July 06, 2010, 15:00
Draft, don't distribute

# MARKOV DECISION PROCESSES

NICOLE BÄUERLE* AND ULRICH RIEDER‡

**Abstract:** The theory of Markov Decision Processes is the theory of controlled Markov chains. Its origins can be traced back to R. Bellman and L. Shapley in the 1950's. During the decades of the last century this theory has grown dramatically. It has found applications in various areas like e.g. computer science, engineering, operations research, biology and economics. In this article we give a short introduction to parts of this theory. We treat Markov Decision Processes with finite and infinite time horizon where we will restrict the presentation to the so-called (generalized) *negative* case. Solution algorithms like Howard's policy improvement and linear programming are also explained. Various examples show the application of the theory. We treat stochastic linear-quadratic control problems, bandit problems and dividend pay-out problems.

**AMS 2010 Classification: 90C40, 60J05, 93E20**

**Keywords and Phrases: Markov Decision Process, Markov Chain, Bellman Equation, Policy Improvement, Linear Programming**

## 1. Introduction

Do you want to play a card game? Yes? Then I will tell you how it works. We have a well-shuffled standard 32-card deck which is also known as a piquet deck. 16 cards are red and 16 cards are black. Initially the card deck lies on the table face down. Then I start to remove the cards and you are able to see its faces. Once you have to say "stop". If the next card is black you win 10 Euro, if it is red you loose 10 Euro. If you do not say "stop" at all, the color of the last card is deciding. Which stopping rule maximizes your expected reward?
Obviously, when you say "stop" before a card is turned over, your expected reward is

$$\frac{1}{2}\cdot 10 \text{ Euro } + \frac{1}{2}\cdot(-10) \text{ Euro } = 0 \text{ Euro}.$$

The same applies when you wait until the last card due to symmetry reasons. But of course you are able to see the cards' faces when turned over and thus always know how many red and how many black cards are still in the deck. So there may be a clever strategy which gives a higher expected reward than zero. How does it look like?
There are now various methods to tackle this problem. We will solve it with the theory of *Markov Decision Processes*. Loosely speaking this is the theory of controlled

We dedicate this paper to Karl Hinderer who passed away on April 17th, 2010. He established the theory of Markov Decision Processes in Germany 40 years ago.

Markov chains. In the general theory a system is given which can be controlled by sequential decisions. The state transitions are random and we assume that the *system state process* is *Markovian* which means that previous states have no influence on future states. In the card game the state of the system is the number of red and black cards which are still in the deck. Given the current state of the system, the controller or decision maker has to choose an *admissible action* (in the card game say "stop" or "go ahead"). Once an action is chosen there is a random system transition according to a stochastic law (removing of next card which either is black or red) which leads to a new state and the controller receives a reward. The task is to control the process such that the expected total (discounted) rewards are maximized.

We will see that problems like this can be solved recursively. When we return to the card game for example it is quite easy to figure out the optimal strategy when there are only 2 cards left in the stack. Knowing the value of the game with 2 cards it can be computed for 3 cards just by considering the two possible actions "stop" and "go ahead" for the next decision. We will see how this formally works in Section 2.3.1.

First books on Markov Decision Processes are Bellman (1957) and Howard (1960). The term 'Markov Decision Process' has been coined by Bellman (1954). Shapley (1953) was the first study of Markov Decision Processes in the context of stochastic games. For more information on the origins of this research area see Puterman (1994). Mathematical rigorous treatments of this optimization theory appeared in Dubins and Savage (1965), Blackwell (1965), Shiryaev (1967), Hinderer (1970), Bertsekas and Shreve (1978) and Dynkin and Yushkevich (1979). More recent textbooks on this topic are Schäl (1990), Puterman (1994), Hernández-Lerma and Lasserre (1996), Bertsekas (2001, 2005), Feinberg and Shwartz (2002), Powell (2007) and Bäuerle and Rieder (2011).

This article is organized as follows: In the next section we introduce Markov Decision Processes with finite time horizon. We show how they can be solved and consider as an example so-called stochastic linear-quadratic control problems. The solution of the card game is also presented. In Section 3 we investigate Markov Decision Processes with infinite time horizon. These models are on the one hand more complicated than the problems with finite time horizon since additional convergence assumptions have to be satisfied, on the other hand the solution is often simpler because the optimal strategy is stationary and the value function can be characterized as the largest $r$-subharmonic function or as the unique fixed point of the maximal reward operator. Here we will restrict the presentation to the so-called (generalized) negative case. Besides some main theorems which characterize the optimal solution we will also formulate two solution techniques, namely Howard's policy improvement and linear programming. As applications we consider a dividend pay-out problem and bandit problems. Further topics on Markov Decision Processes are discussed in the last section. For proofs we refer the reader to the forthcoming book of Bäuerle and Rieder (2011).

## 2. Markov Decision Processes with Finite Time Horizon

In this section we consider Markov Decision Models with a finite time horizon. These models are given by a state space for the system, an action space where the actions can be taken from, a stochastic transition law and reward functions. Hence

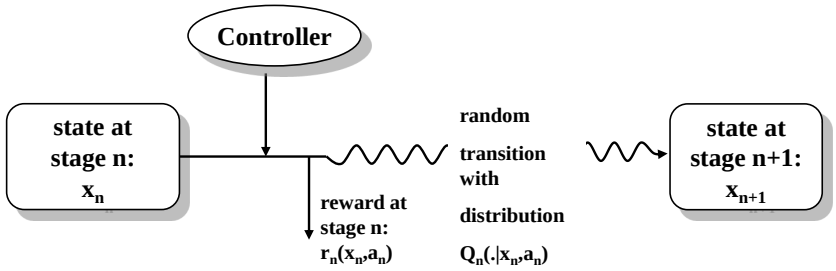


FIGURE 1. General evolution of a Markov Decision Model.

a (non-stationary) *Markov Decision Model* with 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 $\sigma$-algebra $\mathfrak{E}$. The elements (states) are denoted by $x \in E$.
- $A$ is the *action space*, endowed with a $\sigma$-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 admissible state-action pairs at time $n$. In order to have a well-defined problem 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*. If $E$ is discrete we write $q_n(x'|x, a) := Q_n(\{x'\}|x, a)$.
- $r_n : D_n \to \mathbb{R}$ 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 $x$.

Next we introduce the notion of a strategy. Since the system is stochastic, a strategy has to determine actions for every possible state of the system and for every time point. 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 *decision rule* at time $n$. We denote by $F_n$ the set of all decision rules at time $n$. A sequence of decision rules $\pi = (f_0, f_1, \ldots, f_{N-1})$ with $f_n \in F_n$ is called *$N$-stage policy* or *$N$-stage strategy*. If a decision maker follows a policy $\pi = (f_0, f_1, \ldots, f_{N-1})$ and observes at time $n$ the state $x$ of the system, then the action she chooses is $f_n(x)$. This means in particular that the decision at time $n$ depends only on the system state at time $n$. Indeed the decision maker could

also base her decision on the whole history $(x_0, a_0, x_1 \ldots, a_{n-1}, x_n)$. But due to the Markovian character of the problem it can be shown that the optimal policy (which is defined below) is among the smaller class of so called *Markovian policies* we use here.

We consider a Markov Decision Model as an $N$-stage random experiment. The underlying probability space is given by the *canonical construction* 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 $\omega$. 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 $\pi = (f_0, f_1, \ldots, f_{N-1})$ is a fixed policy and $x \in E$ is a fixed initial state. There exists a unique probability measure $\mathbb{P}_x^\pi$ on $(\Omega, \mathcal{F})$ with

$\mathbb{P}_x^\pi(X_0 \in B) = \varepsilon_x(B)$ for all $B \in \mathfrak{E}$,

$\mathbb{P}_x^\pi(X_{n+1} \in B | X_1, \ldots, X_n) = \mathbb{P}_x^\pi(X_{n+1} \in B | X_n) = Q_n\big(B | X_n, f_n(X_n)\big)$,

where $\varepsilon_x$ is the one-point measure concentrated in $x$. The second equation is the so-called *Markov property*, i.e. the sequence of random variables $X_0, X_1, \ldots, X_n$ 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.

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 $n = 0, 1, \ldots, N$*

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

We assume that ($\mathrm{A}_N$) holds for the $N$-stage Markov Decision Problems throughout this section. Obviously Assumption ($\mathrm{A}_N$) is satisfied if all $r_n$ and $g_N$ are bounded from above. We can now introduce the expected discounted reward of a policy and the $N$-stage optimization problem. For $n = 0, 1, \ldots, N$ and a policy $\pi = (f_0, \ldots, f_{N-1})$ let $V_{n\pi}(x)$ be defined by

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

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

$$V_n(x) := \sup_\pi V_{n\pi}(x), \quad x \in E, \tag{2.1}$$

and gives 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 (measure) theoretic inconveniences. Some further assumptions are needed to imply this. 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 $x \in E$.

2.1. **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*. First we introduce some important operators which simplify the notation. In what follows let us denote

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

Due to our assumptions we have $V_{n\pi} \in \mathbb{M}(E)$ for all $\pi$ and $n$.
We define the following operators for $n = 0, 1, \ldots, N-1$ and $v \in \mathbb{M}(E)$:

$$\begin{aligned}
(L_n v)(x,a) &:= r_n(x,a) + \int v(x') Q_n(dx'|x,a), \quad (x,a) \in D_n, \\
(\mathcal{T}_{nf} v)(x) &:= (L_n v)(x, f(x)), \quad x \in E,\ f \in F_n, \\
(\mathcal{T}_n v)(x) &:= \sup_{a \in D_n(x)} (L_n v)(x,a), \quad x \in E
\end{aligned}$$

whenever the integrals exist. $\mathcal{T}_n$ is called the *maximal reward operator at time n*. The operators $\mathcal{T}_{nf}$ can now be used to compute the value of a policy recursively.

**Theorem 2.1** (**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) $V_{n\pi} = T_{nf_n} \ldots T_{N-1 f_{N-1}} g_N$.

For the solution of Markov Decision Problems the following notion will be important.

**Definition 2.2.** 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 $a \mapsto (L_n v)(x,a),\ a \in D_n(x)$.

Below we will see that Markov Decision Problems 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). Here we use a different approach and state at first the following verification theorem. The proof is by recursion.

**Theorem 2.3** (**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 $(f_0^*, f_1^*, \ldots, f_{N-1}^*)$ is optimal for the $N$-stage Markov Decision Problem.*

Theorem 2.3 states that whenever we have a solution of the Bellman equation, together with the maximizers, then we have found a solution of the Markov Decision Problem. Next we consider a general approach to Markov Decision Problems under the following structure assumption. An important case where this assumption is satisfied is given in Section 2.2.

**Structure Assumption ($\mathbf{SA}_N$):** *There exist sets $\mathbb{M}_n \subset \mathbb{M}(E)$ of measurable functions and sets $\Delta_n \subset F_n$ of decision rules 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 shows how Markov Decision Problems can be solved recursively by solving $N$ (one-stage) optimization problems.

**Theorem 2.4 (Structure Theorem).** *Let ($SA_N$) be satisfied. Then it holds:*

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

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

*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 exists a maximizer $f_n$ 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 ($\mathrm{SA}_N$) (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)

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

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

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

where we use the fact that $\mathcal{T}_{nf_n}$ is order preserving, i.e. $v \leq w$ implies $\mathcal{T}_{nf_n} v \leq \mathcal{T}_{nf_n} w$. 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 ($\mathrm{SA}_N$), $V_n \in \mathbb{M}_n$. □

2.2. **Semicontinuous Markov Decision Processes.** In this section we give sufficient conditions under which assumptions ($\mathrm{A}_N$) and ($\mathrm{SA}_N$) are satisfied and thus imply the validity of the Bellman equation and the existence of optimal policies. The simplest case arises when state and action spaces are finite in which case ($\mathrm{A}_N$) is obviously satisfied and ($\mathrm{SA}_N$) is satisfied with $\mathbb{M}_n$ and $\Delta_n$ being the set of all functions $v : E \to [-\infty, \infty)$ and $f : S \to A$ respectively. We assume now that $E$ and $A$ are Borel spaces, i.e. Borel subsets of Polish spaces (i.e. complete, separable, metric spaces). Also $D_n$ is assumed to be a Borel subset of $E \times A$. Let us first consider the Integrability Assumption ($\mathrm{A}_N$). It is fulfilled when the Markov Decision Model has a so-called upper bounding function.

**Definition 2.5.** 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 $(x,a) \in D_n$.

When an upper bounding function exists we denote in the sequel

$$\alpha_b := \sup_{(x,a)\in D} \frac{\int b(x')Q(dx'|x,a)}{b(x)}$$

(with the convention $\frac{0}{0} := 0$). If $r_n$ and $g_N$ are bounded from above, then obviously $b \equiv 1$ is an upper bounding function. For $v \in \mathbb{M}(E)$ we define the *weighted supremum norm* by

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

and introduce the set

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

The next result is fundamental for many applications.

**Proposition 2.6.** *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.*

In order to satisfy ($\mathrm{SA}_N$) we consider so-called semicontinuous models. In the next definition $M$ is supposed to be a Borel space.

**Definition 2.7.** a) A function $v : M \to \bar{\mathbb{R}}$ is called *upper semicontinuous* if for all sequences $(x_n) \subset M$ with $\lim_{n\to\infty} x_n = x \in M$ it holds

$$\limsup_{n\to\infty} v(x_n) \le v(x).$$

b) The set-valued mapping $x \mapsto D(x)$ is called *upper semicontinuous* if it has the following property for all $x \in E$: If $x_n \to x$ and $a_n \in D(x_n)$ for all $n \in \mathbb{N}$, then $(a_n)$ has an accumulation point in $D(x)$.

The next theorem presents easy to check conditions which imply ($\mathrm{SA}_N$).

**Theorem 2.8.** *Suppose a Markov Decision Model with an 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 upper semicontinuous,*

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

*Then the sets* $\mathbb{M}_n := \{v \in \mathbb{M}(E) \mid v^+ \in \mathbb{B}_b, v$ *is upper semicontinuous*$\}$ *and* $\Delta_n := F_n$ *satisfy the Structure Assumption (SA*$_N$*).*

Of course, it is possible to give further conditions which imply (SA$_N$), e.g. other continuity and compactness conditions, monotonicity conditions, concavity or convexity conditions (see Bäuerle and Rieder (2011), Chapter 2).

2.3. **Applications of Finite-Stage Markov Decision Processes.** In this section we present the solution of the card game and investigate stochastic linear-quadratic control problems. Both examples illustrate the solution method for finite-stage Markov Decision Processes.

2.3.1. **Red-and-Black Card-Game.** Let us first reconsider the card game of the introduction. The state of the system is the number of cards which are still uncovered, thus

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

and $N = r_0 + b_0$ where $r_0$ and $b_0$ are the total number of red and black cards in the deck. The state $(0,0)$ will be absorbing. For $x \in E$ and $x \notin \{(0,1),(1,0)\}$ we have $D_n(x) = A = \{0,1\}$ 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 $D_{N-1}\big((0,1)\big) = D_{N-1}\big((1,0)\big) = \{1\}$. The transition probabilities are given by

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

The one-stage reward is given by the expected reward

$$r_n\big((b,r),1\big) := \frac{b-r}{b+r} \quad \text{for } (b,r) \in E \setminus \{(0,0)\},$$

and the reward is zero otherwise. Finally we define

$$g_N(b,r) := \frac{b-r}{b+r} \quad \text{for } (b,r) \in E \setminus \{(0,0)\}$$

and $g_N((0,0)) = 0$. Since $E$ and $A$ are finite, (A$_N$) and also the Structure Assumption (SA$_N$) is clearly satisfied with

$$\mathbb{M}_n = \mathbb{M} := \{v : E \to \mathbb{R} \mid v(0,0) = 0\} \quad \text{and} \quad \Delta := F.$$

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

$$\begin{aligned}
(\mathcal{T}_n v)(b,r) &:= \max\left\{\frac{b-r}{b+r}\ ,\ \frac{r}{r+b}v(r-1,b)+\frac{b}{r+b}v(r,b-1)\right\} \quad \text{for } b+r\geq 2,\\
(\mathcal{T}_{N-1}v)(1,0) &:= 1,\\
(\mathcal{T}_{N-1}v)(0,1) &:= -1,\\
(\mathcal{T}_n v)(0,0) &:= 0.
\end{aligned}$$

It is not difficult to see that $g_N = \mathcal{T}_n g_N$ for $n = 0, 1, \ldots, N-1$. For $x = (b,r) \in E$ with $r + b \geq 2$ the computation is as follows:

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

Since both expressions for $a = 0$ and $a = 1$ are identical, every $f \in F$ is a maximizer of $g_N$. Applying Theorem 2.4 we obtain that $V_n = \mathcal{T}_n \ldots \mathcal{T}_{N-1} g_N = g_N$ and we can formulate the solution of the card game.

**Theorem 2.9.** *The maximal value of the card game is given by*

$$V_0(b_0, r_0) = g_N(b_0, r_0) = \frac{b_0 - r_0}{b_0 + r_0},$$

*and every strategy is optimal.*

Thus, there is no strategy which yields a higher expected reward than the trivial ones discussed in the introduction. The game is fair (i.e. $V_0(b_0, r_0) = 0$) if and only if $r_0 = b_0$. Note that the card game is a stopping problem. The theory of optimal stopping problems can be found e.g. in Peskir and Shiryaev (2006). For more gambling problems see Ross (1983).

2.3.2. **Stochastic Linear-Quadratic Control Problems.** A famous class of control problems with 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 is 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 is given by

$$X_{n+1} := A_{n+1}X_n + B_{n+1}f_n(X_n).$$

We suppose that the random matrices $(A_1, B_1), (A_2, B_2), \ldots$ are independent but not necessarily identically distributed and have finite expectation and covariance. Thus, the law of $X_{n+1}$ is given by the kernel

$$Q_n(B|x,a) := \mathbb{P}\left((A_{n+1}x + B_{n+1}a) \in B\right), \quad B \in \mathcal{B}(\mathbb{R}^m).$$

Moreover, we assume that $\mathbb{E}\left[B_{n+1}^\top R B_{n+1}\right]$ is positive definite for all symmetric positive definite matrices $R$. The one-stage reward is a negative cost function

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

and the terminal reward is

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

with deterministic, symmetric and positive definite matrices $R_0, R_1, \ldots, R_N$. There is no discounting. The aim is to minimize

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

over all $N$-stage policies $\pi$. Thus, the aim is to minimize the expected quadratic distance of the state process to the benchmark zero.

We have $r_n \le 0$ and $b \equiv 1$ is an upper bounding function, thus $(\mathrm{A}_N)$ 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. The minimal cost operator is given by

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

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

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

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 assumption $(\mathrm{SA}_N)$(i): Obviously $x^\top R_N x \in \mathbb{M}_N$. Now let $v(x) = x^\top R x \in \mathbb{M}_{n+1}$. We try to solve the following optimization problem

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

Since $R$ is positive definite, we have by assumption that $\mathbb{E}\left[B_{n+1}^\top R B_{n+1}\right]$ 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

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

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

$$\begin{aligned} \mathcal{T}_n v(x) &= x^\top \Big( R_n + \mathbb{E}\left[A_{n+1}^\top R A_{n+1}\right] - \mathbb{E}\left[A_{n+1}^\top R B_{n+1}\right] \big( \mathbb{E}\left[B_{n+1}^\top R B_{n+1}\right] \big)^{-1} \\ &\qquad \mathbb{E}\left[B_{n+1}^\top R A_{n+1}\right] \Big) x = x^\top \tilde{R} x \end{aligned}$$

where $\tilde{R}$ is defined as the expression in the brackets. Note that $\tilde{R}$ is symmetric and since $x'\tilde{R}x = \mathcal{T}_n v(x) \ge x^\top R_n x$, it is also positive definite. Thus $\mathcal{T} v \in \mathbb{M}_n$ and the Structure Assumption $(\mathrm{SA}_N)$ is satisfied for $\mathbb{M}_n$ and $\Delta_n = \Delta \cap F_n$. Now we can apply Theorem 2.4 to solve the stochastic linear-quadratic control problem.

**Theorem 2.10.** a) *Let the matrices $\tilde{R}_n$ be recursively defined by*

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

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

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

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

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. If the state of the system cannot be observed completely the decision rule is still linear in the state but here the coefficient matrix has to be estimated recursively. This follows from the principle of estimation and control.
Our formulation of the stochastic LQ-problem can be generalized in different ways without leaving the LQ-framework (see e.g. Bertsekas (2001, 2005)). For example the cost function can be extended to

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

where $\hat{R}_k$ are deterministic, symmetric positive semidefinite matrices and $b_k$ are deterministic vectors. In this formulation the control itself is penalized and the expected distance of the state process to the benchmarks $b_k$ has to be kept small.

2.3.3. **Further Applications.** Applications of Markov Decision Processes can be found in stochastic operations research, engineering, computer science, logistics and economics (see e.g. Stokey and Lucas (1989), Bertsekas (2001, 2005), Tijms (2003), Meyn (2008), Bäuerle and Rieder (2011)). Prominent examples are inventory-production control, control of queues (controls can be routing, scheduling), portfolio optimization (utility maximization, index-tracking, indifference pricing, Mean-Variance problems), pricing of American options and resource allocation problems (resources could be manpower, computer capacity, energy, money, water etc.). Recent practical applications are e.g. given in Goto et al. (2004) (Logistics), Enders et al. (2010) (Energy systems) and He et al. (2010) (Health care). Research areas which are closely related to Markov Decision Processes are optimal stopping and multistage (dynamic) game theory.
Markov Decision Problems also arise when continuous-time stochastic control problems are discretized. This numerical procedure is known under the name *approximating Markov chain approach* and is discussed e.g. in Kushner and Dupuis (2001). Stochastic control problems in continuous-time are similar to the theory explained here, however require a quite different mathematical background. There the Bellman equation is replaced by the so-called Hamilton-Jacobi-Bellman equation and tools from stochastic analysis are necessary. Continuous-time Markov Decision Processes are treated in Guo and Hernández-Lerma (2009).

## 3. Markov Decision Processes with Infinite Time Horizon

In this chapter we consider Markov Decision Models with an infinite time horizon. There are situations where problems with infinite time horizon arise in a natural way, e.g. when the random lifetime of a stochastic system is considered. However more important is the fact that Markov Decision Models with finite but large horizon can be approximated by models with infinite time horizon. In what follows we always assume that a stationary Markov Decision Model with infinite horizon is given, i.e. the data does not depend on the time parameter $n$ and we thus have a state space $E$, an action space $A$, a set of admissible state-action pairs $D$, a transition kernel $Q$, a one-stage reward $r$ and a discount factor $\beta \in (0,1]$. By $F$ we denote the set of all decision rules, i.e. measurable functions $f : E \to A$ with $f(x) \in D(x)$ for all $x \in E$.
Let $\pi = (f_0, f_1, \ldots) \in F^\infty$ be a policy for the infinite-stage Markov Decision Model. Then we define

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

which gives the *expected discounted reward* under policy $\pi$ (over an infinite time horizon) when we start in state $x$. The performance criterion is then

$$J_\infty(x) := \sup_\pi J_{\infty\pi}(x), \quad x \in E. \tag{3.1}$$

The function $J_\infty(x)$ gives the *maximal expected discounted reward* (over an infinite time horizon) when we start in state $x$. A policy $\pi^* \in F^\infty$ is called *optimal* if $J_{\infty\pi^*}(x) = J_\infty(x)$ for all $x \in E$. In order to have a well-defined problem we assume

**Integrability Assumption (A):**

$$\delta(x) := \sup_\pi \mathbb{E}_x^\pi \left[ \sum_{k=0}^{\infty} \beta^k r^+\big(X_k, f_k(X_k)\big) \right] < \infty, \quad x \in E.$$

In this stationary setting the operators of the previous section read

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

When we now define for $n \in \mathbb{N}_0$

$$\begin{aligned}
J_{n\pi}(x) &:= \mathcal{T}_{f_0} \ldots \mathcal{T}_{f_{n-1}} 0(x), \quad \pi \in F^\infty \\
J_n(x) &:= \mathcal{T}^n 0(x),
\end{aligned}$$

then the interpretation of $J_n(x)$ is the maximal expected discounted reward over $n$ stages when we start in state $x$ and the terminal reward function is zero, i.e. it holds

$$\begin{aligned}
J_{n\pi}(x) &= \mathbb{E}_x^\pi \Big[ \sum_{k=0}^{n-1} \beta^k r\big(X_k, f_k(X_k)\big) \Big] \\
J_n(x) &= \sup_\pi J_{n\pi}(x), \quad x \in E.
\end{aligned}$$

Moreover, it is convenient to introduce the set

$$\mathbb{B} := \{v \in \mathbb{M}(E) \mid v(x) \leq \delta(x) \text{ for all } x \in E\}.$$

Obviously, we have $J_{\infty\pi} \in \mathbb{B}$ for all policies $\pi$. In order to guarantee that the infinite horizon problem is an approximation of the finite horizon model, we use the following convergence assumption.

**Convergence Assumption (C):**

$$\lim_{n\to\infty} \sup_{\pi} \mathbb{E}_x^{\pi} \left[ \sum_{k=n}^{\infty} \beta^k r^+\big(X_k, f_k(X_k)\big) \right] = 0, \quad x \in E.$$

When assumptions (A) and (C) are satisfied we speak of the so-called (generalized) *negative* case. It is fulfilled e.g. if there exists an upper bounding function $b$ and $\beta\alpha_b < 1$. In particular if $r \leq 0$ or $r$ is bounded from above and $\beta \in (0,1)$. The Convergence Assumption (C) implies that $\lim_{n\to\infty} J_{n\pi}$ and $\lim_{n\to\infty} J_n$ exist. Moreover, for $\pi \in F^\infty$ we obtain

$$J_{\infty\pi} \quad = \quad \lim_{n\to\infty} J_{n\pi}.$$

Next we define the *limit value function* by

$$J(x) \quad := \quad \lim_{n\to\infty} J_n(x) \leq \delta(x), \quad x \in E.$$

By definition it obviously holds that $J_{n\pi} \leq J_n$ for all $n \in \mathbb{N}$, hence $J_{\infty\pi} \leq J$ for all policies $\pi$. Taking the supremum over all $\pi$ implies

$$J_\infty(x) \leq J(x), \quad x \in E.$$

The next example shows that in general $J \neq J_\infty$.

**Example 3.1.** We consider the following Markov Decision Model: Suppose that the state space is $E := \mathbb{N}$ and the action space is $A := \mathbb{N}$. Further let $D(1) := \{3, 4, \ldots\}$ and $D(x) := A$ for $x \geq 2$ be the admissible actions. The transition probabilities are given by

$$\begin{aligned} q(a|1,a) &:= 1, \\ q(2|2,a) &:= 1, \\ q(x-1|x,a) &:= 1 \quad \text{for } x \geq 3. \end{aligned}$$

All other transition probabilities are zero (cf. Figure 2). Note that state 2 is an absorbing state. The discount factor is $\beta = 1$ and the one-stage reward function is given by

$$r(x,a) \quad := \quad -\delta_{x3}, \quad (x,a) \in D.$$

Since the reward is non-positive, assumptions (A) and (C) are satisfied.
We will compute now $J$ and $J_\infty$. Since state 2 is absorbing, we obviously have $J_\infty(2) = 0$ and $J_\infty(x) = -1$ for $x \neq 2$. On the other hand we obtain for $n \in \mathbb{N}$ that

$$J_n(x) = \begin{cases} 0 & , \text{ for } x = 1, 2 \\ -1 & , \text{ for } 3 \leq x \leq n+2 \\ 0 & , \text{ for } x > n+2. \end{cases}$$

Thus, $J_\infty(1) = -1 \neq 0 = J(1) = \lim_{n\to\infty} J_n(1)$.

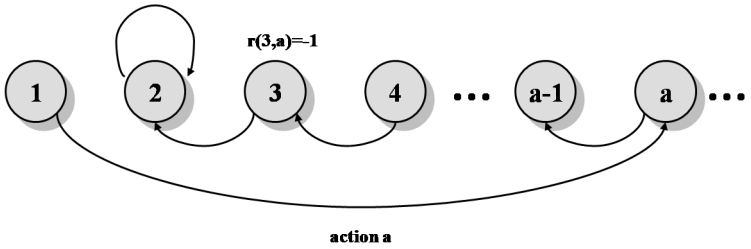


FIGURE 2. Transition diagram of Example 3.1.

As in the finite horizon model the following reward iteration holds where $J_f := J_{\infty(f,f,\ldots)}$ for a stationary policy $(f,f,\ldots)$.

**Theorem 3.2 (Reward Iteration).** *Assume (C) and let $\pi = (f,\sigma) \in F \times F^\infty$. Then it holds:*

a) $J_{\infty\pi} = \mathcal{T}_f J_{\infty\sigma}$.
b) $J_f \in \mathbb{B}$ *and* $J_f = \mathcal{T}_f J_f$.

The functions $J_n, J$ and $J_\infty$ are in general not in $\mathbb{B}$. However, $J_\infty$ and $J$ are analytically measurable and satisfy

$$J_\infty = \mathcal{T} J_\infty \quad \text{and} \quad J \geq \mathcal{T} J,$$

see e.g. Bertsekas and Shreve (1978). As in Section 2 we formulate here a verification theorem in order to avoid the general measurability problems.

**Theorem 3.3 (Verification Theorem).** *Assume (C) and let $v \in \mathbb{B}$ be a fixed point of $\mathcal{T}$ such that $v \geq J_\infty$. If $f^*$ is a maximizer of $v$, then $v = J_\infty$ and the stationary policy $(f^*, f^*, \ldots)$ is optimal for the infinite-stage Markov Decision Problem.*

Natural candidates for a fixed point of $\mathcal{T}$ are the functions $J_\infty$ and $J$. In what follows we want to solve the optimization problem (3.1) and at the same time we would like to have $J_\infty = J$. In order to obtain this statement we require the following structure assumption.

**Structure Assumption (SA):** *There exists a set $\mathbb{M} \subset \mathbb{M}(E)$ of measurable functions and a set $\Delta \subset F$ of decision rules such that:*

(i) $0 \in \mathbb{M}$.
(ii) *If $v \in \mathbb{M}$ then*

$$(\mathcal{T} v)(x) := \sup_{a \in D(x)} \left\{ r(x,a) + \beta \int v(x') Q(dx'|x,a) \right\}, \quad x \in E$$

*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 $v$.*
(iv) $J \in \mathbb{M}$ *and* $J = \mathcal{T} J$.

Note that conditions (i)-(iii) together constitute the Structure Assumption of Section 2 in a stationary model with $g_N \equiv 0$. Condition (iv) imposes additional properties on the limit value function.

**Theorem 3.4** (**Structure Theorem**). *Let (C) and (SA) be satisfied. Then it holds:*

*a)* $J_\infty \in \mathbb{M}$, $J_\infty = \mathcal{T} J_\infty$ *and* $J_\infty = J = \lim_{n\to\infty} J_n$.
*b)* $J_\infty$ *is the largest* $r$-subharmonic function $v$ *in* $\mathbb{M} \cap \mathbb{B}$, *i.e* $J_\infty$ *is the largest function* $v$ *in* $\mathbb{M}$ *with* $v \le \mathcal{T} v$ *and* $v \le \delta$.
*c) There exists a maximizer* $f \in \Delta$ *of* $J_\infty$, *and every maximizer* $f^*$ *of* $J_\infty$ *defines an optimal stationary policy* $(f^*, f^*, \ldots)$ *for the infinite-stage Markov Decision Model.*

The equation $J_\infty = \mathcal{T} J_\infty$ is called *Bellman equation* for the infinite-stage Markov Decision Model. Often this fixed point equation is also called *optimality equation*. Part a) of the preceding theorem shows that $J_\infty$ is approximated by $J_n$ for $n$ large, i.e. the value of the infinite horizon Markov Decision Problem can be obtained by iterating the $\mathcal{T}$-operator. This procedure is called *value iteration*. Part c) shows that an optimal policy can be found among the stationary ones.
As in the case of a finite horizon it is possible to give conditions on the model data under which (SA) and (C) are satisfied. We restrict here to one set of continuity and compactness conditions.
In what follows let $E$ and $A$ be Borel spaces, let $D$ be a Borel subset of $E \times A$ and define

$$D_n^*(x) := \{a \in D(x) \mid a \text{ is a maximum point of } a \mapsto LJ_{n-1}(x,a)\}$$

for $n \in \mathbb{N} \cup \{\infty\}$ and $x \in E$ and

$$\begin{aligned} LsD_n^*(x) \quad := \quad & \{a \in A \mid a \text{ is an accumulation point of a sequence } (a_n) \text{ with} \\ & a_n \in D_n^*(x) \text{ for } n \in \mathbb{N}\}, \end{aligned}$$

the so-called *upper limit of the set sequence* $(D_n^*(x))$.

**Theorem 3.5.** *Suppose there exists an upper bounding function* $b$ *with* $\beta\alpha_b < 1$ *and it holds:*

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

*Then it holds:*

a) $J_\infty = \mathcal{T} J_\infty$ *and* $J_\infty = \lim_{n\to\infty} J_n$. **(Value Iteration)**.
b) *If* $b$ *is upper semicontinuous then* $J_\infty$ *is upper semicontinuous.*
c) $\emptyset \neq LsD_n^*(x) \subset D_\infty^*(x)$ *for all* $x \in E$. **(Policy Iteration)**.
d) *There exists an* $f^* \in F$ *with* $f^*(x) \in LsD_n^*(x)$ *for all* $x \in E$, *and the stationary policy* $(f^*, f^*, \ldots)$ *is optimal.*

Suppose the assumptions of Theorem 3.5 are satisfied and the optimal stationary policy $f^\infty$ is unique, i.e. we obtain $D_\infty^*(x) = \{f(x)\}$. Now suppose $(f_n^*)$ is a sequence of decision rules where $f_n^*$ is a maximizer of $J_{n-1}$. According to part c) we must have $\lim_{n\to\infty} f_n^* = f$. This means that we can approximate the *optimal policy* for the infinite horizon Markov Decision Problem by a sequence of optimal policies for the finite-stage problems. This property is called *policy iteration*.

**Remark 3.6.** If we define

$$\varepsilon_n(x) := \sup_\pi \mathbb{E}_x^\pi \Big[ \sum_{k=n}^\infty \beta^k r^-\big(X_k, f_k(X_k)\big)\Big], \quad x \in E,$$

where $x^- = \max\{0, -x\}$ denotes the negative part of $x$, then instead of (A) and (C) one could require $\varepsilon_0(x) < \infty$ and $\lim_{n\to\infty} \varepsilon_n(x) = 0$ for all $x \in E$. In this case we speak of a (generalized) *positive* Markov Decision Model. This type of optimization problem is not dual to the problems we have discussed so far. In particular, the identification of optimal policies is completely different (see e.g. Bertsekas and Shreve (1978), Schäl (1990)).

3.1. **Contracting Markov Decision Processes.** An advantageous and important situation arises when the operator $\mathcal{T}$ is contracting. To explain this we assume that the Markov Decision Model has a so-called *bounding function* (instead of an upper bounding function which we have considered so far).

**Definition 3.7.** A measurable function $b : E \to \mathbb{R}_+$ is called a *bounding function* for the Markov Decision Model if there exist constants $c_r, \alpha_b \in \mathbb{R}_+$, such that

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

Markov Decision Models with a bounding function $b$ and $\beta\alpha_b < 1$ are called *contracting*. We will see in Lemma 3.8 that $\beta\alpha_b$ is the module of the operator $\mathcal{T}$.
If $r$ is bounded, then $b \equiv 1$ is a bounding function. If moreover $\beta < 1$, then the Markov Decision Model is contracting (the classical *discounted case*). For any contracting Markov Decision Model the assumptions (A) and (C) are satisfied, since $\delta \in I\!B_b$ and there exists a constant $c > 0$ with

$$\lim_{n\to\infty} \sup_\pi \mathbb{E}_x^\pi \left[ \sum_{k=n}^\infty \beta^k r^+\big(X_k, f_k(X_k)\big)\right] \le c \lim_{n\to\infty} (\beta\alpha_b)^n b(x) = 0.$$

**Lemma 3.8.** *Suppose the Markov Decision Model has a bounding function $b$ and let $f \in F$.*

a) *For $v, w \in I\!B_b$ it holds:*

$$\begin{aligned} \|\mathcal{T}_f v - \mathcal{T}_f w\|_b &\le \beta\alpha_b \|v - w\|_b \\ \|\mathcal{T} v - \mathcal{T} w\|_b &\le \beta\alpha_b \|v - w\|_b. \end{aligned}$$

b) *Let $\beta\alpha_b < 1$. Then $J_f = \lim_{n\to\infty} \mathcal{T}_f^n v$ for all $v \in I\!B_b$, and $J_f$ is the unique fixed point of $\mathcal{T}_f$ in $I\!B_b$.*

**Theorem 3.9 (Verification Theorem ).** *Let $b$ be a bounding function, $\beta\alpha_b < 1$ and let $v \in I\!B_b$ be a fixed point of $\mathcal{T} : I\!B_b \to I\!B_b$. If $f^*$ is a maximizer of $v$, then $v = J_\infty = J$ and $(f^*, f^*, \ldots)$ is an optimal stationary policy.*

The next theorem is the main result for contracting Markov Decision Processes. It is a conclusion from Banach's fixed point theorem. Recall that $(I\!B_b, \|\cdot\|_b)$ is a Banach space.

**Theorem 3.10 (Structure Theorem).** *Let $b$ be a bounding function and $\beta\alpha_b < 1$. If there exists a closed subset $I\!M \subset I\!B_b$ and a set $\Delta \subset F$ such that*

(i) $0 \in I\!M$,
(ii) $\mathcal{T} : I\!M \to I\!M$,

(iii) *for all $v \in \mathbb{M}$ there exists a maximizer $f \in \Delta$ of $v$,*

*then it holds:*

a) *$J_\infty \in \mathbb{M}$, $J_\infty = \mathcal{T} J_\infty$ and $J_\infty = \lim_{n\to\infty} J_n$.*
b) *$J_\infty$ is the unique fixed point of $\mathcal{T}$ in $\mathbb{M}$.*
c) *$J_\infty$ is the smallest $r$-superharmonic function $v \in \mathbb{M}$, i.e. $J_\infty$ is the smallest function $v \in \mathbb{M}$ with $v \geq \mathcal{T} v$.*
d) *Let $v \in \mathbb{M}$. Then*

$$\|J_\infty - \mathcal{T}^n v\|_b \leq \frac{\left(\beta\alpha_b\right)^n}{1-\beta\alpha_b} \|\mathcal{T} v - v\|_b.$$

e) *There exists a maximizer $f \in \Delta$ of $J_\infty$, and every maximizer $f^*$ of $J_\infty$ defines an optimal stationary policy $(f^*, f^*, \ldots)$.*

3.2. **Applications of Infinite-Stage Markov Decision Processes.** In this subsection we consider bandit problems and dividend pay-out problems. Applications to finance are investigated in Bäuerle and Rieder (2011). In particular, optimization problems with random horizon can be solved via infinite-stage Markov Decision Processes.

3.2.1. **Bandit Problems.** An important application of Markov Decision Problems are so-called *bandit problems*. We will restrict here to Bernoulli bandits with two-arms. The game is as follows: Imagine we have two slot machines with unknown success probability $\theta_1$ and $\theta_2$. The success probabilities are chosen independently from two prior Beta-distributions. At each stage we have to choose one of the arms. We receive one Euro if the arm wins, else no cash flow appears. The aim is to maximize the expected discounted reward over an infinite number of trials. One of the first (and more serious) applications is to medical trials of a new drug. In the beginning the cure rate of the new drug is not known and may be in competition to well-established drugs with known cure rate (this corresponds to one bandit with known success probability). The problem is not trivial since it is not necessarily optimal to choose the arm with the higher expected success probability. Instead one has to incorporate 'learning effects' which means that sometimes one has to pull one arm just to get some information about its success probability. It is possible to prove the optimality of a so-called *index-policy*, a result which has been generalized further for multi-armed bandits.
The bandit problem can be formulated as a Markov Decision Model as follows. The state is given by the number of successes $m_a$ and failures $n_a$ at both arms $a = 1, 2$ which have appeared so far. Hence $x = (m_1, n_1, m_2, n_2) \in E = \mathbb{N}_0^2 \times \mathbb{N}_0^2$ gives the state. The action space is $A := \{1, 2\}$ where $a$ is the number of the arm which is chosen next. Obviously $D(x) = A$. The transition law is given by

$$\begin{aligned} q(x + e_{2a-1}|x, a) &= \frac{m_a + 1}{m_a + n_a + 2} =: p_a(x) \\ q(x + e_{2a}|x, a) &= 1 - p_a(x) \end{aligned}$$

where $e_a$ is the $a$-th unit vector. The one-stage reward at arm $a$ is $r(x, a) := p_a(x)$ which is the expected reward when we win one Euro in case of success and nothing else, given the information $x = (m_1, n_1, m_2, n_2)$ of successes and failures. We assume that $\beta \in (0, 1)$.

It is convenient to introduce the following notation, where $v : E \to \mathbb{R}$:

$$(Q_a v)(x) := p_a(x)v(x + e_{2a-1}) + (1 - p_a(x))v(x + e_{2a}), \quad x \in E.$$

Observe that since $r$ is bounded (i.e. we can choose $b \equiv 1$) and $\beta < 1$ we have a *contracting Markov Decision Model*. Moreover, the assumptions of Theorem 3.10 are satisfied and we obtain that the value function $J_\infty$ of the infinite horizon Markov Decision Model is the unique solution of

$$J_\infty(x) = \max\Big\{p_1(x) + \beta Q_1 J_\infty(x),\ p_2(x) + \beta Q_2 J_\infty(x)\Big\}, \quad x \in \mathbb{N}_0^2 \times \mathbb{N}_0^2$$

and a maximizer $f^*$ of $J_\infty$ defines an optimal stationary policy $(f^*, f^*, \ldots)$.
A very helpful tool in the solution of the infinite horizon bandit are the so-called *$K$-stopping problems*. In a $K$-stopping problem only one arm of the bandit is considered and the decision maker can decide whether she pulls the arm and continues the game or whether she takes the reward $K$ and quits. The maximal expected reward $J(m, n; K)$ of the $K$-stopping problem is then the unique solution of

$$v(m, n) = \max\Big\{K,\ p(m, n) + \beta\Big(p(m, n)v(m + 1, n) + (1 - p(m, n))v(m, n + 1)\Big)\Big\}$$

for $(m, n) \in \mathbb{N}_0^2$ where $p(m, n) = \frac{m+1}{m+n+2}$. Obviously it holds that $J(\cdot; K) \geq K$ and if $K$ is very large it will be optimal to quit the game, thus $J(m, n; K) = K$ for large $K$.

**Definition 3.11.** For $(m, n) \in \mathbb{N}_0^2$ we define the function

$$I(m, n) := \min\{K \in \mathbb{R} \mid J(m, n; K) = K\}$$

which is called *Gittins-index*.

The main result for the bandit problem is the optimality of the Gittins-index policy.

**Theorem 3.12.** *The stationary Index-policy $(f^*, f^*, \ldots)$ is optimal for the infinite horizon bandit problem where for $x = (m_1, n_1, m_2, n_2)$*

$$f^*(x) := \begin{cases} 2 & \text{if } I(m_2, n_2) \geq I(m_1, n_1) \\ 1 & \text{if } I(m_2, n_2) < I(m_1, n_1). \end{cases}$$

Remarkable about this policy is that we compute for each arm separately its own index (which depends only on the model data of this arm) and choose the arm with the higher index. This reduces the numerical effort enormous since the state space of the separate problems is much smaller. A small state space is crucial because of the curse of dimensionality for the value iteration.
The Bernoulli bandit with infinite horizon is a special case of the multiproject bandit. In a multiproject bandit problem $m$ projects are available which are all in some states. One project has to be selected to work on or one chooses to retire. The project which is selected then changes its state whereas the other projects remain unchanged. Gittins (1989) was the first to show that multiproject bandits can be solved by considering single-projects and that the optimal policy is an index-policy, see also Berry and Fristedt (1985). Various different proofs have been given in the last decades. Further extensions are *restless bandits* where the other projects can change their state too and bandits in continuous-time. Bandit models with applications in finance are e.g. treated in Bank and Föllmer (2003).

3.2.2. **Dividend Pay-out Problems.** Dividend pay-out problems are classical problems in risk theory. There are many different variants of it in discrete and continuous time. Here we consider a completely discrete setting which has the advantage that the structure of the optimal policy can be identified.
Imagine we have an insurance company which earns some premia on the one hand but has to pay out possible claims on the other hand. We denote by $Z_n$ the difference between premia and claim sizes in the $n$-th time interval and assume that $Z_1, Z_2, \ldots$ are independent and identically distributed with distribution $(q_k, k \in \mathbb{Z})$, i.e. $\mathbb{P}(Z_n = k) = q_k$ for $k \in \mathbb{Z}$. At the beginning of each time interval the insurer can decide upon paying a dividend. Of course this can only be done if the risk reserve at that time point is positive. Once the risk reserve got negative (this happens when the claims are larger than the reserve plus premia in that time interval) we say that the company is ruined and has to stop its business. The aim now is to maximize the expected discounted dividend pay out until ruin. In the economic literature this value is sometimes interpreted as the value of the company.
We formulate this problem as a stationary Markov Decision Problem with infinite horizon. The state space is $E := \mathbb{Z}$ where $x \in E$ is the current risk reserve. At the beginning of each period we have to decide upon a possible dividend pay out $a \in A := \mathbb{N}_0$. Of course we have the restriction that $a \in D(x) := \{0, 1, \ldots, x\}$ when $x \geq 0$ and we set $D(x) := \{0\}$ if $x < 0$. The transition probabilities are given by

$$q(x'|x, a) := q_{x'-x+a}, \quad x \geq 0,\ a \in D(x),\ x' \in \mathbb{Z}.$$

In order to make sure that the risk reserve cannot recover from ruin and no further dividend can be paid we have to freeze the risk reserve after ruin. This is done by setting

$$q(x|x, 0) := 1, \quad x < 0.$$

The dividend pay-out is rewarded by $r(x, a) := a$ and the discount factor is $\beta \in (0, 1)$. When we define the *ruin time* by

$$\tau := \inf\{n \in \mathbb{N}_0 \mid X_n < 0\}$$

then for a policy $\pi = (f_0, f_1, \ldots) \in F^\infty$ we obtain

$$J_{\infty\pi}(x) = \mathbb{E}_x^\pi \left[ \sum_{k=0}^{\tau-1} \beta^k f_k(X_k) \right].$$

Obviously $J_{\infty\pi}(x) = 0$ if $x < 0$. In order to have a well-defined and non-trivial model we *assume* that

$$\mathbb{P}(Z_1 < 0) > 0 \quad \text{and} \quad \mathbb{E}\, Z_1^+ < \infty.$$

Then the function $b(x) := 1 + x,\ x \geq 0$ and $b(x) := 0,\ x < 0$ is a bounding function with $\sup_\pi \mathbb{E}_x^\pi[b(X_n)] \leq b(x) + n\, \mathbb{E}\, Z_1^+,\ n \in \mathbb{N}$. Moreover, for $x \geq 0$ we obtain $\delta(x) \leq x + \frac{\beta\, \mathbb{E}\, Z_1^+}{1-\beta}$, and hence $\delta \in I\!B_b$. Thus, the Integrability Assumption (A) and the Convergence Assumption (C) are satisfied and $I\!M := I\!B_b$ fulfills (SA). Moreover, Theorem 3.4 yields that $\lim_{n\to\infty} J_n = J_\infty$ and

$$J_\infty(x) = (\mathcal{T} J_\infty)(x) = \max_{a \in \{0,1,\ldots,x\}} \left\{ a + \beta \sum_{k=a-x}^{\infty} J_\infty(x - a + k) q_k \right\}, \quad x \geq 0.$$

Obviously, $J_\infty(x) = 0$ for $x < 0$. Further, every maximizer of $J_\infty$ (which obviously exists) defines an optimal stationary policy $(f^*, f^*, \ldots)$. In what follows, let $f^*$ be the largest maximizer of $J_\infty$.

**Definition 3.13.** A stationary policy $f^\infty$ is called a *band-policy*, if there exist $n \in \mathbb{N}_0$ and numbers $a_0, \ldots a_n, b_1, \ldots b_n \in \mathbb{N}_0$ such that $b_k - a_{k-1} \geq 2$ for $k = 1, \ldots, n$ and $0 \leq a_0 < b_1 \leq a_1 < b_2 \leq \ldots < b_n \leq a_n$ and

$$f(x) = \begin{cases} 0, & \text{if } x \leq a_0 \\ x - a_k, & \text{if } a_k < x < b_{k+1} \\ 0, & \text{if } b_k \leq x \leq a_k \\ x - a_n, & \text{if } x > a_n \end{cases}$$

A stationary policy $f^\infty$ is called a *barrier-policy* if there exists $b \in \mathbb{N}_0$ such that

$$f(x) = \begin{cases} 0, & \text{if } x \leq b \\ x - b, & \text{if } x > b. \end{cases}$$

**Theorem 3.14.** a) *The stationary policy $(f^*, f^*, \ldots)$ is optimal and is a band-policy.*

b) *If $\mathbb{P}(Z_n \geq -1) = 1$ then the stationary policy $(f^*, f^*, \ldots)$ is a barrier-policy.*

The dividend payout problem has first been considered in the case $Z_n \in \{-1, 1\}$ by de Finetti (1957). Miyasawa (1962) proved the existence of optimal band-policies under the assumption that the profit $Z_n$ takes only a finite number of negative values. Other popular models in insurance consider the reinsurance and/or investment policies and ruin probabilities, see e.g. Martin-Löf (1994), Schäl (2004), Schmidli (2008).

## 4. Solution Algorithms

From Theorem 3.4 we know that the value function and an optimal policy of the infinite horizon Markov Decision Model can be obtained as limits from the finite horizon problem. The *value and policy iteration* already yield first computational methods to obtain a solution for the infinite horizon optimization problem. The use of simulation will become increasingly important in evaluating good policies. Much of the burden of finding an optimal policy surrounds the solution of the Bellman equation, for which now there are several simulation based algorithms such as *approximate dynamic programming*, see e.g. Powell (2007). There are also simulation based versions of both value and policy iteration. In this section we present two other solution methods.

### 4.1. Howard's Policy Improvement Algorithm.

We next formulate *Howard's policy improvement algorithm* which is another tool to compute the value function and an optimal policy. It goes back to Howard (1960) and works well in Markov Decision Models with finite state and action spaces.

**Theorem 4.1.** *Let (C) and (SA) be satisfied. Let $f, h \in F$ be two decision rules with $J_f, J_h \in \mathbb{M}$ and denote*

$$D(x, f) := \{a \in D(x) \mid LJ_f(x, a) > J_f(x)\}, \quad x \in E.$$

*Then it holds:*

*a) If for some subset* $E_0 \subset E$

$$h(x) \in D(x,f) \qquad \text{for } x \in E_0,$$
$$h(x) = f(x) \qquad \text{for } x \notin E_0,$$

*then* $J_h \geq J_f$ *and* $J_h(x) > J_f(x)$ *for* $x \in E_0$. *In this case the decision rule* $h$ *is called an* improvement *of* $f$.

*b) If* $D(x,f) = \emptyset$ *for all* $x \in E$ *and* $J_f \geq 0$, *then* $J_f = J_\infty$, *i.e. the stationary policy* $(f,f,\ldots) \in F^\infty$ *is optimal.*

*c) Let the Markov Decision Model be contracting. If* $D(x,f) = \emptyset$ *for all* $x \in E$, *then the stationary policy* $(f,f,\ldots) \in F^\infty$ *is optimal.*

If $F$ is finite then an optimal stationary policy can be obtained in a finite number of steps. Obviously it holds that $f \in F$ defines an optimal stationary policy $(f,f,\ldots)$ if and only if $f$ cannot be improved by the algorithm.

4.2. **Linear Programming.** Markov Decision Problems can also be solved by linear programming. We restrict here to the contracting case i.e. $\beta < 1$ and assume that state and action space are finite. We consider the following linear programs:

$$(P) \begin{cases} \sum_{x\in E} v(x) \to \min \\ v(x) - \beta \sum_y q(y|x,a) v(y) \geq r(x,a), \quad (x,a) \in D, \\ v(x) \in \mathbb{R},\ x \in E. \end{cases}$$

$$(D) \begin{cases} \sum_{(x,a)\in D} r(x,a)\mu(x,a) \to \max \\ \sum_{(x,a)} \big(\varepsilon_{xy} - \beta q(y|x,a)\big)\mu(x,a) = 1, \quad y \in E, \\ \mu(x,a) \geq 0, (x,a) \in D. \end{cases}$$

Note that $(D)$ is the dual program of $(P)$. Then we obtain the following result.

**Theorem 4.2.** *Suppose the Markov Decision Model is contracting and has finite state and action spaces. Then it holds:*

a) $(P)$ *has an optimal solution* $v^*$ *and* $v^* = J_\infty$.

b) $(D)$ *has an optimal solution* $\mu^*$. *Let* $\mu^*$ *be an optimal vertex. Then for all* $x \in E$, *there exists a unique* $a_x \in D(x)$ *such that* $\mu^*(x,a_x) > 0$ *and the stationary policy* $(f^*,f^*,\ldots)$ *with* $f^*(x) := a_x$, $x \in E$, *is optimal.*

Using so-called occupation measures general Markov Decision Problems with Borel state and action spaces can be solved by infinite dimensional linear programs, see e.g. Altman (1999), Hernández-Lerma and Lasserre (2002).

## 5. Further Topics on Markov Decision Processes

So far we have assumed that the decision maker has full knowledge about the distributional laws of the system. However, there might be cases where the decision maker has only partial information and cannot observe all driving factors of the model. Then the system is called a *Partially Observable Markov Decision Process*. Special cases are *Hidden Markov models*. Using results from filtering theory such models can be solved by a Markov Decision model (in the sense of sections 2 and 3) with an enlarged state space. This approach can be found in Bäuerle and Rieder (2011). Also the control of *Piecewise Deterministic Markov Processes* can be investigated via discrete-time Markov Decision Processes.

The presentation of the infinite horizon Markov Decision Processes is here restricted to the total reward criterion. However, there are many other optimality criteria like e.g. average-reward and risk-sensitive criteria. Average-reward criteria can be defined in various ways, a standard one is to maximize

$$\liminf_{n\to\infty} \frac{1}{n}\, \mathbb{E}_x^\pi \Big[ \sum_{k=0}^{n-1} r\big(X_k, f_k(X_k)\big) \Big].$$

This problem can be solved via the ergodic Bellman equation (sometimes also called Poisson equation). Under some conditions this equation can be derived from the discounted Bellman equation when we let $\beta \to 1$ (see e.g. Hernández-Lerma and Lasserre (1996)). This approach is called *vanishing discount approach*. The risk sensitive criterion is given by

$$\liminf_{n\to\infty} \frac{1}{n} \log \Big( \mathbb{E}_x^\pi \Big[ \exp\big(\gamma \sum_{k=0}^{n-1} r(X_k, f_k(X_k))\big) \Big] \Big)$$

where the "risk factor" $\gamma$ is assumed to be a small positive number in the risk-averse case. This optimization problem has attracted more recent attention because of the interesting connections between risk-sensitive control and game theory and has also important applications in financial optimization (see e.g. Bielecki et al. (1999), Borkar and Meyn (2002)).

## References


Altman, E. (1999) *Constrained Markov decision processes*. Chapman & Hall/CRC, Boca Raton, FL. 21

Bank, P. and Föllmer, H. (2003) American options, multi-armed bandits, and optimal consumption plans: a unifying view. In *Paris-Princeton Lectures on Mathematical Finance, 2002*, 1–42. Springer, Berlin. 18

Bäuerle, N. and Rieder, U. (2011) *Markov Decision Processes with applications to finance*. To appear: Springer-Verlag, Heidelberg. 2, 8, 11, 17, 21

Bellman, R. (1954) The theory of dynamic programming. *Bull. Amer. Math. Soc.* **60**, 503–515. 2

Bellman, R. (1957) *Dynamic programming*. Princeton University Press, Princeton, N.J. 2

Berry, D. A. and Fristedt, B. (1985) *Bandit problems*. Chapman & Hall, London. 18

Bertsekas, D. P. (2001) *Dynamic programming and optimal control. Vol. II*. Athena Scientific, Belmont, MA, second edition. 2, 11

Bertsekas, D. P. (2005) *Dynamic programming and optimal control. Vol. I*. Athena Scientific, Belmont, MA, third edition. 2, 11

Bertsekas, D. P. and Shreve, S. E. (1978) *Stochastic optimal control*. Academic Press Inc., New York. 2, 5, 14, 16

Bielecki, T., Hernández-Hernández, D., and Pliska, S. R. (1999) Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. *Math. Methods Oper. Res.* **50**, 167–188. 22

Blackwell, D. (1965) Discounted dynamic programming. *Ann. Math. Statist.* **36**, 226–235. 2

Borkar, V. and Meyn, S. (2002) Risk-sensitive optimal control for Markov decision processes with monotone cost. *Math. Oper. Res.* **27**, 192–209. 22

Dubins, L. E. and Savage, L. J. (1965) *How to gamble if you must. Inequalities for stochastic processes*. McGraw-Hill Book Co., New York. 2

Dynkin, E. B. and Yushkevich, A. A. (1979) *Controlled Markov processes*. Springer-Verlag, Berlin. 2

Enders, J., Powell, W., and Egan, D. (2010) A Dynamic Model for the Failure Replacement of Aging High-Voltage Transformers. *Energy Systems Journal* **1**, 31–59. 11

Feinberg, E. A. and Shwartz, A. (eds.) (2002) *Handbook of Markov decision processes*. Kluwer Academic Publishers, Boston, MA. 2

de Finetti, B. (1957) Su unímpostazione alternativa della teoria collettiva del rischio. *Transactions of the XVth International Congress of Actuaries* **2**, 433–443. 20

Gittins, J. C. (1989) *Multi-armed bandit allocation indices*. John Wiley & Sons Ltd., Chichester. 18

Goto, J., Lewis, M., and Puterman, M. (2004) Coffee, Tea or ...? A Markov Decision Process Model for Airline Meal Provisioning. *Transportation Science* **38**, 107–118. 11

Guo, X. and Hernández-Lerma, O. (2009) *Continuous-time Markov Decision Processes*. Springer-Verlag, New York. 11

He, M., Zhao, L., and Powell, W. (2010) Optimal control of dosage decisions in controlled ovarian hyperstimulation,. *Ann. Oper. Res.* 223–245. 11

Hernández-Lerma, O. and Lasserre, J. B. (1996) *Discrete-time Markov control processes*. Springer-Verlag, New York. 2, 22

Hernández-Lerma, O. and Lasserre, J. B. (2002) The linear programming approach. In *Handbook of Markov decision processes*, 377–408. Kluwer Acad. Publ., Boston, MA. 21

Hinderer, K. (1970) *Foundations of non-stationary dynamic programming with discrete time parameter*. Springer-Verlag, Berlin. 2

Howard, R. A. (1960) *Dynamic programming and Markov processes*. The Technology Press of M.I.T., Cambridge, Mass. 2, 20

Kushner, H. J. and Dupuis, P. (2001) *Numerical methods for stochastic control problems in continuous time*. Springer-Verlag, New York. 11

Martin-Löf, A. (1994) Lectures on the use of control theory in insurance. *Scand. Actuar. J.* 1–25. 20

Meyn, S. (2008) *Control techniques for complex networks*. Cambridge University Press, Cambridge. 11

Miyasawa, K. (1962) An economic survival game. *Operations Research Society of Japan* **4**, 95–113. 20

Peskir, G. and Shiryaev, A. (2006) *Optimal stopping and free-boundary problems*. Birkhäuser Verlag, Basel. 9

Powell, W. (2007) *Approximate dynamic programming*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. 2, 20

Puterman, M. L. (1994) *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons Inc., New York. 2

Ross, S. (1983) *Introduction to stochastic dynamic programming.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. 9

Schäl, M. (1990) *Markoffsche Entscheidungsprozesse.* B. G. Teubner, Stuttgart. 2, 16

Schäl, M. (2004) On discrete-time dynamic programming in insurance: exponential utility and minimizing the ruin probability. *Scand. Actuar. J.* 189–210. 20

Schmidli, H. (2008) *Stochastic control in insurance.* Springer, London. 20

Shapley, L. S. (1953) Stochastic games. *Proc. Nat. Acad. Sci.* **39**, 1095–1100. 2

Shiryaev, A. N. (1967) Some new results in the theory of controlled random processes. In *Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1965)*, 131–203. Academia, Prague. 2

Stokey, N. L. and Lucas, Jr., R. E. (1989) *Recursive methods in economic dynamics.* Harvard University Press, Cambridge, MA. 11

Tijms, H. (2003) *A first course in stochastic models.* John Wiley & Sons Ltd., Chichester. 11

(N. Bäuerle) Institute for Stochastics, Karlsruhe Institute of Technology, D-76128 Karlsruhe, Germany
*E-mail address*: nicole.baeuerle@kit.edu

(U. Rieder) Department of Optimization and Operations Research, University of Ulm, D-89069 Ulm, Germany
*E-mail address*: ulrich.rieder@uni-ulm.de