Jahresber Dtsch Math-Ver (2010) 112: 193–194
DOI 10.1365/s13291-010-0009-0

Vorwort Heft 4-10

Hans-Christoph Grunau

Online publiziert: 6. Oktober 2010

Wie schon im letzten Heft angekündigt erlauben uns die druckfertig vorliegenden Manuskripte auch dieses Mal wieder die Gestaltung eines Themenheftes, nun mit dem Schwerpunkt „Stochastik".

Juri Hinz erläutert in seinem Übersichtsartikel „Quantitative Modeling of Emission Markets" zunächst ganz allgemein verständlich die Grundlagen und Ziele des Handels mit Emissionszertifikaten. Er geht dabei auch auf problematische Aspekte der Marktmechanismen wie z. B. „windfall profits" (Mitnahmeeffekte) ein. Schritt für Schritt an Komplexität zunehmend und unter Verzicht auf technische Details führt Juri Hinz dann in die stochastische Modellierung des Handels mit Emissionszertifikaten ein.

Nicole Bäuerle und Ulrich Rieder stellen in ihrem Übersichtsartikel „Markov Decision Processes" eine Theorie vor, mit deren Hilfe man optimale Entscheidungsstrategien für Markovketten (zufällige Prozesse ohne Gedächtnis) finden kann. Die Autoren geben zunächst eine Einführung in die mathematischen Grundlagen dieser Theorie und wenden sich dann konkreten Anwendungen und Lösungsalgorithmen zu.

Wie auch im letzten Heft bilden die Buchbesprechungen einen inhaltlichen Kontrapunkt zu den Übersichtsartikeln. Hier liegt der Akzent auf der „Topologie" und insbesondere auf Nachbeben zu den Perelmanschen Durchbrüchen.

In der Regel erhalte ich – leider – nur wenige Reaktionen von Ihnen, den Leserinnen und Lesern des Jahresberichts, auf unsere Beiträge. Anders ist das bei dem Artikel von Roman Duda über „Die Lemberger Mathematikerschule", der im ersten

H.-Ch. Grunau (✉)
Institut für Analysis und Numerik, Fakultät für Mathematik, Otto-von-Guericke-Universität,
Postfach 4120, 39016 Magdeburg, Deutschland
e-mail: hans-christoph.grunau@ovgu.de

Heft dieses Jahres erschienen ist und der offenbar manche von Ihnen ebenso berührt hat wie mich. Dieses war natürlich nicht der erste Beitrag zu der bis zu deren brutalem Ende mathematisch so einflussreichen Gruppe um Banach und Steinhaus. Besonders interessant ist vielleicht der Hinweis eines Lesers auf „Banach und die Lemberger Schule der Funktionsanalysis“ von Gottfried Köthe, erschienen in den Mathematischen Semesterberichten 36 (1989), 145–158. Roman Duda hat zu derselben Thematik eine andere Perspektive gewählt, und so ergänzen sich beide Aufsätze sehr gut.

Jahresber Dtsch Math-Ver (2010) 112: 195–216
DOI 10.1365/s13291-010-0008-1

ÜBERSICHTSARTIKEL

Quantitative Modeling of Emission Markets

Juri Hinz

Received: 22 March 2010 / Published online: 18 September 2010

Abstract The introduction of marketable pollution rights is considered as an appropriate way to combat environmental problems on a global scale. According to theoretical arguments, a properly designed emission trading system should help reaching pollution reduction at low social costs. Nowadays, environmental markets are being established around the world. Their practice provides a stress test for the underlying economic theory and raises a lively discussion about advantages and shortcomings of emission trading. In this work, we highlight some core principles underlying quantitative understanding of emission markets and elaborate on mathematical problems and applications, arising in this context.

Keywords Stochastic modeling · Emission trading · Environmental finance · Risk-neutral pricing · Market equilibrium

Mathematics Subject Classification (2000) 91B70 · 91B60 · 91B76 · 93E20

J. Hinz (✉)
Department of Mathematics, National University of Singapore, 2 Science Drive, 117543 Singapore, Singapore
e-mail: mathj@nus.edu.sg

1 Introduction

During the last decades, market-based environmental instruments have attracted attention of policy makers all over the world. The role of these regulations is to institutionalize the creation of incentives for the use of cleaner technologies by the introduction of appropriate market mechanisms. In a generic design, a cap-and-trade mechanism works as follows: A central authority sets the quantity of emissions it will allow (the so-called cap) within a pre-determined compliance period and then allocates the corresponding amount of fully tradable pollution rights to businesses. Each source of emissions participating in the scheme must have sufficient permits to cover all its emissions by the end of the compliance period to avoid penalty which applies for each unit of pollutant not covered by permits.

Under the regulatory framework of a cap-and-trade system, the potential penalty payment creates a demand for allowances, which determines their price. Effectively, the buyer of certificates is charged for pollution, whereas the seller is rewarded for emission reduction. Based on this observation, economists argue that due to emission trading, the market price of emission certificates helps to identify and to exercise the cheapest reduction sources, ensuring so that the global pollution reduction can be reached at the lowest possible costs. Let us illustrate the underlying philosophy by the following example.

Example Consider two producers with the following characteristics

emissions/reductions	producer A	producer B
nominal emission p.a.	10,000 tonnes of CO_2	10,000 tonnes of CO_2
reduction costs	40 EURO per tonne of CO_2	10 EURO per tonne of CO_2

Suppose that the regulator decides to reduce the total emissions by 10%. To reach this goal, the central authority allocates allowances covering 9,000 tonnes of CO_2 to each of the producer and sets a penalty of 100 EURO for each tonne of pollutant not covered by allowances. To highlight the effect of cost reduction triggered by allowance trading, we compare the scheme with non-transferable emission rights to that with marketable permits. In the first case, both producers realize that it is cheaper to save 1,000 tonnes of carbon dioxide (fulfilling the compliance) than to pay penalties. Hence the emission reduction scheme with non-transferable pollution rights yields the following result

emissions/reductions	producer A	producer B
realized emission p.a.	9,000 tonnes of CO_2	9,000 tonnes of CO_2
total reduction costs	40,000 EURO	10,0000 EURO

Thus, the overall reduction with non-transferable rights costs 50,000 EURO. However, if the pollution rights are marketable, then the same overall reduction is cheaper, since the agents behave differently. For the producer A, it is cheaper to buy allowances at any price below 40 EURO than to reduce emissions. At the same time, the agent

B is willing to sell allowances at any price above 10 EURO facing the opportunity to reduce the own emission at 10 EURO per tonne. Although the price, at which the agents trade certificates can not be determined within this simple framework, the effect of marketable pollution rights is clear: The agent A buys certificates from the agent B instead of reducing own emissions. This gives the following result:

emissions/reductions	producer A	producer B
realized emission p.a.	10,000 tonnes of CO_2	8,000 tonnes of CO_2
total reduction costs	0 EURO	20,0000 EURO

That is, marketable emission rights yield overall reduction costs of 20,000 EURO instead of 50,000 EURO in the previous case. In this example, the marketability of permits ensures remarkable savings. As we see, emission trading forces the agent B who owns the cheapest reduction sources to exercise the own abatement potential beyond the individual targets. In other words, the market rules help identifying and using the cheapest way to fulfill the overall reduction target.

Despite problems with existing emission trading schemes, a widely accepted economic viewpoint considers market mechanism as an appropriate way to optimally allocate emission abatement potential within the entire economy. The hope is that a properly designed cap-and-trade system should help reaching pollution reduction at the lowest social costs. Furthermore, due to the success of the U.S. Acid Rain Program, emission trading is now widely considered as one of the most promising policy instruments to combat environmental pollution on a large scale.

Still, emissions trading is subject of a lively discussion: Proponents of the liberalized markets argue in favor of market mechanisms emphasizing the role of price signals which should help identifying efficient pollution reduction measures, whereas opponents advise against the introduction of marketable pollution permits and prefer emission taxation, believing that the certificate price fluctuations may disturb individual firms in the implementation of correct abatement strategies.

To date, the important examples of reduction mechanisms include the emissions trading under the U.S. Acid Rain Program and EU ETS (European Union Emission Trading Scheme), which is the largest among the existing emission markets. The EU ETS covers more than 10,000 installations in 25 countries, responsible for nearly half of the European Union's emissions. Currently, this market operates in the second phase (2008–2012), this time frame coincides with the Kyoto period.

2 New Problems and Challenges

After emission trading became reality, new problems have occurred.

Financial Risk The major point here is that each emissions market participant is exposed to a risk resulting from the fluctuations of the allowance prices. The need for appropriate risk management leads to the creation of an accompanying market for derivative contracts. Before we explain the nature of the emission-related financial products listed to date, let us elaborate on the very philosophy of risk reduction.

Market players trade financial instruments to reduce their exposure from potential price changes. Say, to secure a future production, the owner of a fuel-consuming business (airline company) needs to buy the fuel in advance. To avoid storage and transportation problems, most of the market participants prefer trading on the so-called forward basis.

Let us put aside a relatively complex exact definition of a forward contract. We only point out that in this agreement, the delivery and the price are negotiated at the moment the contract is signed, whereas the physical delivery and payment is scheduled for a future date. Thus, the owner of the forward contract is protected against the price increase of the fuel in the following way: if the fuel becomes expensive at the delivery date, then the market price of the forward contract also increases. A futures contract is very similar to a forward contract but is more standardized and provides a slightly different cash flow. Beyond forwards and futures, options are popular in the area of risk management. The simplest option type is the so-called European Call with a pre-specified strike price $K \in]0, \infty[$ and maturity date $\tau \in [0, T]$. This instrument works as follows: if the price of the underlying security (say, fuel) at date τ exceeds the strike price K, then the owner of the Call receives the difference between underlying's price and the strike price K. Such contract may be issued on an arbitrary security. For instance, it can be written on a futures price rather than on the physical fuel price. In any case, the underlying asset (forward, future) of the Call must be pre-determined in the contract specification.

The major idea of risk management by financial instruments is that agents buy complex financial products which provide in particular market situations appropriate payments, giving so a certain protection. However, this type of risk handling requires a detailed understanding of derivatives price evolution. Although its principles are based on sound mathematical cornerstones, their adaptation to specific situations is involving. Thus, to determine the so-called fair prices of emission-linked derivatives, one needs to develop appropriate mathematical models. This is a challenging task, since market participants already now face a notable sophistication due to a significant regulatory complexity inherent to the real-world emission markets.

Before we proceed with quantitative modeling, let us list some of the most important financial agreements, traded to date:

EUAs (European Union Allowances) are certificates which cover the emission of one tonne of carbon dioxide (or an equivalent greenhouse gas) within EU ETS. This is the prime asset of the European Emission Tranding Scheme. Physically, EUAs are realized by entries in appropriate electronic registries.

CERs (Certified Emission Reductions) are certificates issued by bodies of the UN Framework Convention on Climate Change and the Kyoto Protocol. They are given upon a successful completion of the climate protection projects (in the sense of the so-called Clean Development Mechanism, specified in the Kyoto protocol). In this sense, CERs can be considered as international environmental assets. The European market EU ETS is linked to the international market since in the second phase of the EU ETS, market players are allowed to cover their emissions not only by EUAs but also by CERs, subject to certain restrictions. On this account, EUAs and CERs are traded at a similar price level.

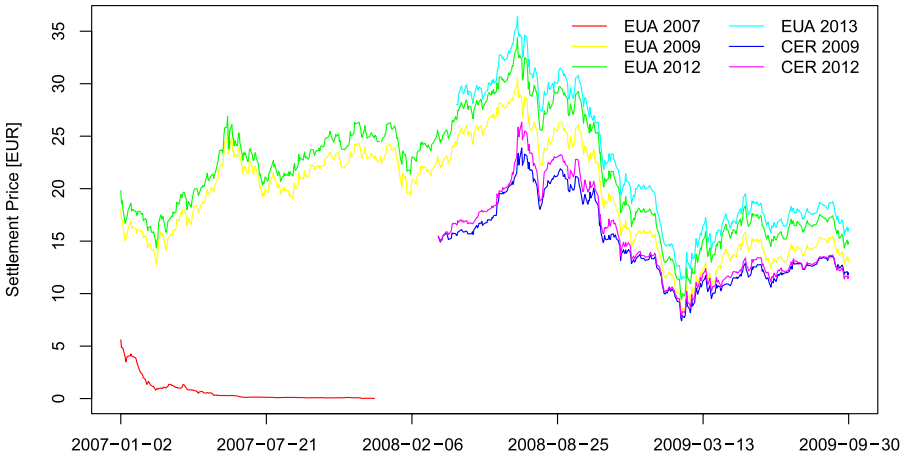

Fig. 1 Future prices on EUA with maturity Dec. 2012

Futures on EUAs and CERs are traded on several exchanges. For instance, the European Climate Exchange (ECX) lists EUA and CER futures with expiry date on the first Monday of March, June, September and December.

Options such as Standard European Calls whose underlying assets are EUA and CER futures, are listed, too. At the ECX, these options are available with maturity date three business days before the expiry date of the underlying future. The strikes are ranging from 1 to 55 EURO with an interval of 0.5 EURO. The traded volume is quoted in tonnes of carbon dioxide and is approximatively 450,000 tonnes per day.

The Fig. 1 shows the historical price evolution of EUA and CER futures listed at ECX.

Windfall Profits This is a problem of a notable gravity, addressing possible shortcomings of the current market design and its potential improvements. Let us highlight the main point, avoiding again formal arguments.

The key notion in this context is the so-called *opportunity costs*. In the economic literature, it stands for the forgone benefit from using a certain strategy compared to the next best alternative. For example, the opportunity costs of farming own land is the amount which could be obtained by renting the land to someone else.

When facing electricity generation, producers consider a profit, which could be potentially realized when instead of production, unused emission allowances were sold to the market. For instance, if the price of the emission certificate is 12 EURO per tonne of CO_2 and the production of one Megawatt-hour emits two tonnes of CO_2, (say, using a coal-fired steam turbine), then the producer must decide between two strategies which are equivalent in terms of their emission certificate balance:

- produce and sell one MWh to the market,
- not produce this MWh and sell allowances covering two tonnes of CO_2.

In this situation, the opportunity costs of producing one Megawatt-hour is $24 = 2 \times 12$ EURO. Obviously, the agent produces energy only if the first strategy is at

least as profitable that the second. Thereby, both, the production and the opportunity costs must be considered in the formation of the electricity market price. Clearly, if the production costs of electricity are 30 EURO per MWh, then the energy will be produced only if its price covers both, the production and the opportunity costs. Thus, electricity can only be delivered at the price exceeding $54 = 30 + 2 \times 12$ EURO. That is, in order to trigger the electricity production, the opportunity costs must be added to the production costs giving the lowest possible wholesales price. In the scientific community, this phenomena is well-known under the name of *cost-pass-through*. An empirical analysis [19], confirms that the strategy of *cost-pass-through* is currently followed by the European energy producers. Furthermore, the detailed investigation of mathematical market models shows that the *cost-pass-through* is the only possible strategy in the so-called equilibrium state of the market. This can be interpreted as follows: When behaving optimally, the energy producers must pass the allowance price on the consumers. We can not blame the energy producers for this, even if the emission credits are allocated free of charge.

More importantly, it turns out that the *cost-pass-through* is nothing but the core mechanism, responsible for the emission savings. Namely, due to the opportunity costs, clean technologies appear cheaper than emission-intense production strategies. For instance, an alternative generation (gas turbine) which yields energy at the price of 40 EURO and emits only one tonne of carbon dioxide, hardly competes with coal-fired steam turbine under generic regime (without emissions regulation). Namely, if there is no regulatory framework, then the coal-fired steam turbine is scheduled first and the gas turbine has to wait until the energy demand can not be covered by coal-fired steam technology. However, given emission regulation, the opposite is true: Say, if the allowance price equals to 12 EURO as above, then the gas technology appears cheaper, operating at full costs of $52 = 40 + 1 \times 12$ EURO. Thus, the gas turbine is scheduled first, followed by the coal-fired steam turbine which runs only if the installed gas turbine capacity does not cover the energy demand.

That is, we obtain the following picture: The opportunity costs is equal to the allowance price times the specific emission rate. This costs must be added to the original electricity production costs, which changes the merit order of technologies. The emission savings are triggered automatically, making clean technologies cheaper than those which are emission-intense. The electricity price increases through the costs-pass through. The burden to the consumers is *partially* justified, since the energy production becomes cleaner, hence its production more expensive.

A surprising and disappointing fact here is that the consumer's burden can be justified only partially! Analyzing a typical energy market, one realizes that the consumers have usually to pay far more than what is spent due to emission reduction. The difference yields additional revenues to the energy producers and is frequently referred to as the so-called *windfall profit*. Quantitatively, it depends on the available technologies, their capacities and demand fluctuation, so let us highlight the effect of *windfall profit* by considering an artificially simple market example.

Example Suppose that the energy consumption does not depend on its price. In this market, high electricity price triggers no demand reduction and so there will be no emission savings because of high energy price. In addition, we assume that the entire

market consists of a single electricity production technology (say, coal-fired steam turbines, as above). In such a market, any emissions regulation does not yield a decrease of pollution. Indeed, there is no reaction on the production side due to the lack of alternative technologies, and there is also no response on the consumption side, since there is no demand reduction. In this market, the energy producers pass certificates costs on the consumer and just pocket the windfall profits without any emission reduction!

Of course, real markets do have different technologies and the energy demand reacts to its price. However, the ability to re-schedule the energy production may be limited, particularly during high load, when all production capacities become busy. Furthermore, the short-term price dependence of the electricity demand on its price is known to be very low. This explains why energy producers within EU ETS realize significant additional revenues, exclusively due to the emissions trading.

Fortunately, the problem of windfall profits may have a sound solution in terms of a correction to the existing market rules. Following ideas of proportional allowance allocation, the work [5] investigates in detail an emission market model where the amount of allocated allowances is linked to the production activity. To clarify the basic idea, consider in the above example a market rule which subsidizes each produced MWh by an additional allocation of allowances covering $1/2$ tonnes of pollutant. If the allowance price is 12 EURO per tonne of CO_2, then the opportunity costs are reduced by 6 EURO for each Megawatt-hour and the electricity may become cheaper. However, within this scheme, another difficult problem occurs. A priori, it is not clear how to adjust the upfront allowance allocation such that the market following a subsidized scheme reaches the same performance in emission reduction. The work [5] addresses this and related questions and develops a mathematical methodology for the quantitative analysis, comparison, and optimization of emission market rules with respect to a wide range of criteria.

In this paper, we let aside the market design questions and refer the interested reader to [5] and to the literature cited therein. The goal of the present work is to illustrate how the mathematical analysis of stochastic models may help understanding the mechanics of allowance price formation to tackle diverse questions in the area of risk management, which arise from the viewpoint of the individual market player.

3 A Toy Model of Emission Market

To explain the emission price mechanism, we present a toy market model where a finite set I of the agents are confronted with abatement of pollution. The key assumptions are:

- We consider a trading scheme in isolation, without credit transfer from and to other markets. That is, unused emission allowances expire worthless.
- There is no strategy adjustment within the compliance period $[0, T]$. This means that the agents schedule their production plans and trade allowances only at the beginning. At the compliance date T, they trade certificates one last time before emission reports are surrendered to the regulator.

- For the sake of simplicity, we set the interest rate to zero.

To describe uncertainties, we realize this model on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ by introducing the following ingredients.

Emission dynamics. For each agent $i \in I$, consider the quantity $E_0^i \in [0, \infty[$ which describes the total pollution of agent i emitted within the entire compliance period $[0, T]$ in the case of the so-called 'business-as-usual' scenario (where no abatement measures are applied).

Abatement. Consider the opportunity to reduce emissions. Each agent i can decide at the beginning $t = 0$ of the compliance period to reduce its emissions by $\xi_0^i \in [0, E_0^i]$ pollutant units.

Abatement costs. We assume that the costs of abatement is a function of the reduced volume. Thus, if the agent i decides reducing own emissions by $x \in [0, \infty[$ units, then this causes costs of $C^i(x)$ currency units. This dependence is described by a strictly convex and continuous function $x \mapsto C^i(x)$ on $]0, \infty[$.

Allowance trading. Suppose that at any time $t \in \{0, T\}$, credits can be exchanged between agents by trading at the price A_t at time $t = 0, T$. Denote by ϑ_t^i the change at time t in allowance number, held by agent i. That is, given the allowance prices $(A_t)_{t\in\{0,T\}}$, the allowance trading $(\vartheta_t^i)_{t\in\{0,T\}}$ yields total costs of

$$\vartheta_0^i A_0 + \vartheta_T^i A_T. \tag{1}$$

Penalty payment. The total pollution of the agent i can be expressed as a difference

$$E_0^i - \xi_0^i$$

of the 'business-as-usual' emission E_0^i less the entire reduction ξ_0^i. As mentioned above, a penalty $\pi \in]0, \infty[$ is being paid at maturity T for each unit of pollutant which is not covered by allowances. Considering the total change in the allowance position $\vartheta_0^i + \vartheta_T^i$ effected by trading, the loss of the agent i resulting from potential penalty payment is

$$\pi(E_0^i - \xi_0^i - \vartheta_0^i - \gamma^i - \vartheta_T^i)^+ \tag{2}$$

where

$$\gamma^i,\ i \in I \text{ is the initial allowance allocation of the agent } i \in I. \tag{3}$$

Uncertainty. We need to take into account that due to uncertainty in the emission control, the actual emission realized at the end of the period $[0, T]$ may slightly differ from E_0^i. It is convenient to subtract this deviation from the initial allocation. Hence, we interpret γ^i as the credits allocated to the agent i less emissions which become known with certainty only at time T. With this interpretation, γ^i stands for allowances effectively available for compliance and is modeled by a random variable.

Individual wealth. In view of (1) and (2), the revenue of the agent i following trading strategy $(\vartheta_t^i)_{t\in\{0,T\}}$ and abatement policy ξ_0^i equals

$$L^{A,i}(\vartheta_0^i, \vartheta_T^i, \xi_0^i) = -\vartheta_0^i A_0 - C^i(\xi_0^i) - \vartheta_T^i A_T - \pi(E_0^i - \xi_0^i - \vartheta_0^i - \gamma^i - \vartheta_T^i)^+. \tag{4}$$

Note that the quantities ξ_0^i, ϑ_0^i, A_0, E_0^i, carrying the subscript $t = 0$ are observed at the beginning and are modeled by deterministic quantities whereas ξ_T^i, ϑ_T^i, γ^i, A_T are observable at the end $t = T$ and are thus described by random variables.

Risk aversion. To face risk preferences, suppose that risk attitudes of the agents $i \in I$ are described by utility functions $U^i : \mathbb{R} \mapsto \mathbb{R}$, which are continuous, strictly increasing and concave. Consider the utility functional $u^i(X) = \mathbb{E}(U^i(X))$, which is assumed to be defined for each random variable X where the expectation is finite or $+\infty$. Given an allowance price process $A = (A_t)_{t\in\{0,T\}}$, each agent i behaves rationally, maximizing $(\vartheta_0^i, \vartheta_T^i, \xi_0^i) \mapsto u^i(L^{A,i}(\vartheta_0^i, \vartheta_T^i, \xi_0^i))$ by an appropriate choice of the own policy $(\vartheta_0^{i*}, \vartheta_T^{i*}, \xi_0^{i*})$. This approach is common in financial modeling, where the risk averse behavior is described by the attempt to maximize a certain nonlinear functional applied to the random variable which describes the agent's wealth, rather than to maximize the expectation of the wealth.

Market equilibrium. In the economic literature, a steady market state in a real situation is frequently represented by the so-called equilibrium within a mathematical market model. Basically, such an equilibrium describes a situation where the price is determined by the rules of supply and demand with a price given such that each of the market participants is satisfied with the own production and trading. The equilibrium modeling of financial markets is a widespread area, which combines diverse research approaches and yields numerous application. The question whether it is reasonable to describe a realistic market operation by an equilibrium state requires a case-by-case discussion. In the situation of an emission market, we may assume that after a certain period, all market participants have adjusted their behavior to the new regulatory environment. Doing so, they determine own production and emissions trading and dynamically respond to the allowance prices in order to maximize the own revenue.

Let us agree that in our framework, a realistic steady market state is given by the following dynamic Nash-type equilibrium:

Definition 1 The price evolution $A^* = (A_t^*)_{t\in\{0,T\}}$ describes an equilibrium of emission market if for each $i \in I$, there exist an emission trading strategy $(\vartheta_t^{*i})_{t\in\{0,T\}}$ and an abatement policy ξ_0^{*i}, such that $u^i(L^{A^*,i}(\vartheta_0^{*i}, \vartheta_T^{*i}, \xi_0^{*i}))$ is finite and

(i) the cumulative changes in positions are in zero net supply, that is,

$$\sum_{i\in I} \vartheta_t^{*i} = 0, \quad \text{for all } t = 0, T, \tag{5}$$

(ii) each agent $i \in I$ is satisfied by the own action in the sense that

$$\begin{aligned} &u^i\left(L^{A^*,i}(\vartheta_0^{*i}, \vartheta_T^{*i}, \xi_0^{*i})\right) \geq u^i\left(L^{A^*,i}(\vartheta_0^i, \vartheta_T^i, \xi_0^i)\right) \\ &\quad \text{if } u^i\left(L^{A^*,i}(\vartheta_0^i, \vartheta_T^i, \xi_0^i)\right) \text{ exists.} \end{aligned} \tag{6}$$

Note that (ii) states that each participant does not have any incentive to change the own strategy, whereas (i) represents the restriction that the total number of allowances in the market remains the same (as issued at the beginning).

The following folk principle

the equilibrium allowance price equals to the marginal abatement costs (7)

is considered as crucial in the economic analysis of emissions markets. This principle means that in a realistic market state, the certificates are traded at a price which coincides with the costs of reduction of an additional unit of pollutant. More precisely, it states an important intrinsic connection between the price of certificates and the abatement activity applied in the market. Let us illustrate this idea by an example.

Example Assume that emission allowances are traded at a price of 12 EURO. According to (7), this means that all abatement sources with reduction price less than 12 EURO per tonne of pollutant are already active. That is, if one additional tonne of pollutant needs to be saved, then this would cost 12 EURO or more, since all measures giving a cheaper reduction are already exhausted. More importantly, (7) shows that in a steady market state, the individual abatement looks like it was externally driven by allowance price. Namely, for an individual market player, the correct behavior is determined by the recent allowance price, traded at the market, as follows: Having noticed that the allowances are traded at 12 EURO per tonne of pollutant, this agent derives the own optimal strategy: The best individual action is to apply all abatement measures whose reduction price is lower than 12 EURO per pollutant unit and not to use any other abatement activity. This decision is very intuitive since the corresponding profit is obvious: Having saved emission at a price less than 12 EURO, the corresponding amount of allowances can be sold at the market price of 12 EURO which leaves the same emission balance in the books but gives a real profit.

It turns out that the principle (7) holds in the above equilibrium. To see this, let us transform (7) in mathematical terms.

Abatement volume. First, we determine the volume of abatement measures available to the agent i at a price a. Remember the strictly convex and continuous function $x \mapsto C^i(x)$ which describes agent's i costs of diminishing own emission by $x \in [0, \infty[$ pollutant units. For simplicity, assume for the moment that C^i is continuously differentiable, with derivative $\dot{C}^i(x)$ interpreted as the costs (per unit of pollutant) of an infinitesimally small reduction of beyond x. This quantity is referred to as the *marginal* reduction costs. The marginal reduction costs $x \mapsto \dot{C}^i(x)$ is increasing in the reduction volume x. The intuition behind this is that the more reduction is applied, the more measures are exercised and become unavailable. Thus, the more expansive it will be to further increase pollution abatement. In particular, given reduction costs $a \in [0, \infty[$, the point x^* at which the minimum of $x \mapsto C^i(x) - ax$ is reached, stands exactly for the abatement volume x^*, where the marginal abatement costs are equal to a, this point x^* is uniquely determined by $\dot{C}^i(x^*) = a$. Denoting by *argmin* the point where the minimum is reached, the definition

$$c^i(a) := \operatorname{argmin}\{C^i(x) - ax : x \in [0, E_0^i]\} \tag{8}$$

stands for cumulative abatement volume of the agent i, available at price a. Note that this point exists and is uniquely defined due to strict convexity of $x \mapsto C^i(x) - ax$

considered as a function on the compact interval $[0, E_0^i]$. (Note that we have assumed the differentiability of C^i for explanation only, the smoothness of C^i is not required in (8)). Using (8), the principle (7) is reflected in the following

Proposition 1 *Suppose that $(A_t^*)_{t\in\{0,T\}}$ is an equilibrium allowance price and that ξ_0^{i*} for $i \in I$ are the corresponding equilibrium abatement policies, then*

$$\xi_0^{i*} = c^i(A_0^*) \quad \textit{holds for each } i \in I. \tag{9}$$

The connection between abatement activity and emission allowance price is an important piece of the puzzle. It can be considered as a feedback relation: The higher is the allowance price, the more abatement measures are active, the more emission savings is triggered. The increasing price of emission certificates lowers the chance that at the end of period, the market will need all issued allowances to cover the emission. Some agents may expect an oversupply of allowances at the end and face the chance to buy them later, at time $t = T$, at a cheaper price. Such market participants tend to sell allowances in advance, with the intention to repurchase them later. The market recognizes this trend and the inial allowance price tends to lower agian. This explains the mechanism driving the price to an equilibrium level by a feedback between certificate price and abatement activity.

To face this mechanism in quantitative form, our model follows the following insights:

(a) *No arbitrage.* The major ingredient of the rational action is the anticipation of the future price change. This natural behavior is already covered by model assumptions. It turns out that in the above equilibrium framework, there is no *arbitrage* opportunity. On the mathematical level, it means that there is no riskless gain from trading. A central result from financial mathematics states that this property is equivalent to the assumption that $(A_t)_{t\in\{0,T\}}$ follows a martingale with respect to a measure $\mathbb{Q}$ which is equivalent to the given probability measure $\mathbb{P}$.

(b) *Price triggers abatement.* As explained above, the marginal abatement costs coincide with the allowance price. Thus, the allowance price triggers all abatement measures whose costs are below the allowance price. If we summarize the total volume of abatement measures available in the market at a price $a \in [0, \infty[$ as

$$c(a) := \sum_{i\in I} c^i(a), \quad a \in [0, \infty[, \tag{10}$$

then the total abatement in the market equals to $c(A_0^*)$, given allowance price $(A_t^*)_{t\in\{0,T\}}$.

(c) *Terminal price is digital.* The idea is that at maturity, the price must fall to zero if there is an excess in allowances, whereas in the case of their shortage, the price will rise, reaching penalty. Under the mild mathematical assumption

$$\text{distribution of } \sum_i \gamma_i \text{ possesses no point masses} \tag{11}$$

an exact coincidence of allowance demand and supply occurs with zero probability and can be neglected. In this case, defining the overall allowance shortage by

$$\mathcal{E}_T = \sum_{i \in I} (E_0^i - \gamma^i), \tag{12}$$

the non-compliance event is written as

$$\mathcal{E}_T - c(A_0^*) > 0.$$

Let us put the insights (a), (b) and (c) together to close the circle.

Proposition 2 *Suppose that* (11) *holds. Given an equilibrium allowance price* $(A_t^*)_{t\in\{0,T\}}$, *there exists a measure* $\mathbb{Q}$ *which is equivalent to* $\mathbb{P}$ *such that* $(A_t^*)_{t\in\{0,T\}}$ *follows a* $\mathbb{Q}$*-martingale whose terminal value is given by*

$$A_T^* = \pi 1_{\{\mathcal{E}_T - c(A_0^*) \geq 0\}}. \tag{13}$$

Although individual market attributes and actions of all agents seem to be irrelevant in the above statement, the reader should notice that this picture appears only from the risk-neutral viewpoint. In line with standard aggregation theorems, the equilibrium market state heavily depends on and is determined by market architecture, rules, risk attitudes and uncertainty. However, once equilibrium is reached and all arbitrage opportunities are exhausted, asset dynamics can be considered under a risk neutral measure. With respect to this measure, market evolution appears as it was driven by cumulative quantities only.

According to the core principles of financial mathematics, any realistic modeling of asset price dynamics must exclude arbitrage. In practice, this is achieved by proposing stochastic models equipped with a built-in measure $\mathbb{Q}$ such that all relevant price processes follow martingales with respect to $\mathbb{Q}$. This approach is frequently referred to as *martingale modeling*. In view of our equilibrium analysis, the *martingale modeling* of emission allowance prices can be stated as the following stand-alone problem

$$\left.\begin{array}{|c|}\hline \text{given measure } \mathbb{Q} \sim \mathbb{P}, \text{ random variable } \mathcal{E}_T, \\ \text{and market abatement volume functions } c, \\ \text{determine a } \mathbb{Q}\text{-martingale } (A_t^*)_{t\in\{0,T\}} \text{ with} \\ A_T^* = \pi 1_{\{\mathcal{E}_T - c(A_0^*) \geq 0\}}. \\ \hline \end{array}\right\} \tag{14}$$

Essentially, this problem addresses the calculation of the fixed point $A_0^* \in [0, \infty[$ to

$$A_0^* = E^{\mathbb{Q}}\left(\pi 1_{\{\mathcal{E}_T - c(A_0^*) \geq 0\}}\right) = \pi \mathbb{Q}(\mathcal{E}_T \geq c(A_0^*)).$$

Note that the function on the right $A_0 \mapsto \pi \mathbb{Q}(\mathcal{E}_T \geq c(A_0))$ is non-increasing, since the market abatement volume $A_0 \mapsto c(A_0)$ is non-decreasing in the allowance price A_0. Under the mild assumption (11) this function is continuous. By the intermediate

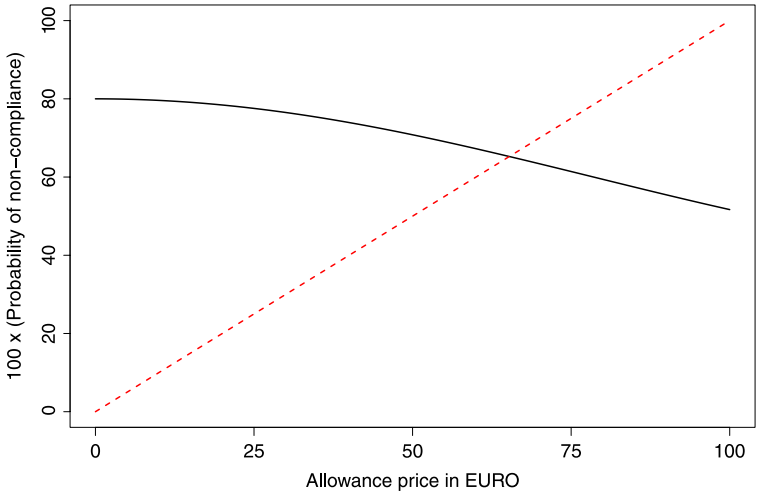

Fig. 2 *Solid graph* $A_0 \mapsto \pi \mathbb{Q}(\mathcal{E}_T \geq c(A_0))$ intersects *dashed graph* $A_0 \mapsto A_0$ (we assume $\pi = 100$)

value theorem, its graph possesses a unique intersection point A_0^* with the graph of the identical function $A_0 \mapsto A_0$, as shown in the Fig. 2.

With this approach, the solution $(A_t^*)_{t\in\{0,T\}}$ of (14) suggests an evolution of allowance price dynamics which responds to the no-arbitrage principle and inherits important properties from equilibrium. Note that the market abatement volume function c can be reasonably estimated from analysis of real-world data (from available production technologies and installed production capacities). In difference to this, the random variable $\mathcal{E}_T$ must be interpreted as an exogenous ingredient of the risk-neural model. The point here is that although the total business-as-usual emission $\sum_{i\in I} E_0^i$ and total allowance allocation $\sum_{i\in I} \gamma^i$ can be reasonably modeled from market data, the distribution of their difference $\mathcal{E}_T$ needs to be described from the risk-neutral perspective, thus it is not directly accessible. In any case, the random variable $\mathcal{E}_T$ should be chosen appropriately, such that the solution $(A_t^*)_{t\in\{0,T\}}$ to the martingale modeling problem (14) shows desirable properties of allowance price evolution. For instance, the distribution of $\mathcal{E}_T$ with respect to $\mathbb{Q}$ should be adjusted such that A_0^* matches the observed initial allowance prices. Further restrictions may be stated, if practitioners require that the listed initial prices of important emission-linked derivatives must also be matched. Such techniques are well-accepted in the financial industry and are known as *implied model calibration*.

4 Dynamic Risk Neutral Modeling

Let us now extend the above model to a finite number $\{0, 1, 2, \ldots, T\}$ of discrete time points at which the agents apply abatements and trade allowances. As above, we re-state the martingale modeling problem and address its solution. Although it is less obvious, the idea is the same and the existence proof is based on a recursive application of the intermediate value theorem starting with the last point $T - 1$ prior

to the compliance date T. In this setting, all distributions are conditional and one also needs to take into account abatement, applied in the past.

Consider a market with a finite number I of the agents confronted with emission reduction. Assume that $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=0}^T)$ is a filtered probability space and agree to consider only stochastic processes adapted to $(\mathcal{F}_t)_{t=0}^T$. For each agent $i \in I$, introduce $(E_t^i)_{t=0}^{T-1}$ which describes the dynamics of the so-called 'business-as-usual' emission with the interpretation that E_t^i stands for the total pollution of agent i which is emitted within the time interval $[t, t+1]$ if no abatement is applied. Suppose that each agent i can decide at any time $t = 0, \dots, T-1$ to reduce its emissions within $[t, t+1]$ by ξ_t^i pollutant units and to buy and to sell credits at price A_t. Denote by ϑ_t^i the change at time t in allowance number, held by agent i. Similarly to the one-period model, the revenue of the agent i following allowance trading $\vartheta^i = (\vartheta_t^i)_{t=0}^T$ and abatement policy $\xi^i = (\xi_t^i)_{t=0}^{T-1}$ equals

$$L^{A,i}(\vartheta^i, \xi^i) = -\sum_{t=0}^{T-1}(\vartheta_t^i A_t + C^i(\xi_t^i)) - \vartheta_T^i A_T - \pi\left(\sum_{t=0}^{T-1}(E_t^i - \xi_t^i - \vartheta_t^i) - \gamma^i - \vartheta_T^i\right)^+. \tag{15}$$

In a complete analogy to (14) one defines the overall allowance shortage

$$\mathcal{E}_T = \sum_{i\in I}\left(\sum_{t=0}^{T-1} E_t^i - \gamma^i\right)$$

and shows that an analysis of the equilibrium situation in the framework of this dynamical model leads to the feedback relation

$$\left.\begin{array}{|c|}\hline \text{given measure } \mathbb{Q} \sim \mathbb{P}, \text{ random variable } \mathcal{E}_T, \\ \text{and market abatement volume functions } c, \\ \text{determine a } \mathbb{Q}\text{-martingale } (A_t^*)_{t=0}^T \text{ with} \\ A_T^* = \pi 1_{\{\mathcal{E}_T - \sum_{t=0}^{T-1} c(A_t^*) \geq 0\}}. \\ \hline \end{array}\right\} \tag{16}$$

To discuss a solution $(A_t^*)_{t=0}^T$ to this problem, introduce the martingale $(\mathcal{E}_t)_{t=0}^T$ and consider its increments $(\varepsilon_t)_{t=1}^T$

$$\mathcal{E}_t = \mathbb{E}^Q(\mathcal{E}_T \mid \mathcal{F}_t), \quad t = 0, \dots, T, \qquad \varepsilon_t = \mathcal{E}_t - \mathcal{E}_{t-1}, \quad t = 1, \dots, T.$$

It turns out that under the standing assumption (11), the problem (16) possesses a solution, which can be written as

$$A_t^* = \alpha_t\left(\mathcal{E}_t - \sum_{s=0}^{t-1} c(A_s^*)\right), \quad \omega \in \Omega, \ \ t = 0, \dots, T. \tag{17}$$

The natural interpretation of this form is that the emission allowance price depends on

- the present time t
- the global situation ω
- the market position $G_t = \mathcal{E}_t - \sum_{s=0}^{t-1} c(A_s^*)$

through appropriate functionals

$$\alpha_t : \mathbb{R} \times \Omega \to [0, \pi], \quad \mathcal{B}(\mathbb{R}) \otimes \mathcal{F}_t\text{-measurable, for } t = 0, \dots T \tag{18}$$

whose recursive calculation is discussed in [11]. The work [11] demonstrates the advantages and the importance of the 'Markovian' case, where (18) are actually path-independent, being true functions $\alpha_t : \mathbb{R} \to [0, \pi]$. This situation occurs if for each $t = 0, \dots, T$, the martingale increment ε_{t+1} is independent of $\mathcal{F}_t$.

Given (18), the price process $(A_t^*)_{t=0}^T$ is well-defined by an application of the following recursion

$$\text{set} \quad G_0 := \mathcal{E}_0, \quad \text{then for } t = 0, \dots, T \text{ define} \tag{19}$$

$$A_t^*(\omega) := \alpha_t(G_t(\omega))(\omega), \tag{20}$$

$$G_{t+1}(\omega) := G_t(\omega) - c(A_t^*(\omega))(\omega) + \varepsilon_{t+1}(\omega). \tag{21}$$

The resulting risk-neutral dynamics (20)–(21) is examined in [11], where valuation of emission-related derivatives is investigated in the framework of the Monte Carlo methodology, for the Markovian case. Although this work demonstrates computational tractability of discrete time models, they are hardly suitable for real-world applications in the present form. However, the gained understanding helps to generalize the martingale approach to modeling of emission allowance prices in continuous time.

5 Continuous Time Models

The formulation (16) provides a natural passage to

$$\left.\begin{array}{|c|}\hline \text{On the probability space } (\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t\in[0,T]}) \\ \text{given measure } \mathbb{Q} \sim \mathbb{P}, \text{ random variable } \mathcal{E}_T, \\ \text{and market abatement volume function } c, \\ \text{determine a } \mathbb{Q}\text{-martingale } (A_t^*)_{t\in[0,T]} \text{ with} \\ A_T^* = \pi 1_{\{\mathcal{E}_T - \int_0^T c(A_t^*)dt \geq 0\}}. \\ \hline \end{array}\right\} \tag{22}$$

Although this problem has not yet been addressed in a sufficient generality, reasonable solutions can be constructed by an ad-hoc analogy. Following insights from the discrete time, the continuous-time counterpart can be addressed in terms of partial differential equations. Namely, the results of the discrete-time analysis given in [11] suggest that, if the increments of the martingale

$$(\mathcal{E}_t := \mathbb{E}^{\mathbb{Q}}[\mathcal{E}_T \mid \mathcal{F}_t])_{t\in[0,T]} \tag{23}$$

are independent, then one can reasonably expect that a solution to (22) can have the functional form

$$A_t = \alpha(t, G_t), \quad t \in [0, T],$$

with an appropriate deterministic function

$$\alpha : [0, T] \times \mathbb{R} \mapsto \mathbb{R} \tag{24}$$

and state process $(G_t)_{t\in[0,T]}$ given by

$$G_t := \mathcal{E}_t - \int_0^t c(A_s)ds, \quad t \in [0, T]. \tag{25}$$

The work [2] follows this approach in the framework of jump-diffusion processes. For simplicity, let us illustrate their results in the pure diffusion settings.

Given the process $(W_t, \mathcal{F}_t)_{t\in[0,T]}$ of a standard Brownian motion, suppose that the martingale (23) is modeled as

$$d\mathcal{E}_t = \sigma_t dW_t$$

where $(\sigma_t)_{t\in[0,T]}$ is deterministic. Further, assume that we are given a continuous non-decreasing abatement volume function c. To ensure the martingale property of the allowance price process $(A_t = \alpha(t, G_t))_{t\in[0,T]}$, we use Itô's formula and (25) to write the stochastic differential of this process as

$$\begin{aligned} dA_t = d\alpha(t, G_t) &= \partial_{(1,0)}\alpha(t, G_t)dt + \partial_{(0,1)}\alpha(t, G_t)dG_t + \frac{1}{2}\partial_{(0,2)}\alpha(t, G_t)d[\mathcal{E}]_t \\ &= \partial_{(1,0)}\alpha(t, G_t)dt - \partial_{(0,1)}\alpha(t, G_t)c(\alpha(t, G_t))dt + \frac{1}{2}\partial_{(0,2)}\alpha(t, G_t)\sigma^2(t)dt \\ &\quad + \partial_{(0,1)}\alpha(t, G_t)\sigma(t)dW_t. \end{aligned}$$

Here $[\mathcal{E}]_t$ stands for the quadratic variation of the martingale $(\mathcal{E}_t)_{t\in[0,T]}$ and $\partial_{(i,j)}$ denotes the respective partial derivatives. Now, we observe that to ensure the martingale property of $(A_t)_{t\in[0,T]}$ it is necessary that the function α solves the partial differential equation

$$\partial_{(1,0)}\alpha(t, x) - c(\alpha(t, x))\partial_{(0,1)}\alpha(t, x) + \frac{1}{2}\sigma^2(t)\partial_{(0,2)}\alpha(t, x) = 0 \tag{26}$$

in $]0, T[\times\mathbb{R}$ with the boundary condition

$$\alpha(T, x) = \pi \mathbf{1}_{[0,\infty[}(x), \quad x \in \mathbb{R}, \tag{27}$$

justified by the digital terminal allowance price. (Note that the boundary value is given at the final time $t = T$. The change of variable $T - t$ transforms this backward initial value problem to the standard form.) The following summarizes the above-presented approach.

1. Given a continuous non-decreasing function $c : \mathbb{R} \to \mathbb{R}_+$ and a deterministic $(\sigma_t)_{t\in[0,T]}$, determine a solution α to the boundary value problem (26), (27).
2. Verify that there is a unique strong solution to

$$dG_t = d\mathcal{E}_t - c(\alpha(t, G_t))dt, \qquad G_0 = \mathcal{E}_0. \tag{28}$$

3. Introduce the allowance price by $(A_t := \alpha(t, G_t))_{t\in[0,T]}$.

Having constructed the allowance price process $(A_t)_{t\in[0,T]}$ in this way, one obtains a standard procedure for the valuation of European options. Indeed, observe that, due to the Markov property of the strong solution to (28), the fair time t price of a European Call option written on the allowance price at (maturity) date $\tau \in]t, T]$ is given in terms of an appropriate function of the state variable:

$$C_t = \mathbb{E}^{\mathbb{Q}}\left[(A_\tau - K)^+ \mid \mathcal{F}_t\right] = \mathbb{E}^{\mathbb{Q}}\left[(\alpha(\tau, G_\tau) - K)^+ \mid \mathcal{F}_t\right] = f^\tau(t, G_t).$$

To ensure that $(C_t = f^\tau(t, G_t))_{t\in[0,\tau]}$ is a martingale, the function $f^\tau : [0, \tau[\times\mathbb{R} \to \mathbb{R}$ is to be taken as a solution to the linear partial differential equation

$$\partial_{(1,0)} f^\tau(t,x) - c(\alpha(t,x))\partial_{(0,1)} f^\tau(t,x) + \frac{1}{2}\partial_{(0,2)} f^\tau(t,x)\sigma^2(t) = 0, \tag{29}$$

in $]0, \tau[\times\mathbb{R}$. However, the boundary condition in this case will be

$$f^\tau(\tau, x) = (\alpha(\tau, x) - K)^+, \quad x \in \mathbb{R}. \tag{30}$$

Summarizing, we obtain the following description for the procedure.

1. Find the function α as above.
2. Given the strike price $K \geq 0$ and maturity time $\tau \in [0, T]$ of a European Call, calculate f^τ as the solution to the backward initial value problem (29), (30).
3. Given the allowance price $a \in [0, \pi]$ at recent time $t \in [0, \tau]$, obtain x as the solution to $\alpha(t, x) = a$.
4. Substitute t and the thus obtained x into the function f^τ to obtain the time t price of the European Call as $f^\tau(t, x)$.

6 Reduced-form Models

Instead of characterizing the non-compliance event N by

$$N := \left\{\mathcal{E}_T - \int_0^T c(A_s)ds \geq 0\right\}$$

in terms of the exogenously specified random variable $\mathcal{E}_T$ and function c, the reduced form approach focuses on a direct modeling

Specify the non-complance event $N \subset \Omega$ and model the allowance prices by a $\mathbb{Q}$-martingale $(A_t)_{t\in[0,T]}$ with terminal value $A_T = \pi 1_N$.	(31)

Although there are many candidates for such a process, it is not obvious how to satisfy the requirements from the practical side. For a practitioner trying to calibrate a model at time $\tau \in [0, T]$, the minimum requirement is to match the price observed at time τ, as well as the observed price fluctuation intensity up to this time τ. Further desired

properties include the existence of closed-form formulas or at least fast valuation schemes for European options, a small number of parameters providing sufficient model flexibility, and reliable and fast parameter identification from historical data.

One natural way to describe the non-compliance in the emission market can be formulated in terms of the process of geometric Brownian motion

$$d\Gamma_t = \Gamma_t \sigma dW_t, \quad \Gamma_0, \sigma \in]0, \infty[, \text{ with standard Brownian motion } (W_t)_{t\in[0,T]}.$$

A first attempt to model the non-compliance is to set

$$N = \{\Gamma_T \geq 1\}.$$

However, in this case, the martingale

$$a_t := \mathbb{E}^{\mathbb{Q}}(1_{\{\Gamma_T \geq 1\}} | \mathcal{F}_t), \quad t \in [0, T]$$

follows

$$da_t = \Phi'(\Phi^{-1}(a_t)) \frac{1}{\sqrt{T-t}} dW_t \tag{32}$$

independently on the choice of σ. Such approach lacks calibration capability, since it is not possible to fit the fluctuation intensity of emission certificates $(A_t = \pi a_t)_{t\in[0,T]}$ to the actual situation observed in the market. From this viewpoint, one needs to introduce additional degrees of freedom to (32). In [3], it is suggested to consider instead of (32) the dynamics

$$da_t = \Phi'(\Phi^{-1}(a_t)) \sqrt{\beta (T-t)^{-\alpha}} dW_t, \quad t \in [0, T] \tag{33}$$

with two additional parameters $\alpha \in \mathbb{R}$ and $\beta \in]0, \infty[$.

This approach is a particular case of the following general construction. Given a Brownian motion $(\tilde{W}_t, \tilde{\mathcal{F}}_t)_{t\in[0,\infty[}$ consider the standard normal distribution function Φ in order to define

$$Y_t = \Phi\left(e^{\frac{t}{2}} \left(x_0 + \int_0^t e^{-\frac{s}{2}} d\tilde{W}_s \right) \right), \quad t \in [0, \infty[.$$

Further, transform this process to

$$a_t = Y_{\psi(t)}, \quad t \in [0, T[,$$

by time change

$$\psi(t) = \int_0^t z_s ds < +\infty, \quad t \in [0, T[\tag{34}$$

which is defined in terms of a pre-specified positive-valued deterministic function $(z_t)_{t\in[0,T]}$. It turns out that the time-changed process possesses a stochastic differential

$$da_t = \Phi'(\Phi^{-1}(a_t)) \sqrt{z_t} dW_t, \quad t \in [0, T[\tag{35}$$

with respect to another Brownian motion $(W_t, \mathcal{F}_t)_{t\in[0,T[}$ given by

$$dW_t = \frac{1}{\sqrt{z_t}} d\tilde{W}_{\psi(t)}, \qquad \mathcal{F}_t = \tilde{\mathcal{F}}_{\psi(t)}, \quad t \in [0, T[.$$

Such construction yields a sufficiently rich class of martingales $(a_t)_{t\in[0,T]}$ with values in $[0, 1]$. This martingale family is parameterized via positive-valued deterministic functions $(z_t)_{t\in[0,T]}$ which must satisfy (34). It turns out that the terminal value is digital $a_T \in \{0, 1\}$ if and only if $\lim_{t\uparrow T} \psi(t) = +\infty$. In the parameterization (33), this is the case if $\alpha \geq 1$.

The main advantage of the present modeling is that the allowance price appears as a function of a Gaussian process. Due to this form, the parameters α and β can be efficiently estimated from historical data. Furthermore, the European Call options can be calculated via simple numerical integration. Namely, under the assumption that the interest rate $r \in]0, \infty[$ is deterministic, the price of a European Call option with strike price $K \geq 0$ and maturity date $\tau_o \in [0, T]$ written on allowance futures price maturing at $\tau_f \in [\tau_o, T]$, is given at time $t \in [0, \tau_o]$ by

$$C_t = e^{-r(\tau_o - t)} \int_{\mathbb{R}} (\pi\, \Phi(x) e^{-r(T-t_f)} - K)^+ N(\mu_{t,\tau_o}, \nu_{t,\tau_o})(dx) \tag{36}$$

where the parameters of the normal distribution are given for $\beta > 0, \alpha = 1$ by

$$\mu_{t,\tau_o} = \Phi^{-1}\left(\frac{A_t(\tau_f)}{\pi} e^{-r(T-t_f)}\right)\left(\frac{T-t}{T-\tau_o}\right)^{\beta/2},$$

$$\nu_{t,\tau_o} = \left(\frac{T-t}{T-\tau_o}\right)^{\beta} - 1$$

and for $\beta > 0, \alpha \in \mathbb{R} \setminus \{1\}$ by

$$\mu_{t,\tau_o} = \Phi^{-1}\left(\frac{A_t(\tau_f)}{\pi} e^{-r(T-t_f)}\right) e^{-\beta \frac{(T-\tau_o)^{-\alpha+1} - (T-t)^{-\alpha+1}}{2(-\alpha+1)}},$$

$$\nu_{t,\tau_o} = e^{-\beta \frac{(T-\tau_o)^{-\alpha+1} - (T-t)^{-\alpha+1}}{-\alpha+1}} - 1.$$

Here $A_t(\tau_f)$ denotes the market price at time t of the emission allowance future with maturity date t_f.

7 Outlook and Conclusion

So far, we focused on a generic cap and trade scheme modeled after the first phase of the EU ETS, namely limited to one compliance period and without banking in the sense that unused allowances become worthless at the end of the period. This is a strong simplification since as already mentioned above, real-world markets are operating in a multi-period framework. Furthermore, subsequent periods are connected

by market specific regulations. Presently, there are three regulatory mechanisms connecting successive compliance periods in a cap-and-trade scheme. Their rules go under the names of *borrowing*, *banking* and *withdrawal*.

- Borrowing allows for the transfer of a (limited) number of allowances from the next period into the present one;
- Banking allows for the transfer of a (limited) number of (unused) allowances from the present period into the next;
- Withdrawal penalizes firms which fail to comply in two ways: by penalty payment for each unit of pollutant which is not covered by credits and by withdrawal of the missing allowances from their allocation for the next period.

From the nature of the existing markets and the designs touted for possible implementation, it seems that policy makers tend to favor unlimited banking and forbid borrowing. Furthermore, the withdrawal rule is most likely to be included. Banking and withdrawal seem to be reasonable rules to reach an emission target within a fixed number of periods, because each success (resp. failure) in the previous period results in stronger (resp. weaker) abatement in the subsequent periods. The pricing of allowance options within multi-period reduced-form models is addressed in [3].

The introduction of mandatory emission trading schemes all over the world opens up perspectives for environmental protection but rises also difficile questions in the area of quantitative financial modeling. These problems encompass diverse aspects of market design, game theoretic market modeling, econometric and statistical techniques. Nowadays, numerous financial places trade a large volume of emission allowances and allowance-linked derivatives. The trading activity is increasing, although market participants seem to lack theoretical principles for pricing of these contracts. In this situation, practitioners require reliable and sound solutions. That is, there is an insisting need for adaptation of methodologies, developed within financial mathematics, to the new asset class of environmental financial instruments. With this work, we intend to highlight some of the recent questions in this area and to encourage a further research.

Appendix: Literature Overview

The publications on quantitative aspects of emission trading are rather extensive, and we refer the interested reader to [21] which provides a valuable guide to publications, however far from being complete. The *economic theory* of allowance trading goes back to [9] and [14], where the public good 'environment' was proposed by means of transferable permits. Important results in *dynamic allowance trading* were obtained in [8, 12, 13, 16, 17, 20, 22] and in the literature cited therein. Recently, after the introduction of the real-world emission market EU ETS, *empirical evidence* has become available. The experience gained from market operation is discussed in [10], and a detailed analysis of allowance prices from this market is given in [23] and [24]. The contributions [1] and [15] are devoted to *econometric modeling* of emission allowance prices. The modeling of *dynamic price equilibrium* is addressed in [4] and [5], which provide a mathematical analysis of the market equilibrium and use optimal stochastic control theory to show social optimality of emission trading schemes. A recent

work [11] considers equilibrium of risk averse market players and elaborates on risk neutral dynamics. The problems of *derivative valuation* in emission markets are also addressed. The paper [7] discusses an endogenous emission permit price dynamics within an equilibrium setting and elaborates on the valuation of European options on emission allowances. The dissertation [25] and the paper [18] deal with the risk-neutral allowance price formation within the EU ETS. The work [6] is also devoted to option pricing within EU ETS. Finally, the recent work [3] presents an approach where emission certificate futures are modeled in terms of a deterministic time change applied to a certain class of interval-valued diffusion processes.

References

1. Benz, E., Trueck, S.: Modeling the price dynamics of CO_2 emission allowances. Energy Econ. **31**, 4–15 (2008)
2. Borovkov, K., Decrouez, G., Hinz, J.: Jump-diffusion modeling in emission markets. Stoch. Models **27**(1) (2011, to appear)
3. Carmona, R., Hinz, J.: Risk neutral modeling of emission allowance prices and option valuation. Technical report, Princeton University (2009)
4. Carmona, R., Fehr, F., Hinz, J.: Optimal stochastic control and carbon price formation. SIAM J. Control Optim. **48**, 2168–2190 (2009)
5. Carmona, R., Fehr, F., Hinz, J., Porchet, A.: Market designs for emissions trading schemes. SIAM Rev. **52**(3), 403–452 (2010)
6. Cetin, U., Verschuere, M.: Pricing and hedging in carbon emissions markets. Int. J. Theor. Appl. Finance (IJTAF) **12**(7), 949–967 (2009)
7. Chesney, M., Taschini, L.: The endogenous price dynamics of the emission allowances: An application to CO_2 option pricing. Technical report (2008)
8. Cronshaw, M., Kruse, J.B.: Regulated firms in pollution permit markets with banking. J. Regul. Econ. **9**(2), 179–189 (1996)
9. Dales, J.H.: Pollution, Property and Prices. University of Toronto Press, Toronto (1968)
10. Daskalakis, G., Psychoyios, D., Markellos, R.N.: Modeling CO_2 emission allowance prices and derivatives: Evidence from the European Trading Scheme. J. Bank. Finance **33**(7), 1230–1241 (2009)
11. Hinz, J., Novikov, A.: On fair pricing of emission-related derivatives. Bernoulli (to appear). Available online: http://www.bernoulli-society.org/index.php/publications/bernoulli-journal/bernoulli-journal-papers
12. Leiby, P., Rubin, J.: Intertemporal permit trading for the control of greenhouse gas emissions. Environ. Resour. Econ. **19**(3), 229–256 (2001)
13. Maeda, A.: Impact of banking and forward contracts on tradable permit markets. Environ. Econ. Policy Stud. **6**(2), 81–102 (2004)
14. Montgomery, W.D.: Markets in licenses and efficient pollution control programs. J. Econ. Theory **5**(3), 395–418 (1972)
15. Paolella, M.S., Taschini, L.: An econometric analysis of emissions trading allowances. J. Bank. Finance **32**(10), 2022–2032 (2008)
16. Rubin, J.: A model of intertemporal emission trading, banking and borrowing. J. Environ. Econ. Manag. **31**(3), 269–286 (1996)
17. Schennach, S.M.: The economics of pollution permit banking in the context of title iv of the 1990 clean air act amendments. J. Environ. Econ. Manag. **40**(3), 189–21 (2000)
18. Seifert, J., Uhrig-Homburg, M., Wagner, M.: Dynamic behavior of CO_2 spot prices. J. Environ. Econ. Manag. **56**(2), 180–194 (2008)
19. Sijm, J., Neuhoff, K., Chen, Y.: CO_2 cost pass-through and windfall profits in the power. Climate Policy **6**, 49–72 (2006)
20. Stevens, B., Rose, A.: A dynamic analysis of the marketable permits approach to global warming policy: a comparison of spatial and temporal flexibility. J. Environ. Econ. Manag. **44**, 45–69 (2002)
21. Taschini, L.: Environmental economics and modeling marketable permits: a survey. Preprint (2009)
22. Tietenberg, T.: Emissions Trading: An Exercise in Reforming Pollution Policy. Resources for the Future, Boston (1985)

23. Uhrig-Homburg, M., Wagner, M.: Derivatives instruments in the EU emissions trading scheme, and early market perspective. Energy Environ. **19**(5), 635–655 (2008)
24. Uhrig-Homburg, M., Wagner, M.: Futures price dynamics of CO_2 emissions certificates: an empirical analysis. J. Deriv. **17**(2), 73–88 (2009)
25. Wagner, M.: CO_2-Emissionszertifikate, Preismodellierung und Derivatebewertung. PhD thesis, Universität Karlsruhe (2006)

Juri Hinz completed his Ph.D. in Mathematics in 1997 and became then an Assistant at the University of Tübingen. From 2003 to 2007 he worked as Senior Scientist at ETH Zurich, where he led several research projects. In 2007 the Association of European Operations Research Society (EURO) awarded one of his projects the Prize of Excellence in Practice. Since 2007, Dr. Juri Hinz is an Associate Professor for Financial Mathematics, Probability Theory and Mathematical Statistics at the National University of Singapore. His publications deal with portfolio optimization, real-time auctions on electricity, modeling day-ahead electricity prices, and pricing of commodity derivatives.

Jahresber Dtsch Math-Ver (2010) 112: 217–243
DOI 10.1365/s13291-010-0007-2

ÜBERSICHTSARTIKEL

Markov Decision Processes

Nicole Bäuerle · Ulrich Rieder

Received: 14 April 2010 / Published online: 8 September 2010

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.

Keywords Markov decision process · Markov chain · Bellman equation · Policy improvement · Linear programming

Mathematics Subject Classification (2010) 90C40 · 60J05 · 93E20

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.

N. Bäuerle (✉)
Institute for Stochastics, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany
e-mail: nicole.baeuerle@kit.edu

U. Rieder
Department of Optimization and Operations Research, University of Ulm, 89069 Ulm, Germany
e-mail: ulrich.rieder@uni-ulm.de

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 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 Sect. 2.3.1.

First books on Markov Decision Processes are [5] and [25]. The term 'Markov Decision Process' has been coined by [4]. Shapley [37] 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 [32]. Mathematical rigorous treatments of this optimization theory appeared in [9, 11, 13, 24, 38] and [14]. More recent textbooks on this topic are [7, 8, 16, 22, 31, 32, 34] and [3].

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 Sect. 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 [3].

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 (for a general evolution see Fig. 1). Hence 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 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 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 $q_n(x'|x, a) := Q_n(\{x'\}|x, a)$.

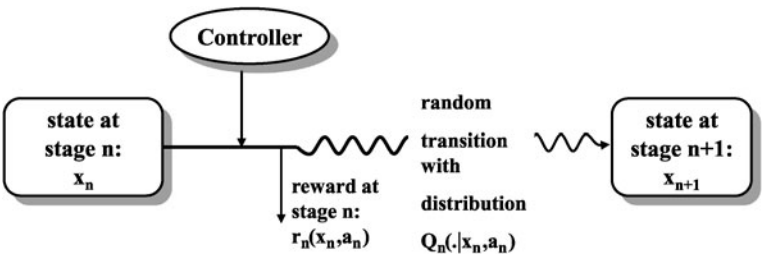

Fig. 1 General evolution of a Markov Decision Model

- $r_n : D_n \to \mathbb{R}$ 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

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

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 $\pi = (f_0, f_1, \ldots, f_{N-1})$ 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

$$\mathbb{P}_x^\pi(X_0 \in B) = \varepsilon_x(B) \quad \text{for all } B \in \mathfrak{E},$$
$$\mathbb{P}_x^\pi(X_{n+1} \in B | X_1, \ldots, X_n) = \mathbb{P}_x^\pi(X_{n+1} \in B | X_n) = Q_n\big(B | X_n, f_n(X_n)\big),$$

where ε_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 (A_N) holds for the N-stage Markov Decision Problems throughout this section. Obviously Assumption (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 π and start in state $x \in E$ at time n. The *value function* V_n is defined by

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

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 $\pi \in F_0 \times \cdots \times F_{N-1}$ 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 $\pi \in F_0 \times \cdots \times F_{N-1}$ 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 π and n.

We define the following operators for $n = 0, 1, \ldots, N-1$ and $v \in \mathbb{M}(E)$:

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

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

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

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*

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

see e.g. [9]. 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 Sect. 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* $n = 0, 1, \ldots, N-1$

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

(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 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 (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 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 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 (SA_N), $V_n \in \mathbb{M}_n$. □

2.2 Semicontinuous Markov Decision Processes

In this section we give sufficient conditions under which assumptions (A_N) and (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 (A_N) is obviously satisfied and (SA_N) is satisfied with $\mathbb{M}_n$ and Δ_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 (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 (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 (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 $\mathbb{M}_n := \{v \in \mathbb{M}(E) \mid v^+ \in \mathbb{B}_b, v \text{ is upper semicontinuous}\}$ *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 [3], Chap. 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 $D_{N-1}((0, 1)) = D_{N-1}((1, 0)) = \{1\}$. 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

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

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 [30]. For more gambling problems see [33].

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

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

Moreover, we assume that $\mathbb{E}[B_{n+1}^\top R B_{n+1}]$ 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 π. 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 (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 (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 (SA_N) (i): Obviously $x^\top R_N x \in \mathbb{M}_N$. Now let $v(x) = x^\top Rx \in \mathbb{M}_{n+1}$. We try to solve the following optimization problem

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

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

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

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

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

where $\tilde{R}$ is defined as the expression in the brackets. Note that $\tilde{R}$ is symmetric and since $x'\tilde{R}x = \mathcal{T}_n v(x) \geq x^\top R_n x$, 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 2.4 to solve the stochastic linear-quadratic control problem.

Theorem 2.10

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

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

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

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

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. [7, 8]). 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. [3, 7, 8, 28, 39, 40]). 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 [19] (Logistics), [15] (Energy systems) and [21] (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 [26]. 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 [20].

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 π (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

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

Moreover, it is convenient to introduce the set

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

Obviously, we have $J_{\infty\pi} \in \mathbb{B}$ for all policies π. 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

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

Next we define the *limit value function* by

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

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 π. Taking the supremum over all π 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

$$q(a|1, a) := 1,$$

$$q(2|2, a) := 1,$$

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

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

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

Since the reward is non-positive, assumptions (A) and (C) are satisfied.

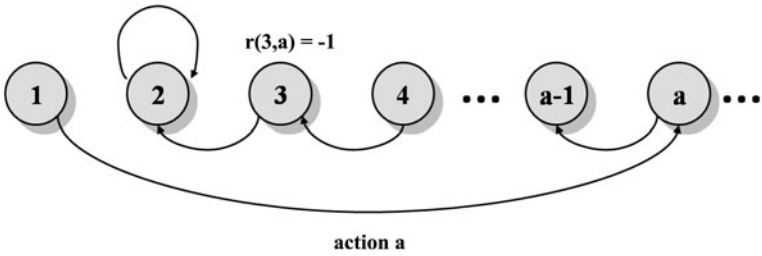

Fig. 2 Transition diagram of Example 3.1

We will compute now J and J_∞. 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, $J_\infty(1) = -1 \neq 0 = J(1) = \lim_{n\to\infty} J_n(1)$.

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_∞ are in general not in $\mathbb{B}$. However, J_∞ 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. [9]. As in Sect. 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_∞ 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 Sect. 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_∞ *is the largest* r-subharmonic function v *in* $\mathbb{M} \cap \mathbb{B}$, *i.e* J_∞ *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_∞, *and every maximizer* f^* *of* J_∞ *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_∞ 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

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

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_∞ *is upper semicontinuous.*
(c) $\emptyset \neq Ls\, D_n^*(x) \subset D_\infty^*(x)$ *for all* $x \in E$ (Policy Iteration).
(d) *There exists an* $f^* \in F$ *with* $f^*(x) \in Ls\, D_n^*(x)$ *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^∞ 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

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

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. [9, 34]).

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

$\delta \in \mathbb{B}_b$ 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 $v, w \in \mathbb{B}_b$ it holds*:

$$\|\mathcal{T}_f v - \mathcal{T}_f w\|_b \le \beta\alpha_b \|v - w\|_b,$$
$$\|\mathcal{T} v - \mathcal{T} w\|_b \le \beta\alpha_b \|v - w\|_b.$$

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

Theorem 3.9 (Verification Theorem) *Let b be a bounding function, $\beta\alpha_b < 1$ and let $v \in \mathbb{B}_b$ be a fixed point of $\mathcal{T} : \mathbb{B}_b \to \mathbb{B}_b$. 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 $(\mathbb{B}_b, \|\cdot\|_b)$ 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 $\mathbb{M} \subset \mathbb{B}_b$ and a set $\Delta \subset F$ such that*

(i) $0 \in \mathbb{M}$,
(ii) $\mathcal{T} : \mathbb{M} \to \mathbb{M}$,
(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_∞ *is the unique fixed point of $\mathcal{T}$ in $\mathbb{M}$.*
(c) J_∞ *is the smallest r-superharmonic function $v \in \mathbb{M}$, i.e. J_∞ is the smallest function $v \in \mathbb{M}$ with $v \ge \mathcal{T} v$.*
(d) *Let $v \in \mathbb{M}$. Then*

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

(e) *There exists a maximizer $f \in \Delta$ of J_∞, and every maximizer f^* of J_∞ 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 [3]. 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 θ_1 and θ_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_∞ 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_∞ 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 [18] 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 [6]. 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 [2].

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 $\delta \in \mathbb{B}_b$. Thus, the Integrability Assumption (A) and the Convergence Assumption (C) are satisfied and $\mathbb{M} := \mathbb{B}_b$ 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_∞ (which obviously exists) defines an optimal stationary policy $(f^*, f^*, \ldots)$. In what follows, let f^* be the largest maximizer of J_∞.

Definition 3.13 A stationary policy f^∞ 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 $0 \leq a_0 < b_1 \leq a_1 < b_2 \leq \cdots < b_n \leq a_n$ 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^∞ 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 [17]. Miyasawa [29] 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. [27, 35, 36].

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. [31]. 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 [25] 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* $E_0 \subset E$

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

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:

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

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

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* μ^*. *Let* μ^* *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. [1, 23].

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 Sects. 2 and 3) with an enlarged state space. This approach can be found in [3]. 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

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

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. [22]). This approach is called *vanishing discount approach*. The risk sensitive criterion is given by

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

where the "risk factor" γ 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. [10, 12]).

References

1. Altman, E.: Constrained Markov Decision Processes. Chapman & Hall/CRC, Boca Raton (1999)
2. Bank, P., Föllmer, H.: American options, multi-armed bandits, and optimal consumption plans: a unifying view. In: Paris-Princeton Lectures on Mathematical Finance, 2002, pp. 1–42. Springer, Berlin (2003)
3. Bäuerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Heidelberg (2011, to appear)
4. Bellman, R.: The theory of dynamic programming. Bull. Am. Math. Soc. **60**, 503–515 (1954)
5. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
6. Berry, D.A., Fristedt, B.: Bandit Problems. Chapman & Hall, London (1985)
7. Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. II, 2nd edn. Athena Scientific, Belmont (2001)
8. Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. I, 3rd edn. Athena Scientific, Belmont (2005)
9. Bertsekas, D.P., Shreve, S.E.: Stochastic Optimal Control. Academic Press, New York (1978)

10. Bielecki, T., Hernández-Hernández, D., Pliska, S.R.: Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. Math. Methods Oper. Res. **50**, 167–188 (1999)
11. Blackwell, D.: Discounted dynamic programming. Ann. Math. Stat. **36**, 226–235 (1965)
12. Borkar, V., Meyn, S.: Risk-sensitive optimal control for Markov decision processes with monotone cost. Math. Oper. Res. **27**, 192–209 (2002)
13. Dubins, L.E., Savage, L.J.: How to Gamble if You Must. Inequalities for Stochastic Processes. McGraw-Hill, New York (1965)
14. Dynkin, E.B., Yushkevich, A.A.: Controlled Markov Processes. Springer, Berlin (1979)
15. Enders, J., Powell, W., Egan, D.: A dynamic model for the failure replacement of aging high-voltage transformers. Energy Syst. J. **1**, 31–59 (2010)
16. Feinberg, E.A., Shwartz, A. (eds.): Handbook of Markov Decision Processes. Kluwer Academic, Boston (2002)
17. de Finetti, B.: Su unímpostazione alternativa della teoria collettiva del rischio. In: Transactions of the XVth International Congress of Actuaries, vol. 2, pp. 433–443 (1957)
18. Gittins, J.C.: Multi-armed Bandit Allocation Indices. Wiley, Chichester (1989)
19. Goto, J., Lewis, M., Puterman, M.: Coffee, tea or . . . ? A Markov decision process model for airline meal provisioning. Transp. Sci. **38**, 107–118 (2004)
20. Guo, X., Hernández-Lerma, O.: Continuous-time Markov Decision Processes. Springer, New York (2009)
21. He, M., Zhao, L., Powell, W.: Optimal control of dosage decisions in controlled ovarian hyperstimulation. Ann. Oper. Res. 223–245 (2010)
22. Hernández-Lerma, O., Lasserre, J.B.: Discrete-time Markov Control Processes. Springer, New York (1996)
23. Hernández-Lerma, O., Lasserre, J.B.: The linear programming approach. In: Handbook of Markov Decision Processes, pp. 377–408. Kluwer Acad., Boston (2002)
24. Hinderer, K.: Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Springer, Berlin (1970)
25. Howard, R.A.: Dynamic Programming and Markov Processes. The Technology Press of MIT, Cambridge (1960)
26. Kushner, H.J., Dupuis, P.: Numerical Methods for Stochastic Control Problems in Continuous Time. Springer, New York (2001)
27. Martin-Löf, A.: Lectures on the use of control theory in insurance. Scand. Actuar. J. **1**, 1–25 (1994)
28. Meyn, S.: Control Techniques for Complex Networks. Cambridge University Press, Cambridge (2008)
29. Miyasawa, K.: An economic survival game. Oper. Res. Soc. Jpn. **4**, 95–113 (1962)
30. Peskir, G., Shiryaev, A.: Optimal Stopping and Free-boundary Problems. Birkhäuser, Basel (2006)
31. Powell, W.: Approximate Dynamic Programming. Wiley-Interscience, Hoboken (2007)
32. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
33. Ross, S.: Introduction to Stochastic Dynamic Programming. Academic Press, New York (1983)
34. Schäl, M.: Markoffsche Entscheidungsprozesse. Teubner, Stuttgart (1990)
35. Schäl, M.: On discrete-time dynamic programming in insurance: exponential utility and minimizing the ruin probability. Scand. Actuar. J. 189–210 (2004)
36. Schmidli, H.: Stochastic Control in Insurance. Springer, London (2008)
37. Shapley, L.S.: Stochastic games. Proc. Natl. Acad. Sci. **39**, 1095–1100 (1953)
38. Shiryaev, A.N.: Some new results in the theory of controlled random processes. In: Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions Random Processes, Prague, pp. 131–203. Academia, Prague (1965)
39. Stokey, N.L., Lucas, E.E. Jr.: Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge (1989)
40. Tijms, H.: A First Course in Stochastic Models. Wiley, Chichester (2003)

Nicole Bäuerle is full professor for Stochastics at the Karlsruhe Institute of Technology. Currently she is in the board of the Fachgruppe Stochastik and the DGVFM (Deutsche Gesellschaft für Versicherungs- und Finanzmathematik). Her research areas include applied probability, stochastic processes and control as well as financial and actuarial mathematics. She is editor of the journals "Stochastic Models" and "Mathematical Methods of Operations Research".

Ulrich Rieder is full professor for Optimization and Operations Research at the University of Ulm since 1980. He helped to establish a new program in applied mathematics at Ulm, called Wirtschaftsmathematik. His research interests include applied probability, optimization and control of stochastic processes in operations research, finance and insurance. From 1990–2008 he was editor-in-chief of "Mathematical Methods of Operations Research". He is editor of several journals in the areas of operations research and finance.