (SA_N) is satisfied and we can apply Theorem 2.3.8. Thus, we obtain $V_N(x) = \log x$ and

$$\begin{aligned} V_{N-1}(x) = \mathcal{T}_{N-1} V_N(x) &= \sup_{a \in [0,x]} \Big\{ \log a + \log(x-a) + \mathbb{E} \log Z \Big\} \\ &= 2 \log x + 2 \log 0.5 + \mathbb{E} \log Z \end{aligned}$$

where the maximizer is given by $f^*_{N-1}(x) = 0.5x$. Now by induction we obtain

$$V_n(x) = (N - n + 1) \log x + d_n, \quad 0 \le n \le N$$

where $d_N = 0$ and

$$d_n = d_{n+1} + (N-n)\, \mathbb{E} \log Z - \log\big(N - n + 1\big) + (N-n) \log\Big(\frac{N-n}{N-n+1}\Big),$$

and we obtain the maximizer $f^*_n(x) = \frac{1}{N-n+1} x$. Thus, the optimal fraction which is consumed is independent of the wealth and increases over time. Finally it remains to show that (SA_N) is satisfied. But this can now be easily verified by choosing

$$\mathbb{M}_n := \{v \in \mathbb{M}(E) \mid v(x) = b \log x + d \text{ for constants } b, d \in \mathbb{R}\}$$

$$\Delta_n := \{f \in F_n \mid f(x) = cx \text{ for } c \in \mathbb{R}\}.$$

Indeed, the necessary calculations are pretty much the same as we have performed before. ♦

In order to obtain a good guess about how $\mathbb{M}_n$ and Δ_n look like it is reasonable to compute the first steps of the Backward Induction Algorithm and investigate the structure of the value functions.

The solution method in Theorem 2.3.8 relies on a very simple but general observation which is called the *Principle of Dynamic Programming*. Informally it says that whenever we have an optimal policy π^* over a certain horizon N and consider the process now only on a subinterval of $[0, N]$, then the

corresponding policy which is obtained by restricting π^* to this subinterval is again optimal. This can be formalized as follows.

Theorem 2.3.12 (Principle of Dynamic Programming). *Let (SA_N) be satisfied. Then it holds for $n \leq m \leq N$:*

$$V_{n\pi^*}(x) = V_n(x) \quad \Rightarrow \quad V_{m\pi^*} = V_m \ \ \mathbb{P}_{nx}^{\pi^*} - a.s.,$$

i.e. if $(f_n^*, \ldots, f_{N-1}^*)$ is optimal for the time period $[n, N]$ then $(f_m^*, \ldots, f_{N-1}^*)$ is optimal for $[m, N]$.

Proof. It follows from the Reward Iteration (Theorem 2.3.4) and the definition of V_m that

$$\begin{aligned} V_n(x) = V_{n\pi^*}(x) &= \mathcal{T}_{nf_n^*} \ldots \mathcal{T}_{m-1,f_{m-1}^*} V_{m\pi^*}(x) \\ &= \mathbb{E}_{nx}^{\pi^*}\Big[\sum_{k=n}^{m-1} r_k\big(X_k, f_k^*(X_k)\big) + V_{m\pi^*}(X_m)\Big] \\ &\leq \mathbb{E}_{nx}^{\pi^*}\Big[\sum_{k=n}^{m-1} r_k\big(X_k, f_k^*(X_k)\big) + V_m(X_m)\Big] \leq V_n(x) \end{aligned}$$

where we have used Corollary 2.3.9 for the last inequality. This implies that we have indeed equality and that $\mathbb{E}_{nx}^{\pi^*}\big[V_m(X_m) - V_{m\pi^*}(X_m)\big] = 0$ which means that $V_{m\pi^*} = V_m \ \mathbb{P}_{nx}^{\pi^*} - a.s.$ □

Sometimes Markov Decision Models are such that the state space contains an absorbing subset ('cemetery' subset) which will never be left once it is reached and where we obtain no reward. Let us call this set $G \subset E$. Obviously such a set is not very interesting and can in principle be neglected when formulating the Markov Decision Model. However this leads at least for some stages to a substochastic transition law. This is explained in the next example.

Example 2.3.13 (Absorbing Markov Decision Model). Suppose that a Markov Decision Model is given, where E is the state space, $\emptyset \neq G \subset E$ and A is the action space. We suppose that $Q_n(G|x, a) = 1$ for $x \in G$, i.e. G will never be left, once it is entered. Moreover $r_n(x, a) = 0$ and $g_N(x) = 0$ for all $x \in G$ and $a \in D_n(x)$. Then it is sufficient to consider the following *substochastic* Markov Decision Model $(\tilde{E}, \tilde{A}, \tilde{D}_n, \tilde{Q}_n, \tilde{r}_n, \tilde{g}_N)$ with

- $\tilde{E} := E \setminus G$,
- $\tilde{A} := A$,
- $\tilde{D}_n(x) := D_n(x)$ for $x \in \tilde{E}$ and $\tilde{D}_n := \{(x, a) \mid x \in \tilde{E}, a \in \tilde{D}_n(x)\}$,
- $\tilde{Q}_n(B|x, a) := Q_n(B|x, a)$ for $(x, a) \in \tilde{D}_n$ and $B \in \mathfrak{E}, B \subset \tilde{E}$,
- $\tilde{r}_n(x, a) := r_n(x, a)$ for $(x, a) \in \tilde{D}$,

- $\tilde{g}_N(x) := g_N(x)$ for $x \in \tilde{E}$.

Typically $\tilde{Q}_n(\tilde{E}|x,a) < 1$ for some $x \in E$, since G is deleted from the state space and may be entered with positive probability. Moreover, by definition we have $V_N(x) = 0$ for $x \in G$. It is now easy to see by induction using Theorem 2.3.8 that $V_n(x) = 0$ for all $x \in G$. This implies that it holds for all $x \in \tilde{E}$:

$$\begin{aligned} V_n(x) &= \sup_{a \in D_n(x)} \Big\{ r_n(x,a) + \int V_{n+1}(x') Q_n(dx'|x,a) \Big\} \\ &= \sup_{a \in \tilde{D}_n(x)} \Big\{ \tilde{r}_n(x,a) + \int_{\tilde{E}} V_{n+1}(x') \tilde{Q}_n(dx'|x,a) \Big\} \end{aligned}$$

where we have used that V_{n+1} vanishes on G. As a consequence, we can restrict ourselves to consider the Bellman equation on $\tilde{E}$. For an application see Section 2.6.1. ⧫

Remark 2.3.14. a) The statements in Theorem 2.3.8 remain valid when we replace the Assumption (A_N) by the following weaker assumption: Assume that for $n = 0, 1, \ldots, N$

$$\sum_{k=n}^{N-1} r_k^+(X_k, f_k(X_k)) + g_N^+(X_N)$$

is P_{nx}^π-integrable for all π and $x \in E$. However, in this case we might have $V_n(x) = +\infty$ for some $x \in E$.

b) It is well known that the Bellman equation holds under much weaker assumptions than in Theorem 2.3.8 (see Hinderer (1970), Theorem 14.4). In particular, if the reward functions r_n and g_N are non-negative (and without any further assumptions), the value functions V_n satisfy the Bellman equation. ◊

Remark 2.3.15 (Minimizing Cost). Instead of one-stage rewards r_n and a terminal reward g_N, sometimes problems are given where we have a one-stage cost c_n and a terminal cost h_N. In this case we want to minimize

$$\mathbb{E}_{nx}^\pi \left[\sum_{k=n}^{N-1} c_k\big(X_k, f_k(X_k)\big) + h_N(X_N) \right], \quad x \in E$$

for $\pi = (f_0, \ldots, f_{N-1})$. But this problem can be transformed into a reward maximization problem by setting $r_n(x,a) := -c_n(x,a)$, $g_N(x) := -h_N(x)$. Thus, all the statements so far remain valid. We will use the same notation $V_{n\pi}$ and V_n for the cost functions under policy π and the minimal cost function. Moreover, the minimal cost operator $\mathcal{T}_n$ has the form

$$(\mathcal{T}_n v)(x) = \inf_{a \in D_n(x)} \Big\{ c_n(x,a) + \int v(x') Q_n(dx'|x,a) \Big\}.$$

In this case V_n is also called *cost-to-go function.* $\Diamond$

2.4 Structured Markov Decision Models

In this section we give sufficient conditions under which assumptions (A_N) and (SA_N) in the preceding section are satisfied and thus imply the validity of the Bellman equation and the existence of optimal policies. For (SA_N) we will identify conditions which imply that special sets $\mathbb{M}_n$ and Δ_n can be chosen. Of course it is interesting to choose the sets $\mathbb{M}_n$ and Δ_n as small as possible. The smaller the sets, the more information we have about the value functions and the optimal policy. On the other hand, small sets imply that we have to prove a lot of properties of $\mathcal{T}_n v$ if $v \in \mathbb{M}_{n+1}$.
Let us first consider the Integrability Assumption (A_N). It is fulfilled when the Markov Decision Model has a so-called upper bounding function.

Definition 2.4.1. A measurable function $b : E \to \mathbb{R}_+$ is called an *upper bounding function* for the Markov Decision Model if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$ such that for all $n = 0, 1, \ldots, N-1$:

(i) $r_n^+(x,a) \le c_r b(x)$ for all $(x,a) \in D_n$,
(ii) $g_N^+(x) \le c_g b(x)$ for all $x \in E$,
(iii) $\int b(x') Q_n(dx'|x,a) \le \alpha_b b(x)$ for all $(x,a) \in D_n$.

If r_n and g_N are bounded from above, then obviously $b \equiv 1$ is an upper bounding function.
Let b be an upper bounding function for the Markov Decision Model. For $v \in \mathbb{M}(E)$ we denote the *weighted supremum norm* by

$$\|v\|_b := \sup_{x \in E} \frac{|v(x)|}{b(x)}$$

(with the convention $\frac{0}{0} := 0$) and define the set

$$\mathbb{B}_b := \{v \in \mathbb{M}(E) \mid \|v\|_b < \infty\}.$$

Equivalently $\mathbb{B}_b$ can be written as

$$\mathbb{B}_b := \{v \in \mathbb{M}(E) \mid |v(x)| \le c b(x) \text{ for all } x \in E \text{ and for some } c \in \mathbb{R}_+\}.$$

The concept of upper bounding functions is of particular interest for Markov Decision Models with infinite time horizon (cp. Chapter 7). The next result is fundamental for many applications.

Proposition 2.4.2. *If the Markov Decision Model has an upper bounding function b, then $\delta_n^N \in \mathbb{B}_b$ and the Integrability Assumption (A_N) is satisfied.*

Proof. Since $\delta_n^N \geq 0$ we have to show that $\delta_n^N(x) \leq cb(x)$ for some $c \in \mathbb{R}_+$. From the properties of an upper bounding function it follows that

$$\begin{aligned}\mathbb{E}_x^\pi \left[r_k^+\big(X_k, f_k(X_k)\big) \mid X_{k-1} \right] &= \int r_k^+\big(x', f_k(x')\big) Q_k\big(dx' \mid X_{k-1}, f_{k-1}(X_{k-1})\big) \\ &\leq c_r \alpha_b b(X_{k-1})\end{aligned}$$

and by iteration we obtain

$$\mathbb{E}_x^\pi \left[r_k^+\big(X_k, f_k(X_k)\big) \right] \leq c_r \alpha_b^k b(x), \quad x \in E.$$

Analogously we get

$$\mathbb{E}_x^\pi \left[g_N^+(X_N) \right] \leq c_g \alpha_b^N b(x), \quad x \in E$$

and the result follows. □

For the rest of Section 2.4 we assume that E and A are Borel spaces (see Remark 2.1.2 a)). Also D_n is assumed to be a Borel subset of $E \times A$. Further we suppose that our Markov Decision Model has an upper bounding function b and we introduce the set

$$\mathbb{B}_b^+ := \{v \in \mathbb{M}(E) \mid \|v^+\|_b < \infty\}.$$

Equivalently $\mathbb{B}_b^+$ can be written as

$$\mathbb{B}_b^+ := \{v \in \mathbb{M}(E) \mid v^+(x) \leq cb(x) \text{ for all } x \in E \text{ and for some } c \in \mathbb{R}_+\}.$$

2.4.1 Semicontinuous Markov Decision Models

In so-called *semicontinuous* Markov Decision Models the Structure Assumption (SA_N) is fulfilled with $\mathbb{M}_n$ being a subset of upper semicontinuous functions. This is a consequence of the following results (for the definition of upper semicontinuity and properties of set-valued functions the reader is referred to the Appendix A).

Proposition 2.4.3. *Let $v \in \mathbb{B}_b^+$ be upper semicontinuous. Suppose the following assumptions are satisfied:*

(i) *$D_n(x)$ is compact for all $x \in E$ and $x \mapsto D_n(x)$ is upper semicontinuous,*

(ii) $(x,a) \mapsto L_n v(x,a)$ *is upper semicontinuous on* D_n.

Then $\mathcal{T}_n v$ *is upper semicontinuous and there exists a maximizer* f_n *of* v.

Remark 2.4.4. Condition (i) in Proposition 2.4.3 can be replaced by the following condition: For all $x \in E$ the level sets $\{a \in D_n(x) \mid L_n v(x,a) \geq c\}$ are compact for all $c \in \mathbb{R}$ and the set-valued mapping

$$x \mapsto \{a \in D_n(x) \mid L_n v(x,a) = \mathcal{T}_n v(x)\}$$

is upper semicontinuous. ◊

Proof. To ease notation let us define

$$w(x,a) := L_n v(x,a), \quad w^*(x) := \mathcal{T}_n v(x)$$

and $D(x) := D_n(x)$. For $x_0 \in E$ select a sequence $(x_n) \subset E$ converging to x_0 such that the limit of $w^*(x_n)$ exists. We have to show that

$$\lim_{n\to\infty} w^*(x_n) \leq w^*(x_0).$$

Since $a \mapsto w(x,a)$ is upper semicontinuous on the compact set $D(x)$, it attains its supremum on $D(x)$ (see Theorem A.1.2). Let $a_n \in D(x_n)$ be a maximum point of $a \mapsto w(x_n,a)$ on $D(x_n)$. By the upper semicontinuity of $x \mapsto D(x)$ there is a subsequence (a_{n_k}) of (a_n) converging to some $a_0 \in D(x_0)$. The upper semicontinuity of w implies

$$\lim_{n\to\infty} w^*(x_n) = \lim_{k\to\infty} w^*(x_{n_k}) = \lim_{k\to\infty} w(x_{n_k}, a_{n_k}) \leq w(x_0,a_0) \leq w^*(x_0),$$

i.e. w^* is upper semicontinuous.
Since w and w^* are measurable, it follows easily that

$$D^* := \{(x,a) \in D \mid w(x,a) = w^*(x)\}$$

is a Borel subset of $E \times A$ and each $D^*(x)$ is compact since

$$D^*(x) := \{a \in D(x) \mid w(x,a) \geq w^*(x)\}.$$

Then, applying the selection theorem of Kuratowski and Ryll-Nardzewski (see Appendix, Theorem A.2.3), we obtain a Borel measurable selector f for D^*. This is the desired maximizer of v. □

Remark 2.4.5. If $A \subset \mathbb{R}$ then there exists a smallest and a largest maximizer of $v \in \mathbb{B}_b^+$. Note that the set

$$D_n^*(x) := \{a \in D_n(x) \mid L_n v(x,a) = \mathcal{T}_n v(x)\}$$

is compact for $x \in E$. Then $\max D_n^*(x)$ and $\min D_n^*(x)$ are maximizers of v by the Selection Theorem A.2.3. ◊

A set of sufficient conditions on the data of a Markov Decision Model in order to assure that (SA$_N$) is satisfied with $\mathbb{M}_n$ being the set of upper semicontinuous functions $v \in \mathbb{B}_b^+$ and with $\Delta_n := F_n$ is given below. Consequently under these assumptions Theorem 2.3.8 is valid. The proof follows immediately from Proposition 2.4.3.

Theorem 2.4.6. *Suppose the Markov Decision Model has an upper bounding function b and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *$D_n(x)$ is compact for all $x \in E$ and $x \mapsto D_n(x)$ is upper semicontinuous,*
(ii) *$(x,a) \mapsto \int v(x')Q_n(dx'|x,a)$ is upper semicontinuous for all upper semicontinuous $v \in \mathbb{B}_b^+$,*
(iii) *$(x,a) \mapsto r_n(x,a)$ is upper semicontinuous,*
(iv) *$x \mapsto g_N(x)$ is upper semicontinuous.*

Then the sets $\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is upper semicontinuous}\}$ and $\Delta_n := F_n$ satisfy the Structure Assumption (SA$_N$). In particular, $V_n \in \mathbb{M}_n$ and there exists a maximizer $f_n^* \in F_n$ of V_{n+1}. The policy $(f_0^*, \ldots, f_{N-1}^*)$ is optimal.

Instead of checking condition (ii) of Theorem 2.4.6 directly, we can alternatively use the following Lemma:

Lemma 2.4.7. *Let b be a continuous upper bounding function. Then the following statements are equivalent:*

(i) *$(x,a) \mapsto \int v(x')Q(dx'|x,a)$ is upper semicontinuous for all upper semicontinuous $v \in \mathbb{B}_b^+$.*
(ii) *$(x,a) \mapsto \int b(x')Q(dx'|x,a)$ is continuous, and $(x,a) \mapsto \int v(x')Q(dx'|x,a)$ is continuous and bounded for all continuous and bounded v on E.*

A stochastic kernel Q with the last property is called *weakly continuous*.

Proof. The proof that (ii) implies (i) is as follows: Let $v \in \mathbb{B}_b^+$ be upper semicontinuous. Then we have $v - cb \leq 0$ for some $c \in \mathbb{R}_+$ and $x \mapsto v(x) - cb(x)$ is upper semicontinuous. According to Lemma A.1.3 this implies the existence of a sequence $(\tilde{v}_k)$ of continuous and bounded functions such that $\tilde{v}_k \downarrow v - cb$. Due to our assumption the function $(x,a) \mapsto \int \tilde{v}_k(x')Q(dx'|x,a)$ is now continuous and bounded. Moreover, monotone convergence implies that

$$\int \tilde{v}_k(x')Q(dx'|x,a) \downarrow \int (v - cb)(x')Q(dx'|x,a) \quad \text{for } k \to \infty.$$

Thus, we can again conclude with Lemma A.1.3 that the limit is upper semicontinuous. However in view of our assumption this implies $(x,a) \mapsto \int v(x')Q(dx'|x,a)$ is upper semicontinuous.
Next we prove that (i) implies (ii): Since b and $-b$ are in $\mathbb{B}_b^+$, we get

$$(x,a) \mapsto \int b(x')Q(dx'|x,a)$$

is continuous. If v is bounded and continuous, then the functions $v - \|v\|$ and $-v - \|v\|$ belong to $\mathbb{B}_b^+$ and are upper semicontinuous. Hence the function $(x,a) \mapsto \int v(x')Q(dx'|x,a)$ is continuous. □

2.4.2 Continuous Markov Decision Models

Next we investigate when the Structure Assumption (SA_N) is satisfied with $\mathbb{M}_n$ being a subset of continuous functions.

Proposition 2.4.8. *Let $v \in \mathbb{B}_b^+$ be continuous. Suppose the following assumptions are satisfied:*

(i) *$D_n(x)$ is compact for all $x \in E$ and $x \mapsto D_n(x)$ is continuous,*
(ii) *$(x,a) \mapsto L_n v(x,a)$ is continuous on D_n.*

Then $\mathcal{T}_n v$ is continuous and there exists a maximizer $f_n \in F_n$ of v. If v has a unique maximizer $f_n \in F_n$ at time n, then f_n is continuous.

Proof. We use the same notation as in the proof of Proposition 2.4.3. In view of Proposition 2.4.3 it is sufficient to show that w^* is lower semicontinuous, i.e. that $w^*(x_0) \le \liminf_{n\to\infty} w^*(x_n)$ for each sequence $(x_n) \subset E$ which converges to $x_0 \in E$ and for which $\lim_{n\to\infty} w^*(x_n)$ exists. We know by assumption that $w^*(x_0) = w(x_0, a_0)$ for some $a_0 \in D(x_0)$. Since $x \mapsto D(x)$ is continuous, there exists a subsequence (x_{n_k}) of (x_n) and a sequence of points $a_{n_k} \in D(x_{n_k})$ converging to a_0. Hence we have $(x_{n_k}, a_{n_k}) \to (x_0, a_0)$. It follows from the continuity of w that

$$w^*(x_0) = w(x_0,a_0) = \lim_{k\to\infty} w(x_{n_k}, a_{n_k}) \le \lim_{k\to\infty} w^*(x_{n_k}) = \lim_{n\to\infty} w^*(x_n).$$

Since $x \mapsto D(x)$ is upper semicontinuous, D is closed and it follows that

$$D^* := \{(x,a) \in D \mid w(x,a) = w^*(x)\}$$

is a closed subset of D. Then we obtain that $x \mapsto D^*(x)$ is also upper semicontinuous. Thus, if v has a unique maximizer f_n, i.e. $D_n^*(x) = \{f_n(x)\}$ for all $x \in E$, then f_n must be continuous. □

Remark 2.4.9. If $A \subset \mathbb{R}$, then the smallest (largest) maximizer of v is lower semicontinuous (upper semicontinuous). This can be seen as follows: If $A \subset \mathbb{R}$, then $x \mapsto f(x) := \max D^*(x)$ is the largest maximizer. Choose $x_0, x_n \in E$ such that $x_n \to x_0$ and the limit of $(f(x_n))$ exists. Since $a_n := f(x_n) \in D^*(x_n)$ and $x \mapsto D^*(x)$ is upper semicontinuous, (a_n) has an accumulation point in $D^*(x_0)$ which must be $\lim_{n\to\infty} f(x_n)$. It follows that

$$\lim_{n\to\infty} f(x_n) \leq \max D^*(x_0) = f(x_0),$$

i.e. the largest maximizer is upper semicontinuous. In the same way it can be shown that the smallest maximizer is lower semicontinuous. ◊

It is now rather straightforward that the following conditions on the data of a Markov Decision Model imply by using Proposition 2.4.8 that (SA_N) is satisfied and that Theorem 2.3.8 is valid.

Theorem 2.4.10. *Suppose a Markov Decision Model with upper bounding function b is given and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *$D_n(x)$ is compact for all $x \in E$ and $x \mapsto D_n(x)$ is continuous,*
(ii) *$(x, a) \mapsto \int v(x')Q_n(dx'|x, a)$ is continuous for all continuous $v \in \mathbb{B}_b^+$,*
(iii) *$(x, a) \mapsto r_n(x, a)$ is continuous,*
(iv) *$x \mapsto g_N(x)$ is continuous.*

Then the sets $\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is continuous}\}$ and $\Delta_n := F_n$ satisfy the Structure Assumption (SA_N). If the maximizer of V_n is unique, then Δ_n can be chosen as the set of continuous functions.

2.4.3 Measurable Markov Decision Models

Sometimes the Structure Assumption (SA_N) can be fulfilled with $\mathbb{M}_n = \mathbb{B}_b^+$. For this case the following result is useful.

Proposition 2.4.11. *Let $v \in \mathbb{B}_b^+$ and suppose the following assumptions are satisfied:*

(i) *$D_n(x)$ is compact for all $x \in E$,*
(ii) *$a \mapsto L_n v(x, a)$ is upper semicontinuous on $D_n(x)$ for all $x \in E$.*

Then $\mathcal{T}_n v$ is measurable and there exists a maximizer $f_n \in F_n$ of v.

Proof. We use the same notation as in the proof of Proposition 2.4.3. Let $\alpha \in \mathbb{R}$. It is sufficient to prove that $\{x \in E \mid w^*(x) \geq \alpha\}$ is a Borel set. But

$$\begin{aligned}\{x \in E \mid w^*(x) \geq \alpha\} &= \{x \in E \mid w(x,a) \geq \alpha \text{ for some } a \in D(x)\} \\ &= \mathrm{proj}_E\{(x,a) \in D \mid w(x,a) \geq \alpha\}.\end{aligned}$$

This set is Borel by a result of Kunugui and Novikov (see Himmelberg et al. (1976)), since D is Borel with compact values (i.e. compact vertical sections) and $\{(x,a) \in D \mid w(x,a) \geq \alpha\}$ is a Borel subset of D with closed (and therefore compact) values. Actually, Kunugui and Novikov prove that the projection of a Borel subset of $E \times A$ with compact values is a Borel subset of E. The existence of a maximizer can be shown in the same way as in the proof of Proposition 2.4.3. □

Remark 2.4.12. a) If A is countable and $D_n(x)$ is finite for all $x \in E$, then both assumptions (i) and (ii) of Proposition 2.4.11 are fulfilled.
b) Condition (i) in Proposition 2.4.11 can be replaced by: For all $x \in E$ the level set $\{a \in D_n(x) \mid L_n v(x,a) \geq c\}$ is compact for all $c \in \mathbb{R}$. ◊

The following theorem follows directly from Proposition 2.4.11. In particular the main result (Theorem 2.3.8) holds under the assumptions of Theorem 2.4.13.

Theorem 2.4.13. *Suppose a Markov Decision Model with upper bounding function b is given and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *$D_n(x)$ is compact for all $x \in E$,*
(ii) *$a \mapsto \int v(x')Q_n(dx'|x,a)$ is upper semicontinuous for all $v \in \mathbb{B}_b^+$ and for all $x \in E$,*
(iii) *$a \mapsto r_n(x,a)$ is upper semicontinuous for all $x \in E$.*

Then the sets $\mathbb{M}_n := \mathbb{B}_b^+$ and $\Delta_n := F_n$ satisfy the Structure Assumption (SA_N).

In a more general framework one can choose $\mathbb{M}_n$ as the set of upper semianalytic functions (see Bertsekas and Shreve (1978)). But of course, one wants to choose $\mathbb{M}_n$ and Δ_n as small as possible.

2.4.4 Monotone and Convex Markov Decision Models

Structural properties (e.g. monotonicity, concavity and convexity) for the value functions and also for the maximizers are important for applications.

Results like these also simplify numerical solutions. In order to ease the exposition we assume now that $E \subset \mathbb{R}^d$ and $A \subset \mathbb{R}^m$ endowed with the usual preorder $\leq$ of componentwise comparison e.g. $x \leq y$ for $x, y \in \mathbb{R}^d$ if $x_k \leq y_k$ for $k = 1, \ldots, d$.

Theorem 2.4.14. *Suppose a Markov Decision Model with upper bounding function b is given and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *$D_n(\cdot)$ is increasing, i.e. $x \leq x'$ implies $D_n(x) \subset D_n(x')$,*
(ii) *the stochastic kernels $Q_n(\cdot|x,a)$ are stochastically monotone for all $a \in D_n(x)$, i.e. the mapping $x \mapsto \int v(x')Q_n(dx'|x,a)$ is increasing for all increasing $v \in \mathbb{B}_b^+$ and for all $a \in D_n(x)$,*
(iii) *$x \mapsto r_n(x,a)$ is increasing for all a,*
(iv) *g_N is increasing on E,*
(v) *for all increasing $v \in \mathbb{B}_b^+$ there exists a maximizer $f_n \in \Delta_n$ of v.*

Then the sets $\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is increasing}\}$ and Δ_n satisfy the Structure Assumption (SA_N).

Proof. Obviously condition (iv) shows that $g_N \in \mathbb{M}_N$. Let now $v \in \mathbb{M}_{n+1}$. Then conditions (ii) and (iii) imply that $x \mapsto L_n v(x,a)$ is increasing for all a. In view of (i) we obtain $\mathcal{T}_n v \in \mathbb{M}_n$. Condition (v) is equivalent to condition (iii) of (SA_N). Thus, the statement is shown. □

It is more complicated to identify situations in which the maximizers are increasing. For this property we need the following definition.

Definition 2.4.15. A set $D \subset E \times A$ is called *completely monotone* if for all points $(x,a'), (x',a) \in D$ with $x \leq x'$ and $a \leq a'$ it follows that $(x,a), (x',a') \in D$.

An important special case where D is completely monotone is given if $D(x)$ is independent of x. If $A = \mathbb{R}$ and $D(x) = [\underline{d}(x), \bar{d}(x)]$. Then D is completely monotone if and only if $\underline{d} : E \to \mathbb{R}$ and $\bar{d} : E \to \mathbb{R}$ are increasing.

For a definition and properties of *supermodular* functions see Appendix A.3.

Proposition 2.4.16. *Let $v \in \mathbb{B}_b^+$ and suppose the following assumptions are satisfied where $D_n^*(x) := \{a \in D_n(x) \mid L_n v(x,a) = \mathcal{T}_n v(x)\}$ for $x \in E$:*

(i) *D_n is completely monotone,*
(ii) *$L_n v$ is supermodular on D_n,*
(iii) *there exists a largest maximizer f_n^* of v i.e. for all $x \in E$ it holds: $f_n^*(x) \geq a$ for all $a \in D_n^*(x)$ which are comparable with $f_n^*(x)$.*

Then f_n^* is weakly increasing, i.e. $x \leq x'$ implies $f_n^*(x) \leq f_n^*(x')$, whenever $f_n^*(x)$ and $f_n^*(x')$ are comparable.

Proof. Suppose that f_n^* is not increasing, i.e. there exist $x, x' \in E$ with $x \leq x'$ and $f_n^*(x) =: a > a' := f_n^*(x')$. Due to our assumptions (i) and (ii) we know that $(x, a'), (x', a) \in D_n$ and

$$\big(L_n v(x, a') - L_n v(x, a)\big) + \big(L_n v(x', a) - L_n v(x', a')\big) \geq 0.$$

Since a is a maximum point of $b \mapsto L_n v(x, b)$ and a' is a maximum point of $b \mapsto L_n v(x', b)$ the expressions in brackets are non-positive. Thus, we must have $L_n v(x', a') = L_n v(x', a)$ which means in particular that a is also a maximum point of $b \mapsto L_n v(x', b)$. But this contradicts the definition of f_n^* as the largest maximizer of v, and the statement follows. □

Remark 2.4.17. a) A similar result holds for the smallest maximizer of v.
b) If we reverse the relation on the state or the action space we obtain conditions for weakly decreasing maximizers. ◊

If our Markov Decision Model fulfills all assumptions of Theorem 2.4.14 and Proposition 2.4.16, then

$$\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is increasing}\}$$
$$\Delta_n := \{f \in F_n \mid f \text{ is weakly increasing}\}$$

satisfy the Structure Assumption (SA_N).

Of particular interest are concave or convex value functions.

Proposition 2.4.18. *Let $v \in \mathbb{B}_b^+$ and suppose the following assumptions are satisfied:*

(i) *D_n is convex in $E \times A$,*
(ii) *$L_n v(x, a)$ is concave on D_n.*

Then $\mathcal{T}_n v$ is concave on E.

Proof. First note that $\mathcal{T}_n v(x) < \infty$ for all $x \in E$ and that E is convex. Let $x, x' \in E$ and $\alpha \in (0, 1)$. For $\varepsilon > 0$ there exist $a \in D_n(x)$ and $a' \in D_n(x')$ with

$$L_n v(x, a) \geq \mathcal{T}_n v(x) - \varepsilon,$$
$$L_n v(x', a') \geq \mathcal{T}_n v(x') - \varepsilon.$$

The convexity of D_n implies

$$\alpha(x, a) + (1 - \alpha)(x', a') \in D_n$$

which means that $\alpha a + (1 - \alpha)a' \in D_n(\alpha x + (1 - \alpha)x')$. Hence by (ii)

$$\begin{aligned} T_n v(\alpha x + (1-\alpha)x') &\geq L_n v\big(\alpha x + (1-\alpha)x', \alpha a + (1-\alpha)a'\big) \\ &\geq \alpha L_n v(x,a) + (1-\alpha) L_n v(x',a') \\ &\geq \alpha \mathcal{T}_n v(x) + (1-\alpha)\mathcal{T}_n v(x') - \varepsilon. \end{aligned}$$

This is true for all $\varepsilon > 0$, and the statement follows. □

Proposition 2.4.18 now directly implies that the following conditions on the data of the Markov Decision Model guarantee that (SA_N) is satisfied with the set $\mathbb{M}_n$ being a subset of concave functions.

Theorem 2.4.19. *Suppose a Markov Decision Model with upper bounding function b is given and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *D_n is convex in $E \times A$,*
(ii) *the mapping $(x,a) \mapsto \int v(x') Q_n(dx'|x,a)$ is concave for all concave $v \in \mathbb{B}_b^+$,*
(iii) *$(x,a) \mapsto r_n(x,a)$ is concave,*
(iv) *g_N is concave on E,*
(iv) *for all concave $v \in \mathbb{B}_b^+$ there exists a maximizer $f_n \in \Delta_n$ of v.*

Then the sets $\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is concave}\}$ and Δ_n satisfy the Structure Assumption (SA_N).

Remark 2.4.20. If $A = \mathbb{R}$ and $D(x) = [\underline{d}(x), \bar{d}(x)]$ then D is convex in $E \times A$ if and only if E is convex, $\underline{d} : E \to \mathbb{R}$ is convex and $\bar{d} : E \to \mathbb{R}$ is concave. ♢

Proposition 2.4.21. *Let $v \in \mathbb{B}_b^+$ and suppose that the following assumptions are satisfied:*

(i) *E is convex and $D_n := E \times A$,*
(ii) *$x \mapsto L_n v(x,a)$ is convex for all $a \in A$.*

Then $\mathcal{T}_n v$ is convex on E. If moreover A is a polytope and $a \mapsto L_n v(x,a)$ is convex for all $x \in E$, then there exists a so-called bang-bang *maximizer f_n^* of v at time n, i.e. $f_n^*(x)$ is a vertex of A for all $x \in E$.*

Proof. The first statement follows from the fact that the supremum of an arbitrary number of convex functions is again convex (because $\mathcal{T}_n v < \infty$). Now if A is a polytope, the convex function $a \mapsto L_n v(x,a)$ attains its supremum in a vertex and the set of all maximum points of $a \mapsto L_n v(x,a)$ which are vertices, is finite for all $x \in E$. Then, by applying the Selection Theorem A.2.3, we obtain the second statement. □

Proposition 2.4.21 now directly implies that the following conditions on the data of the Markov Decision Model guarantee that (SA_N) is satisfied with the set $\mathbb{M}_n$ being a subset of convex functions.

Theorem 2.4.22. *Suppose a Markov Decision Model with upper bounding function b is given and for all $n = 0, 1, \ldots, N-1$ it holds:*

(i) *E is convex and $D_n := E \times A$,*
(ii) *for all convex $v \in \mathbb{B}_b^+$, $x \mapsto \int v(x')Q_n(dx'|x,a)$ is convex for all $a \in A$,*
(iii) *$x \mapsto r_n(x,a)$ is convex for all $a \in A$,*
(iv) *g_N is convex,*
(v) *for all convex $v \in \mathbb{B}_b^+$ there exists a maximizer $f_n \in \Delta_n$ of v.*

Then the sets $\mathbb{M}_n := \{v \in \mathbb{B}_b^+ \mid v \text{ is convex}\}$ and Δ_n satisfy the Structure Assumption (SA_N).

2.4.5 Comparison of Markov Decision Models

When the value functions of a Markov Decision Model have one of the properties of the last section (e.g. monotonicity, convexity, concavity), then it is possible to discuss the qualitative influence of the transition kernel on the value function. More precisely, we want to know in which direction the value function changes, if the transition kernel $Q_n(\cdot|x,a)$ is replaced by $\tilde{Q}_n(\cdot|x,a)$. To this end, denote

$$\begin{aligned}\mathbb{M}_{st} &:= \{v \in \mathbb{B}_b^+ \mid v \text{ is increasing}\}\\ \mathbb{M}_{cv} &:= \{v \in \mathbb{B}_b^+ \mid v \text{ is concave}\}\\ \mathbb{M}_{cx} &:= \{v \in \mathbb{B}_b^+ \mid v \text{ is convex}\}.\end{aligned}$$

We denote by $(\tilde{V}_n)$ the value functions of the Markov Decision Model with transition kernels $\tilde{Q}_n$. In the following theorem we use stochastic orders for the transition kernels (see Appendix B.3 for details).

Theorem 2.4.23. *Suppose a Markov Decision Model with upper bounding function b is given which satisfies the Structure Assumption (SA_N) with the set $\mathbb{M}_n^\diamond$ where $\diamond \in \{st, cv, cx\}$. If for all $n = 0, 1, \ldots, N-1$*

$$Q_n(\cdot|x,a) \leq_\diamond \tilde{Q}_n(\cdot|x,a), \quad \text{for all } (x,a) \in D$$

then $V_n \leq \tilde{V}_n$ for $n = 0, 1, \ldots, N-1$.

The proof follows directly from the properties of the stochastic orders (see Appendix B.3).

2.5 Stationary Markov Decision Models

In this section we consider stationary Markov Decision Models, i.e. the data does not depend on n and is given by (E, A, D, Q, r_n, g_N) with $r_n := \beta^n r$, $g_N := \beta^N g$ and $\beta \in (0, 1]$.

We denote by F the set of all decision rules $f : E \to A$ with $f(x) \in D(x)$ for $x \in E$. Then F^N is the set of all N-stage policies $\pi = (f_0, \ldots, f_{N-1})$.

The expected discounted reward over n stages under a policy $\pi \in F^n$ is given by

$$J_{n\pi}(x) := \mathbb{E}_x^\pi \left[\sum_{k=0}^{n-1} \beta^k r\big(X_k, f_k(X_k)\big) + \beta^n g(X_n) \right], \quad x \in E$$

when the system starts in state $x \in E$. The maximal expected discounted reward over n stages is defined by

$$\begin{aligned} J_0(x) &:= g(x) \\ J_n(x) &:= \sup_{\pi \in F^n} J_{n\pi}(x), \quad x \in E,\ 1 \le n \le N. \end{aligned}$$

In order to obtain a well-defined stochastic optimization problem we need the following integrability assumption (see Section 2.2):

Assumption ($\mathbf{A}_N$): *For $x \in E$*

$$\delta_N(x) := \sup_\pi \mathbb{E}_x^\pi \left[\sum_{k=0}^{N-1} \beta^k r^+(X_k, f_k(X_k)) + \beta^N g^+(X_N) \right] < \infty.$$

Remark 2.5.1. Since $\delta_0 := g^+ \le \delta_{n-1} \le \delta_n \le \delta_N$ and

$$\delta_n^N = \beta^n \delta_{N-n}, \quad n = 0, 1 \ldots, N,$$

the Integrability Assumption (A_N) is equivalent to the integrability assumption in Section 2.2 when we have a stationary Markov Decision Model. ◊

As explained in Section 2.4 it is convenient to show that (A_N) is satisfied by proving the existence of an upper bounding function. The definition of an upper bounding function for a stationary Markov Decision Model is as follows.

Definition 2.5.2. A measurable mapping $b : E \to \mathbb{R}_+$ is called an *upper bounding function* for the stationary Markov Decision Model if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$, such that:

(i) $r^+(x,a) \le c_r b(x)$ for all $(x,a) \in D$,
(ii) $g^+(x) \le c_g b(x)$ for all $x \in E$,
(iii) $\int b(x')Q(dx'|x,a) \le \alpha_b b(x)$ for all $(x,a) \in D$.

If the stationary Markov Decision Model has an upper bounding function b, then we have $\delta_N \in \mathbb{B}_b$ and the Integrability Assumption (A_N) is satisfied (cf. Proposition 2.4.2). Obviously every stationary model is a special non-stationary model. We obtain the following relation between the value functions J_n and V_n:

$$V_n(x) = \beta^n J_{N-n}(x), \quad x \in E,\ n = 0, 1, \ldots, N.$$

But on the other hand, every non-stationary Markov Decision Model can be formulated as a stationary one. The idea is to extend the state space by including the time parameter.

As in Definition 2.3.1 we introduce the following operators for $v \in \mathbb{M}(E)$:

$$\begin{aligned} Lv(x,a) &:= r(x,a) + \beta \int v(x')Q(dx'|x,a), \quad (x,a) \in D, \\ \mathcal{T}_f v(x) &:= Lv\big(x, f(x)\big), \quad x \in E \\ \mathcal{T} v(x) &:= \sup_{a \in D(x)} Lv(x,a), \quad x \in E. \end{aligned}$$

$\mathcal{T}$ is called the *maximal reward operator*. The reward iteration reads now as follows.

Theorem 2.5.3 (Reward Iteration). *For $\pi = (f_0, \ldots, f_{n-1})$ it holds:*

$$J_{n\pi} = \mathcal{T}_{f_0} \ldots \mathcal{T}_{f_{n-1}} g.$$

The Structure Assumption (SA_N) has to be modified for stationary Markov Decision Models.

Structure Assumption ($\mathbf{SA}_N$): *There exist sets $\mathbb{M} \subset \mathbb{M}(E)$ and $\Delta \subset F$ such that:*

(i) $g \in \mathbb{M}$.
(ii) *If $v \in \mathbb{M}$ then $\mathcal{T}v(x)$ is well-defined and $\mathcal{T}v \in \mathbb{M}$.*
(iii) *For all $v \in \mathbb{M}$ there exists a maximizer $f \in \Delta$ of v, i.e.*

$$\mathcal{T}_f v(x) = \mathcal{T} v(x), \quad x \in E.$$

The main Theorem 2.3.8 about the recursive computation of the optimal value functions has now the following form.

Theorem 2.5.4 (Structure Theorem). *Let (SA_N) be satisfied.*

a) Then $J_n \in \mathbb{M}$ and the Bellman equation $J_n = \mathcal{T} J_{n-1}$ holds, i.e.

$$J_0(x) = g(x)$$
$$J_n(x) = \sup_{a \in D(x)} \left\{ r(x,a) + \beta \int J_{n-1}(x') Q(dx'|x,a) \right\}, \quad x \in E.$$

Moreover, $J_n = \mathcal{T}^n g$.

b) For $n = 1, \ldots, N$ there exist maximizers f_n^* of J_{n-1} with $f_n^* \in \Delta$, and every sequence of maximizers f_n^* of J_{n-1} defines an optimal policy $(f_N^*, \ldots, f_1^*)$ for the stationary N-stage Markov Decision Model.

In many examples we will see that the Structure Assumption is naturally fulfilled. For some conditions which imply (SA_N) see Section 2.4. The simplest case arises when both E and A are finite. Here the Structure Assumption (SA_N) is satisfied with $\mathbb{M} := \{v : E \to \mathbb{R}\}$ because every function is measurable and maximizers exist. Moreover, the transition kernel has a discrete density and we denote

$$q(x'|x,a) := Q(\{x'\}|x,a)$$

for $x, x' \in E$ and $a \in D(x)$.
Analogously to the non-stationary case, Theorem 2.5.4 gives a recursive algorithm to solve Markov Decision Problems. Due to the stationarity of the data however, it is not necessary to formulate the algorithm as a backward algorithm.

Forward Induction Algorithm.

1. Set $n := 0$ and for $x \in E$:
$$J_0(x) := g(x).$$
2. Set $n := n + 1$ and compute for all $x \in E$
$$J_n(x) = \sup_{a \in D(x)} \left\{ r(x,a) + \int J_{n-1}(x') Q(dx'|x,a) \right\}.$$
Compute a maximizer f_n^* of J_{n-1}.
3. If $n = N$, then the value function J_N is computed and the optimal policy π^* is given by $\pi^* = (f_N^*, \ldots, f_1^*)$. Otherwise, go to step 2.

The *induction algorithm* computes the n-stage value functions and the optimal decision rules recursively over the stages, beginning with the terminal reward function. We illustrate this procedure with the following numerical example which is known as Howard's toymaker in the literature.

Example 2.5.5 (Howard's Toymaker). Suppose a Markov Decision Model is given by the following data. The planning horizon is $N = 4$. The state space consists of two states $E = \{1, 2\}$ as well as the action space $A = \{1, 2\}$. We have no restriction on the actions, i.e. $D(x) = A$. The reward is discounted by a factor $\beta \in (0, 1)$ and the one-stage reward is given by $r(1, 1) = 6,\ r(2, 1) = -3,\ r(1, 2) = 4,\ r(2, 2) = -5$. The terminal reward is $g(1) = 105,\ g(2) = 100$. The transition probabilities are denoted by the following matrices (see Figure 2.2)

$$q(\cdot|\cdot, 1) = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{pmatrix} \quad q(\cdot|\cdot, 2) = \begin{pmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{pmatrix}.$$

Note that $q(\cdot|\cdot, 1)$ gives the transition probabilities if action $a = 1$ is chosen and $q(\cdot|\cdot, 2)$ gives the transition probabilities if action $a = 2$ is chosen.

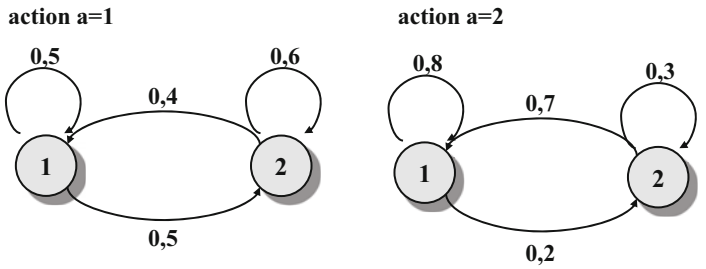

Fig. 2.2 Transition probabilities of Howard's toymaker.

The result of the computation with $\beta = 0.98$ is shown in Table 2.1. By $D_n^*(x)$ we denote the set

$$D_n^*(x) := \{a \in D(x) \mid a \text{ maximizes } \ b \mapsto r(x, b) + \beta \sum_{x' \in E} q(x'|x, b) J_{n-1}(x')\}.$$

In particular, the value function of the 4-stage problem is given by $J_4(1) = 106.303$ and $J_4(2) = 96.326$.
Moreover, if $\Delta_0 := J_0(1) - J_0(2) \geq 0$ then the following *Turnpike Theorem* can be shown:

- If $\beta < \frac{20}{29}$, then there exists an $N^* = N^*(\beta, \Delta_0) \in \mathbb{N}$ such that $D_n^*(1) = D_n^*(2) = \{1\}$ for $n \geq N^*$ and $2 \in D_n^*(1) = D_n^*(2)$ for $n < N^*$.
- If $\beta > \frac{20}{29}$, then there exists an $N^* = N^*(\beta, \Delta_0) \in \mathbb{N}$ such that $D_n^*(1) = D_n^*(2) = \{2\}$ for $n \geq N^*$ and $1 \in D_n^*(1) = D_n^*(2)$ for $n < N^*$.

n	$J_n(1)$	$J_n(2)$	$D_n^*(1)$	$D_n^*(2)$
0	105	100		
1	106.45	96.96	1	1
2	106.461	96.531	2	2
3	106.385	96.412	2	2
4	106.303	96.326	2	2

Table 2.1 Computational results of the Backward induction algorithm.

- If $\beta = \frac{20}{29}$, then

$$\begin{array}{ll} D_n^*(1) = D_n^*(2) = \{1\} \text{ for all } n \in \mathbb{N} & \text{if } \Delta_0 < \frac{29}{3} \\ D_n^*(1) = D_n^*(2) = A \text{ for all } n \in \mathbb{N} & \text{if } \Delta_0 = \frac{29}{3} \\ D_n^*(1) = D_n^*(2) = \{2\} \text{ for all } n \in \mathbb{N} & \text{if } \Delta_0 > \frac{29}{3}. \end{array}$$ ♦

The following Markov Decision Model with random discounting has important applications in finance and insurance.

Example 2.5.6 (Random Discounting). Suppose a stationary Markov Decision Model $(E, A, D, Q, \beta^n r, \beta^N g)$ is given. Sometimes the discount factors for the stages vary randomly. Here we assume that the (non-negative) discount factors (β_n) form a Markov process, given by a transition kernel $Q^\beta(B|y)$ and β_0 is given. We suppose that (β_n) is independent of the state process. (Also the more general case where (β_n) depends on the state process can be dealt with.) Then we are interested in finding the maximal expected discounted reward over all policies $\pi \in F^N$, i.e. we want to maximize the expression

$$\mathbb{E}_x^\pi \left[\sum_{n=0}^{N-1} \Big(\prod_{k=0}^{n-1} \beta_k \Big) r\big(X_n, f_n(\beta_0, \ldots, \beta_{n-1}, X_n)\big) + \Big(\prod_{k=0}^{N-1} \beta_k \Big) g(X_N) \right], \quad x \in E.$$

Of course we assume that the Markov process (β_n) can be observed by the decision maker and thus the decision rules are allowed to depend on it. This problem can again be solved via a standard Markov Decision Model by extending the state space E. Let us define:

- $\tilde{E} := E \times [0, \infty) \times [0, \infty)$ where $(x, \beta, \delta) \in \tilde{E}$ denotes the state x, the new discount factor β and the product δ of the discount factors realized so far,
- $\tilde{A} = A$,
- $\tilde{D}(x, \beta, \delta) = D(x)$ for all $(x, \beta, \delta) \in E$,
- $\tilde{Q}(B \times B_1 \times B_2 | x, \beta, \delta, a) = Q(B|x, a) \otimes Q^\beta(B_1|\beta) \otimes 1_{B_2}(\delta \cdot \beta)$ for all $(x, \beta, \delta) \in \tilde{E}$ and suitable measurable sets B, B_1, B_2,
- $\tilde{r}\big((x, \beta, \delta), a\big) = \delta r(x, a)$ for all $(x, \beta, \delta) \in \tilde{E}$,
- $\tilde{g}(x, \beta, \delta) = \delta g(x)$.

Then the maximal reward operator for $v \in \mathbb{M}(E)$ is given by

$$\mathcal{T}v(x,\beta,\delta) = \sup_{a\in D(x)} \left\{ \delta r(x,a) + \int Q(dx'|x,a) \int Q^{\beta}(dy'|\beta) v(x',y',\delta\beta) \right\}$$

and the solution of the preceding problem with random discounting is given by the value

$$J_N(x,\beta_0,1) = \mathcal{T}^N \tilde{g}(x,\beta_0,1).$$

Note that the optimal decision rule at time n depends on x_n, β_n and the product of the discount factors $\delta_n = \prod_{k=0}^{n-1} \beta_k$. ♦

2.6 Applications and Examples

In this section we collect some applications of the theory presented so far. Further examples with a focus on financial optimization problems can be found in Chapter 4.

2.6.1 Red-and-Black Card Game

Consider the following simple card game: The dealer uncovers successively the cards of a well-shuffled deck which initially contains b_0 black and r_0 red cards. The player can at any time stop the uncovering of the cards. If the next card at the stack is black (red), the player wins (loses) 1 Euro. If the player does not stop the dealer, then the colour of the last card is decisive. When the player says stop right at the beginning then the probability of winning 1 Euro is obviously $\frac{b_0}{b_0+r_0}$ and her expected gain will be $\frac{b_0-r_0}{b_0+r_0}$. The same holds true when the player waits until the last card. Is there a strategy which yields a larger expected gain? The answer is no! We will prove this by formulating the problem as a stationary Markov Decision Model. The state of the system is the number of cards in the stack, thus

$$E := \{x = (b,r) \in \mathbb{N}_0^2 \mid b \le b_0, r \le r_0\}.$$

The state $(0,0)$ will be absorbing, thus in view of Example 2.3.13 we define $G := \{(0,0)\}$ as the set of absorbing states. Once, we have entered G the reward is zero and we stay in G. For $x \in E$ and $x \notin \{(0,1),(1,0)\}$ we have $D(x) = A = \{0,1\}$ with the interpretation that $a = 0$ means 'go ahead' and $a = 1$ means 'stop'. Since the player has to take the last card if she had not stopped before we have $D\big((0,1)\big) = D\big((1,0)\big) = \{1\}$. The transition probabilities are given by

$$\begin{aligned} q\big((b,r-1) \mid (b,r),0\big) &:= \frac{r}{r+b}, \quad r \geq 1, b \geq 0 \\ q\big((b-1,r) \mid (b,r),0\big) &:= \frac{b}{r+b}, \quad r \geq 0, b \geq 1 \\ q\big((0,0) \mid (b,r),1\big) &:= 1, \quad (b,r) \in E. \\ q\big((0,0) \mid (0,0),a\big) &:= 1, \quad a \in A. \end{aligned} \tag{2.1}$$

The one-stage reward is given by the expected reward (according to Remark 2.1.2),

$$r\big((b,r),1\big) := \frac{b-r}{b+r} \quad \text{for } (b,r) \in E \setminus G,$$

and the reward is zero otherwise. Finally we define

$$g(b,r) := \frac{b-r}{b+r} \quad \text{for } (b,r) \in E \setminus G$$

and $g((0,0)) = 0$. We summarize now the data of the stationary Markov Decision Model.

- $E := \{x = (b,r) \in \mathbb{N}_0^2 \mid b \leq b_0, r \leq r_0\}$ where $x = (b,r)$ denotes the number of black and red cards in the stack,
- $A := \{0,1\}$ where $a = 0$ means 'go ahead' and $a = 1$ means 'stop',
- $D(x) = A$ for $x \notin \{(0,1),(1,0)\}$ and $D\big((0,1)\big) = D\big((1,0)\big) = \{1\}$,
- the transition probabilities are given by equation (2.1),
- the one-stage reward is $r\big(x,1\big) = \frac{b-r}{b+r}$ for $x = (b,r) \in E \setminus G$ and 0 otherwise,
- $g(x) = \frac{b-r}{b+r}$ for $x = (b,r) \in E \setminus G$, $g((0,0)) = 0$,
- $N := r_0 + b_0$ and $\beta := 1$.

Since E and A are finite, (A_N) and also the Structure Assumption (SA_N) is clearly satisfied with

$$\mathbb{M} := \{v : E \to \mathbb{R} \mid v(x) = 0 \text{ for } x \in G\} \quad \text{and} \quad \Delta := F.$$

In particular we immediately know that an optimal policy exists. The maximal reward operator is given by

$$\mathcal{T}v(b,r) := \max\left\{\frac{b-r}{b+r}\,,\ \frac{r}{r+b}v(r-1,b) + \frac{b}{r+b}v(r,b-1)\right\} \quad \text{for } b+r \geq 2,$$

$$\mathcal{T}v(1,0) := 1,$$

$$\mathcal{T}v(0,1) := -1,$$

$$\mathcal{T}v(0,0) := 0.$$

It is not difficult to see that $g = \mathcal{T}g$. For $x = (b,r) \in E$ with $r + b \geq 2$ the computation is as follows:

$$\begin{aligned}\mathcal{T}g(b,r) &= \max\left\{\frac{b-r}{b+r}\,,\ \frac{r}{r+b}g(r-1,b)+\frac{b}{r+b}g(r,b-1)\right\}\\ &= \max\left\{\frac{b-r}{b+r}\,,\ \frac{r}{r+b}\cdot\frac{b-r+1}{r+b-1}+\frac{b}{r+b}\cdot\frac{b-r-1}{r+b-1}\right\}\\ &= \max\left\{\frac{b-r}{b+r}\,,\ \frac{b-r}{b+r}\right\} = g(b,r).\end{aligned}$$

Since both expressions for $a = 0$ and $a = 1$ are identical, every $f \in F$ is a maximizer of g. Applying Theorem 2.5.4 we obtain that $J_n = \mathcal{T}^n g = g$ and we can formulate the solution of the card game.

Theorem 2.6.1. *The maximal value of the card game is given by*

$$J_{r_0+b_0}(b_0,r_0) = g(b_0,r_0) = \frac{b_0-r_0}{b_0+r_0},$$

and every strategy is optimal.

Thus, there is no strategy which yields a higher expected reward than the trivial ones discussed above. Note that the game is fair (i.e. $J_N(b_0,r_0) = 0$) if and only if $r_0 = b_0$.

2.6.2 A Cash Balance Problem

The cash balance problem involves the decision about the optimal cash level of a firm over a finite number of periods. The aim is to use the firm's liquid assets efficiently. There is a random stochastic change in the cash reserve each period (which can be both positive and negative). Since the firm does not earn interest from the cash position, there are holding cost or *opportunity cost* for the cash reserve if it is positive. But also in case the cash reserve is negative the firm incurs an out-of-pocket expense and has to pay interest. The cash reserve can be increased or decreased by the management at the beginning of each period which implies transfer costs. To keep the example simple we assume that the random changes in the cash flow are given by independent and identically distributed random variables (Z_n) with finite expectation. The transfer cost are linear. More precisely, let us define a function $c : \mathbb{R} \to \mathbb{R}_+$ by

$$c(z) := c_u z^+ + c_d z^-$$

where $c_u, c_d > 0$. The transfer cost are then given by $c(z)$ if the amount z is transferred. The cost $L(x)$ have to be paid at the beginning of a period for cash level x. We assume that

- $L : \mathbb{R} \to \mathbb{R}_+$, $L(0) = 0$,

- $x \mapsto L(x)$ is convex,
- $\lim_{|x|\to\infty} \frac{L(x)}{|x|} = \infty$.

This problem can be formulated as a Markov Decision Model with disturbances (Z_n) and with state space $E := \mathbb{R}$, where the state $x \in E$ is the current cash level. At the beginning of each period we have to decide upon the new cash level $a \in A := \mathbb{R}$. All actions are admissible, i.e. $D(x) := A$. The reward is then given as the negative cost $r(x,a) := -c(a-x) - L(a)$ of transfer and holding cost. The transition function is given by

$$T(x,a,z) := a - z$$

where z is a realization of the stochastic cash change Z_{n+1}. There is no terminal reward, i.e. $g \equiv 0$ and cost are discounted by a factor $\beta \in (0,1]$. The planning horizon N is given. We summarize the data of the stationary Markov Decision Model:

- $E := \mathbb{R}$ where $x \in E$ denotes the cash level,
- $A := \mathbb{R}$ where $a \in A$ denotes the new cash level after transfer,
- $D(x) := A$,
- $\mathcal{Z} := \mathbb{R}$ where $z \in \mathcal{Z}$ denotes the cash change,
- $T(x,a,z) := a - z$,
- $Q^Z(\cdot|x,a) :=$ distribution of Z_{n+1} (independent of (x,a)),
- $r(x,a) := -c(a-x) - L(a)$,
- $g \equiv 0$,
- $\beta \in (0,1]$.

Obviously the reward is bounded from above, i.e. $b \equiv 1$ is an upper bounding function. In what follows we will treat this problem as one of minimizing cost which seems to be more natural. The minimal cost operator is given by:

$$\mathcal{T}v(x) := \min \left\{ \begin{array}{ll} \inf\limits_{a>x} \Big\{(a-x)c_u + L(a) + \beta\,\mathbb{E}\,v(a-Z)\Big\}, & (2.2) \\ L(x) + \beta\,\mathbb{E}\,v(x-Z), & (2.3) \\ \inf\limits_{a<x} \Big\{(x-a)c_d + L(a) + \beta\,\mathbb{E}\,v(a-Z)\Big\} \Big\} & (2.4) \end{array} \right.$$

where $Z := Z_1$. We will next check the Structure Assumption (SA_N). Thus, we first have to find a reasonable set $\mathbb{M}$. Looking at $\mathcal{T}v$ we choose the Ansatz:

$$\mathbb{M} := \{v : E \to \mathbb{R}_+ \mid v \text{ is convex and } v(x) \le c(-x) + d \text{ for some } d \in \mathbb{R}_+\}.$$

Moreover, we will see below that the set of minimizers is of a special form. Obviously $0 \in \mathbb{M}$. Now let $v \in \mathbb{M}$ and define the functions

$$h_u(a) := (a-x)c_u + L(a) + \beta\,\mathbb{E}\,v(a-Z),$$
$$h_d(a) := (x-a)c_d + L(a) + \beta\,\mathbb{E}\,v(a-Z).$$

By the definition of $\mathbb{M}$ both functions are finite on $\mathbb{R}$, since for $a \in A$ we obtain

$$\mathbb{E}\,v(a-Z) \le d + \mathbb{E}\,c(Z-a) \le d + \mathbb{E}\,|a-Z|(c_u+c_d) < \infty.$$

Also both functions are convex and $\lim_{|a|\to\infty} h_u(a) = \lim_{|a|\to\infty} h_d(a) = \infty$. Thus, both have a well-defined finite minimum point. Moreover, the convexity implies that the right- and left-hand side derivative at each point exist. A minimum point is characterized by a non-negative right-hand side derivative and a non-positive left-hand side derivative. Thus, we define

$$S_- := \inf\Big\{a \in \mathbb{R} \mid \frac{\partial^+}{\partial a} h_u(a) \ge 0\Big\},$$
$$S_+ := \sup\Big\{a \in \mathbb{R} \mid \frac{\partial^-}{\partial a} h_d(a) \le 0\Big\},$$

where $\frac{\partial^+}{\partial a}h$ and $\frac{\partial^-}{\partial a}h$ denote the right- and left-hand side derivative respectively. Since $h_u(a) - h_d(a) = (a-x)(c_u+c_d)$ is increasing in a, we have $S_- \le S_+$. It is important to note that S_- and S_+ do not depend on x. In order to determine a minimum point of $\mathcal{T}v$ we distinguish three cases:

(i) $x < S_-$: In this case the infimum of (2.4) is obtained if we plug in $a = x$ and thus the values of (2.4) and (2.3) are equal. However, the infimum of (2.2) is attained in $a = S_-$ and is less or equal to the value of (2.3) since $h_u(S_-) \le h_u(x) = L(x) + \beta\,\mathbb{E}\,v(x-Z)$.

(ii) $S_- \le x \le S_+$: Here the minimum values of the three expressions are equal and $a = x$ is the global minimum point.

(iii) $S_+ < x$: This case is analogous to the first one and the global minimum is attained in $a = S_+$.

Hence we have shown that a minimizer f^* exists and is of the form

$$f^*(x) = \begin{cases} S_- & \text{if } x < S_- \\ x & \text{if } S_- \le x \le S_+ \\ S_+ & \text{if } x > S_+. \end{cases} \tag{2.5}$$

This means that if the cash level is below S_-, sell enough securities to bring the cash level up to S_-. If the cash level is between the limits do nothing, and if the cash level is above S_+, buy securities to reduce the cash level to this critical level. Note that S_+ and S_- depend on v. As a consequence we define

$$\Delta := \Big\{ f \in F \;\Big|\; \text{there exist } \; S_-, S_+ \in \mathbb{R} \text{ with } \\ S_- \le S_+ \text{ and } f \text{ is of the form (2.5)} \Big\}.$$

Inserting the minimizer gives

$$\mathcal{T}v(x) = \begin{cases} (S_- - x)c_u + L(S_-) + \beta\,\mathbb{E}\,v(S_- - Z) & \text{if } x < S_- \\ L(x) + \beta\,\mathbb{E}\,v(x - Z) & \text{if } S_- \le x \le S_+ \\ (x - S_+)c_d + L(S_+) + \beta\,\mathbb{E}\,v(S_+ - Z) & \text{if } x > S_+. \end{cases}$$

It is not difficult to verify that this function is again in $\mathbb{M}$. First there exists $d \in \mathbb{R}_+$ such that $\mathcal{T}v(x) \le d + c(-x)$. The convexity of $\mathcal{T}v$ on the intervals $(-\infty, S_-)$, (S_-, S_+), (S_+, ∞) is also obvious. It remains to investigate the points S_- and S_+. Here we have to show that the left-hand side derivative is less than or equal to the right-hand side derivative. Due to the definition of S_- and S_+ we obtain

$$c_u + \frac{\partial^+}{\partial x}\Big(L(x) + \beta\,\mathbb{E}\,v(x - Z)\Big)\Big|_{x=S_-} \ge 0$$
$$-c_d + \frac{\partial^-}{\partial x}\Big(L(x) + \beta\,\mathbb{E}\,v(x - Z)\Big)\Big|_{x=S_+} \le 0$$

since S_- and S_+ are the minimum points. This observation yields $\mathcal{T}v \in \mathbb{M}$. Thus, the Structure Assumption (SA_N) is satisfied for $\mathbb{M}$ and Δ. Theorem 2.5.4 can be applied to the cash balance problem and we obtain the following result.

Theorem 2.6.2. *a) There exist critical levels S_{n-} and S_{n+} such that for $n = 1, \dots, N$*

$$J_n(x) = \begin{cases} (S_{n-} - x)c_u + L(S_{n-}) + \beta\,\mathbb{E}\,J_{n-1}(S_{n-} - Z) & \text{if } x < S_{n-} \\ L(x) + \beta\,\mathbb{E}\,J_{n-1}(x - Z) & \text{if } S_{n-} \le x \le S_{n+} \\ (x - S_{n+})c_d + L(S_{n+}) + \beta\,\mathbb{E}\,J_{n-1}(S_{n+} - Z) & \text{if } x > S_{n+}. \end{cases}$$

with $J_0 \equiv 0$.

b) The optimal cash balance policy is given by $(f_N^*, \dots, f_1^*)$ where f_n^* is

$$f_n^*(x) := \begin{cases} S_{n-} & \text{if } x < S_{n-} \\ x & \text{if } S_{n-} \le x \le S_{n+} \\ S_{n+} & \text{if } x > S_{n+}. \end{cases} \tag{2.6}$$

Note that the critical levels which determine the transfer regions (sell, buy, do nothing) depend on n. Obviously the transfer cost imply that it is unlikely

that many transfers occur. Hence the problem is sometimes also called *smoothing problem.*

2.6.3 Stochastic Linear-Quadratic Problems

A famous class of control problems with various different applications are linear-quadratic problems (LQ-problems). The name stems from the linear state transition function and the quadratic cost function. In what follows we suppose that $E := \mathbb{R}^m$ is the state space of the underlying system and $D_n(x) := A := \mathbb{R}^d$, i.e. all actions are admissible. The state transition functions are linear in state and action with random coefficient matrices $A_1, B_1, \ldots, A_N, B_N$ with suitable dimensions, i.e. the system transition functions are given by

$$T_n(x, a, A_{n+1}, B_{n+1}) := A_{n+1}x + B_{n+1}a.$$

Thus, the disturbance in $[n, n+1)$ is given by $Z_{n+1} := (A_{n+1}, B_{n+1})$. The distribution of Z_{n+1} is influenced neither by the state nor by the action, and the random matrices $Z_1, Z_2, \ldots$ are supposed to be independent but not necessarily identically distributed and have finite expectation and covariance. Moreover, we assume that $\mathbb{E}\left[B_{n+1}^\top Q B_{n+1}\right]$ is positive definite for all symmetric positive definite Q. Obviously we obtain a non-stationary problem. The one-stage reward is a negative cost function

$$r_n(x, a) := -x^\top Q_n x$$

and the terminal reward is

$$g_N(x, a) := -x^\top Q_N x$$

with deterministic, symmetric and positive definite matrices $Q_0, Q_1, \ldots, Q_N$. There is no discounting. The aim is to minimize

$$\mathbb{E}_x^\pi \left[\sum_{k=0}^{N} X_k^\top Q_k X_k\right]$$

over all N-stage policies π. Thus, the aim is here to keep the state of the system close to zero. We summarize the data of the Markov Decision Model with disturbances (Z_n) as follows.

- $E := \mathbb{R}^m$ where $x \in E$ denotes the system state,
- $A := \mathbb{R}^d = D_n(x)$ where $a \in A$ denotes the action,
- $\mathcal{Z} := \mathbb{R}^{(m,m)} \times \mathbb{R}^{(m,d)}$ where $Z = (A, B)$ with values in $\mathcal{Z}$ denotes the random transition coefficients of the linear system,

- $T_n(x, a, A, B) := Ax + Ba$,
- $Q^Z(\cdot|x, a) :=$ distribution of $Z_{n+1} := (A_{n+1}, B_{n+1})$ (independent of (x, a)),
- $r_n(x, a) := -x^\top Q_n x$,
- $g_N(x, a) := -x^\top Q_N x$,
- $\beta := 1$.

We have $r \leq 0$ and $b \equiv 1$ is an upper bounding function. Thus, (A_N) is satisfied. We will treat this problem as a cost minimization problem, i.e. we suppose that V_n is the minimal cost in the period $[n, N]$. For the calculation below we assume that all expectations exist. There are various applications of this regulation problem in engineering, but it will turn out that problems of this type are also important for example for quadratic hedging or mean-variance problems. The minimal cost operator is given by

$$\mathcal{T}_n v(x) = \inf_{a \in \mathbb{R}^d} \left\{ x^\top Q_n x + \mathbb{E}\, v\big(A_{n+1}x + B_{n+1}a\big) \right\}.$$

We will next check the Structure Assumption (SA_N). It is reasonable to assume that $\mathbb{M}_n$ is given by

$$\mathbb{M}_n := \{v : \mathbb{R}^m \to \mathbb{R}_+ \mid v(x) = x^\top Q x \text{ with } Q \text{ symmetric, positive definite}\}.$$

It will also turn out that the sets $\Delta_n := \Delta \cap F_n$ can be chosen as the set of all linear functions, i.e.

$$\Delta := \{f : E \to A \mid f(x) = Cx \text{ for some } C \in \mathbb{R}^{(d,m)}\}.$$

Let us start with (SA_N)(i): Obviously $x^\top Q_N x \in \mathbb{M}_N$. Now let $v(x) = x^\top Q x \in \mathbb{M}_{n+1}$. We try to solve the following optimization problem

$$\begin{aligned}\mathcal{T}_n v(x) &= \inf_{a \in \mathbb{R}^d} \left\{ x^\top Q_n x + \mathbb{E}\, v\big(A_{n+1}x + B_{n+1}a\big) \right\} \\ &= \inf_{a \in \mathbb{R}^d} \Big\{ x^\top Q_n x + x^\top \mathbb{E}\left[A_{n+1}^\top Q A_{n+1}\right] x + 2x^\top \mathbb{E}\left[A_{n+1}^\top Q B_{n+1}\right] a \\ &\qquad + a^\top \mathbb{E}\left[B_{n+1}^\top Q B_{n+1}\right] a \Big\}.\end{aligned}$$

Since Q is positive definite, we have by assumption that $\mathbb{E}\left[B_{n+1}^\top Q B_{n+1}\right]$ is also positive definite and thus regular and the function in brackets is convex in a (for fixed $x \in E$). Differentiating with respect to a and setting the derivative equal to zero, we obtain that the unique minimum point is given by

$$f_n^*(x) = -\Big(\mathbb{E}\left[B_{n+1}^\top Q B_{n+1}\right] \Big)^{-1} \mathbb{E}\left[B_{n+1}^\top Q A_{n+1}\right] x.$$

Inserting the minimum point into the equation for $\mathcal{T}_n v$ yields

$$\mathcal{T}_n v(x) = x^\top \Big(Q_n + \mathbb{E}\big[A_{n+1}^\top Q A_{n+1}\big] - \mathbb{E}\big[A_{n+1}^\top Q B_{n+1}\big] \big(\mathbb{E}\big[B_{n+1}^\top Q B_{n+1}\big] \big)^{-1} \mathbb{E}\left[B_{n+1}^\top Q A_{n+1}\right] \Big) x = x^\top \tilde{Q} x$$

where $\tilde{Q}$ is defined as the expression in the brackets. Note that $\tilde{Q}$ is symmetric and since $x'\tilde{Q}x = \mathcal{T}_n v(x) \geq x^\top Q_n x$, it is also positive definite. Thus $\mathcal{T} v \in \mathbb{M}_n$ and the Structure Assumption (SA_N) is satisfied for $\mathbb{M}_n$ and $\Delta_n = \Delta \cap F_n$. Now we can apply Theorem 2.3.8 to solve the stochastic LQ-problem.

Theorem 2.6.3. *a) Let the matrices $\tilde{Q}_n$ be recursively defined by*

$$\begin{aligned} \tilde{Q}_N &:= Q_N \\ \tilde{Q}_n &:= Q_n + \mathbb{E}\left[A_{n+1}^\top \tilde{Q}_{n+1} A_{n+1}\right] \\ &\quad - \mathbb{E}\big[A_{n+1}^\top \tilde{Q}_{n+1} B_{n+1}\big] \big(\mathbb{E}\big[B_{n+1}^\top \tilde{Q}_{n+1} B_{n+1}\big]\big)^{-1} \mathbb{E}\big[B_{n+1}^\top \tilde{Q}_{n+1} A_{n+1}\big]. \end{aligned}$$

Then $\tilde{Q}_n$ are symmetric, positive semidefinite and $V_n(x) = x^\top \tilde{Q}_n x$ for $x \in E$.

b) The optimal policy $(f_0^*, \ldots, f_{N-1}^*)$ is given by

$$f_n^*(x) := -\Big(\mathbb{E}\left[B_{n+1}^\top \tilde{Q}_{n+1} B_{n+1}\right] \Big)^{-1} \mathbb{E}\left[B_{n+1}^\top \tilde{Q}_{n+1} A_{n+1}\right] x.$$

Note that the optimal decision rule is a linear function of the state and the coefficient matrix can be computed off-line. The minimal cost function is quadratic.

Our formulation of the stochastic LQ-problem can be generalized in different ways without leaving the LQ-framework. For example the transition function can be extended to

$$T_n(x, a, A_{n+1}, B_{n+1}, C_{n+1}) := A_{n+1} x + B_{n+1} a + C_{n+1}$$

where C_n are vectors with random entries. Thus, the stochastic disturbance variable is extended to $Z_n := (A_n, B_n, C_n)$ with the usual independence assumptions. Moreover, the cost function can be generalized to

$$\mathbb{E}_x^\pi \left[\sum_{k=0}^{N} (X_k - b_k)^\top Q_k (X_k - b_k) + \sum_{k=0}^{N-1} f_k(X_k)^\top \hat{Q}_k f_k(X_k) \right]$$

where $\hat{Q}_k$ are deterministic, symmetric positive semidefinite matrices and b_k are deterministic vectors. In this formulation the control itself is penalized and the distance of the state process to the benchmarks b_k has to be kept small. Note that in both generalizations the value functions remain of linear-quadratic form.

2.7 Exercises

Exercise 2.7.1 (Howard's Toymaker). Consider Howard's toymaker of Example 2.5.5. Show the stated Turnpike Theorem by conducting the following steps:

a) For a function $v : E \to \mathbb{R}$ let $\Delta v := v(1) - v(2)$ and show for all $a \in A$ that $Lv(1,a) - Lv(2,a) = 9 + 0.1\beta\Delta v$.
b) Show that $\Delta J_n = 9\sum_{k=0}^{n-1}(0.1\beta)^k + (0.1\beta)^n \Delta J_0$ for $n \in \mathbb{N}$.

Exercise 2.7.2. Suppose a stationary Markov Decision Model with planning horizon N is given which satisfies (A_N) and (SA_N). We define $J_n = \mathcal{T}^n g$ and $\hat{J}_n = \mathcal{T}^n \hat{g}$ for two terminal reward functions $g, \hat{g} \in \mathbb{M}$. Show:

a) For all $k = 1, \ldots, N$ it holds:

$$J_N - \hat{J}_k \leq \beta^N \sup_x \big(g(x) - \hat{g}(x)\big) + \sup_x \big(\hat{J}_k(x) - \hat{J}_{k-1}(x)\big) \sum_{j=1}^{N-k} \beta^j.$$

b) For all $k = 1, \ldots, N$ it holds:

$$J_N - \hat{J}_k \geq \beta^N \inf_x \big(g(x) - \hat{g}(x)\big) + \inf_x \big(\hat{J}_k(x) - \hat{J}_{k-1}(x)\big) \sum_{j=1}^{N-k} \beta^j.$$

c) The bounds for J_N in a) and b) are decreasing in k.

Exercise 2.7.3 (Card Game). Consider the following variant of the red-and-black card game. Suppose we have a deck of 52 cards which is turned over and cards are uncovered one by one. The player has to say 'stop' when she thinks that the next card is the ace of spades. Which strategy maximizes the probability of a correct guess? This example is taken from Ross (1983).

Exercise 2.7.4 (Casino Game). Imagine you enter a casino and are allowed to play N times the same game. The probability of winning one game is $p \in (0,1)$ and the games are independent. You have an initial wealth $x > 0$ and are allowed to stake any amount in the interval $[0, x]$. When you win, you obtains twice your stake otherwise it is lost. The aim is to maximize the expected wealth $\mathbb{E}_x^\pi[X_N]$ after N games.

a) Set this up as a Markov Decision Model.
b) Find an upper bounding function and show that (SA_N) can be satisfied.
c) Determine an optimal strategy for the cases $p < \frac{1}{2}$, $p = \frac{1}{2}$ and $p > \frac{1}{2}$.
d) What changes if you want to maximize $\mathbb{E}_x^\pi[U(X_N)]$ where $U : \mathbb{R}_+ \to \mathbb{R}$ is a strictly increasing and strictly concave utility function?

Exercise 2.7.5 (LQ-problem). Consider the following special LQ-problem (see Bertsekas (2005) Section 4.1 for more details): The transition function is given by

$$T_n(x,a,z) := A_{n+1}x + B_{n+1}a + z$$

where $x \in \mathbb{R}^m, a \in \mathbb{R}^d$ and A_n, B_n are deterministic matrices of appropriate dimension. The disturbances $Z_1, Z_2, \ldots$ are independent and identically distributed with finite expectation and covariance matrix. The cost

$$\mathbb{E}_x^\pi \left[\sum_{k=0}^{N} X_k^\top Q_k X_k \right]$$

have to be minimized where the Q_n are positive definite.

a) Show that (A_N) and (SA_N) are satisfied.
b) Show that the minimal cost-to-go function is given by

$$V_0(x) = x^\top \tilde{Q}_0 x + \sum_{k=1}^{N} \mathbb{E}[Z_k^\top \tilde{Q}_k Z_k], \quad x \in \mathbb{R}^m$$

where

$$\begin{aligned} \tilde{Q}_N &:= Q_N \\ \tilde{Q}_n &:= Q_n + A_{n+1}^\top \tilde{Q}_{n+1} A_{n+1} \\ &\quad - A_{n+1}^\top \tilde{Q}_{n+1} B_{n+1} \big(B_{n+1}^\top \tilde{Q}_{n+1} B_{n+1} \big)^{-1} B_{n+1}^\top \tilde{Q}_{n+1} A_{n+1} \end{aligned}$$

and the optimal policy $(f_0^*, \ldots, f_{N-1}^*)$ is given by

$$f_n^*(x) = -\Big(B_{n+1}^\top \tilde{Q}_{n+1} B_{n+1} \Big)^{-1} B_{n+1}^\top \tilde{Q}_{n+1} A_{n+1} x.$$

c) Let now $A_k = A$, $B_k = B$ and $Q_k = Q$ for all k and consider the so-called *discrete Riccati equation*

$$\begin{aligned} \tilde{Q}_N &:= Q \\ \tilde{Q}_n &:= Q + A^\top \tilde{Q}_{n+1} A - A^\top \tilde{Q}_{n+1} B \big(B^\top \tilde{Q}_{n+1} B \big)^{-1} B^\top \tilde{Q}_{n+1} A. \end{aligned}$$

Moreover, assume that the matrix

$$[B, AB, A^2B, \ldots, A^{N-1}B]$$

has full rank. Show that there exists a positive definite matrix $\tilde{Q}$ such that $\lim_{n\to\infty} \tilde{Q}_n = \tilde{Q}$. Moreover, $\tilde{Q}$ is the unique solution of

$$\tilde{Q} = Q + A^\top \tilde{Q} A - A^\top \tilde{Q} B \big(B^\top \tilde{Q} B \big)^{-1} B^\top \tilde{Q} A$$

within the class of positive semidefinite matrices.

Remark: The convergence of $\tilde{Q}_k$ in the case of stochastic coefficients is more delicate.

Exercise 2.7.6 (Binary Markov Decision Model). Suppose a stationary Markov Decision Model is given with $A = \{0, 1\}$. Such a model is called a *binary* Markov Decision Model. We suppose that the reward functions r and g are bounded. Define $r(x, 1) = r_1(x)$ and $r(x, 0) = r_0(x)$ for $x \in E$. For $v : E \to \mathbb{R}$ measurable and bounded and $a \in A$ we denote

$$(Q_a v)(x) := \int v(y) Q(dy|x, a), \quad x \in E.$$

a) Show that (A_N) and (SA_N) are satisfied.
b) Show that the value function satisfies

$$J_n = \max\{r_0 + \beta Q_0 J_{n-1}, r_1 + \beta Q_1 J_{n-1}\} =: \max\{L_0 J_{n-1}, L_1 J_{n-1}\}.$$

c) If we denote $d_n(x) = L_1 J_{n-1}(x) - L_0 J_{n-1}(x), n \in \mathbb{N}, x \in E$ show that

$$d_{n+1} = L_1 L_0 J_{n-1} - L_0 L_1 J_{n-1} + \beta Q_1 d_n^+ - \beta Q_0 d_n^- .$$

Exercise 2.7.7 (Replacement Problem). A machine is in use over several periods. The state of the machine is randomly deteriorating and the reward which is obtained depends on the state of the machine. When should the machine be replaced by a new one? The new machine costs a fix amount $K \geq 0$.
We assume that the evolution of the state of the machine is a Markov process with state space $E = \mathbb{R}_+$ and transition kernel Q where $Q([x, \infty)|x) = 1$. A large state x refers to a worse condition/quality of the machine. The reward is $r(x)$ if the state of the machine is x. We assume that the measurable function $r : E \to \mathbb{R}$ is bounded and for the terminal reward $g = r$. In what follows we use the abbreviation

$$(Qv)(x) := \int v(x') Q(dx'|x), \quad (Q_0 v)(x) := \int v(x') Q(dx'|0).$$

Note that $(Q_0 v)(x)$ does not depend on x!

a) Show that (A_N) and (SA_N) are satisfied.
b) Show that the maximal reward operator is given by

$$\mathcal{T} v(x) = r(x) + \max\{\beta(Qv)(x), -K + \beta(Q_0 v)(x)\}, \quad x \geq 0.$$

c) Let $d_n(x) := -K + \beta(Q_0 J_{n-1})(x) - \beta(Q J_{n-1})(x), n \in \mathbb{N}, x \in E$. Show that

$$d_{n+1} = -(1 - \beta)K - \beta Q r - \beta Q d_n^- + c_n$$

where $c_n := \beta Q_0 J_n - \beta^2 Q_0 J_{n-1}$ is independent of $x \in E$.
d) If r is decreasing and the transition kernel Q is stochastically monotone prove that a maximizer f_n^* of J_{n-1} is of the form

$$f_n^*(x) = \begin{cases} \text{replace} & \text{if } x \geq x_n^* \\ \text{not replace} & \text{if } x < x_n^* \end{cases}$$

for $x_n^* \in \mathbb{R}_+$. The value x_n^* is called the threshold or control limit.

Exercise 2.7.8 (Terminating Markov Decision Model). Suppose a stationary Markov Decision Model with $\beta = 1$ is given with the following properties:

- There exists a set $G \subset E$ such that $r(x,a) = 0$ and $Q(\{x\}|x,a) = 1$ for all $x \in G$ and $a \in D(x)$.
- For all $x \in E$ there exists an $N(x) \leq N$ such that $\mathbb{P}_x^\pi(X_{N(x)} \in G) = 1$ for all policies π.

Define $J(x) := J_{N(x)}(x)$ for all $x \in E$. Such a Markov Decision Model is called *terminating*.

a) Show that $J(x) = g(x)$ for $x \in G$ and $J(x) = \mathcal{T}J(x)$ for $x \notin G$.
b) If $f \in F$ satisfies $\mathcal{T}_f J(x) = \mathcal{T}J(x)$ for $x \notin G$ and $f(x) \in D(x)$ arbitrary for $x \in G$, then the stationary policy $(f, f, \ldots, f) \in F^N$ is optimal.
c) Show that the Red-and-Black card game of Section 2.6.1 is a terminating Markov Decision Model.

Exercise 2.7.9. You are leaving your office late in the evening when it suddenly starts raining and you realize that you have lost your umbrella somewhere during the day. The umbrella can only be at a finite number of places, labelled $1, \ldots, m$. The probability that it is at place i is p_i with $\sum_{i=1}^m p_i = 1$. The distance between two places is given by d_{ij} where $i, j \in \{0, 1, \ldots, m\}$ and label 0 is your office. In which sequence do you visit the places in order to minimize the expected length of the journey until you find your umbrella? Set this up as a Markov Decision Problem or as a terminating Markov Decision Model and write a computer program to solve it.

2.8 Remarks and References

In this chapter we consider Markov Decision Models with Borel state and action spaces and unbounded reward functions. In order to increase the readability and to reduce the mathematical framework (e.g. measurability and existence problems) we introduce the Structure Assumption (SA_N) and the notion of an (upper) bounding function in Section 2.4. This framework is very useful for applications in finance (see Chapter 4) where the state spaces are often uncountable subsets of Euclidean spaces and the utility functions are unbounded. A similar approach is also used in Schäl (1990) and Puterman (1994).

Section 2.4: Semi-continuous Markov Decision Models have been investigated e.g. in Bertsekas and Shreve (1978), Dynkin and Yushkevich (1979)

and Hernández-Lerma and Lasserre (1996). Properties of the value functions like *increasing*, *concave*, *convex* and combinations thereof were first rigorously studied in Hinderer (1985). For a recent paper see Smith and McCardle (2002). Moreover, Hinderer (2005) and Müller (1997) discuss Lipschitz-continuity of the value functions. The fact that supermodular functions are important for obtaining monotone maximizers was first discussed by Topkis (1978). Altman and Stidham (1995) investigate so-called *binary* Markov Decision Models with two actions (e.g. in replacement problems) and derive general conditions for the existence of threshold policies. The comparison results for Markov Decision Problems can been found in Müller and Stoyan (2002), see also Bäuerle and Rieder (1997).

Section 2.6: The 'Red-and-Black' card game was presented by Connelly (1974) under the name 'Say red'. It can also be solved by martingale arguments. Other interesting game problems and examples can be found in Ross (1970, 1983). Heyman and Sobel (2004a,b) consider various stochastic optimization problems in Operations Research, in particular cash balance models, inventory and queueing problems. For a recent extension of the cash balance problem, where the cash changes depend on an underlying Markov process, see Hinderer and Waldmann (2001). Stochastic LQ-problems have been investigated by many authors. For a comprehensive treatment see e.g. Bertsekas (2001).

http://www.springer.com/978-3-642-18323-2

Markov Decision Processes with Applications to Finance

Bäuerle, N.; Rieder, U.

2011, XVI, 388 p. 24 illus., Softcover
ISBN: 978-3-642-18323-2