de Gruyter Studies in Mathematics 27

de Gruyter Studies in Mathematics

1 Riemannian Geometry, 2nd rev. ed., *Wilhelm P. A. Klingenberg*
2 Semimartingales, *Michel Métivier*
3 Holomorphic Functions of Several Variables, *Ludger Kaup and Burchard Kaup*
4 Spaces of Measures, *Corneliu Constantinescu*
5 Knots, 2nd rev. and ext. ed., *Gerhard Burde and Heiner Zieschang*
6 Ergodic Theorems, *Ulrich Krengel*
7 Mathematical Theory of Statistics, *Helmut Strasser*
8 Transformation Groups, *Tammo tom Dieck*
9 Gibbs Measures and Phase Transitions, *Hans-Otto Georgii*
10 Analyticity in Infinite Dimensional Spaces, *Michel Hervé*
11 Elementary Geometry in Hyperbolic Space, *Werner Fenchel*
12 Transcendental Numbers, *Andrei B. Shidlovskii*
13 Ordinary Differential Equations, *Herbert Amann*
14 Dirichlet Forms and Analysis on Wiener Space, *Nicolas Bouleau and Francis Hirsch*
15 Nevanlinna Theory and Complex Differential Equations, *Ilpo Laine*
16 Rational Iteration, *Norbert Steinmetz*
17 Korovkin-type Approximation Theory and its Applications, *Francesco Altomare and Michele Campiti*
18 Quantum Invariants of Knots and 3-Manifolds, *Vladimir G. Turaev*
19 Dirichlet Forms and Symmetric Markov Processes, *Masatoshi Fukushima, Yoichi Oshima and Masayoshi Takeda*
20 Harmonic Analysis of Probability Measures on Hypergroups, *Walter R. Bloom and Herbert Heyer*
21 Potential Theory on Infinite-Dimensional Abelian Groups, *Alexander Bendikov*
22 Methods of Noncommutative Analysis, *Vladimir E. Nazaikinskii, Victor E. Shatalov and Boris Yu. Sternin*
23 Probability Theory, *Heinz Bauer*
24 Variational Methods for Potential Operator Equations, *Jan Chabrowski*
25 The Structure of Compact Groups, *Karl H. Hofmann and Sidney A. Morris*
26 Measure and Integration Theory, *Heinz Bauer*
27 Stochastic Finance, 2nd rev. and ext. ed., *Hans Föllmer and Alexander Schied*
28 Painlevé Differential Equations in the Complex Plane, *Valerii I. Gromak, Ilpo Laine and Shun Shimomura*
29 Discontinuous Groups of Isometries in the Hyperbolic Plane, *Werner Fenchel and Jakob Nielsen*
30 The Reidemeister Torsion of 3-Manifolds, *Liviu I. Nicolaescu*
31 Elliptic Curves, *Susanne Schmitt and Horst G. Zimmer*

Hans Föllmer · Alexander Schied

Stochastic Finance

An Introduction in Discrete Time

Second Revised and Extended Edition

Walter de Gruyter
Berlin · New York

Authors

Hans Föllmer
Institut für Mathematik
Humboldt Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

Alexander Schied
Institut für Mathematik, MA 7−4
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
Germany

Mathematics Subject Classification 2000: Primary: 60-01, 91-01, 91-02; secondary: 46N10, 60E15, 60G40, 60G42, 91B08, 91B16, 91B28, 91B30, 91B50, 91B52, 91B70

Keywords: Martingales, arbitrage, contingent claims, options, hedging, preferences, optimization, equilibrium, value at risk, risk measures

♾ Printed on acid-free paper which falls within the guidelines of the ANSI to ensure permanence and durability.

Library of Congress − Cataloging-in-Publication Data

Föllmer, Hans.
Stochastic finance : an introduction in discrete time / by Hans Föllmer, Alexander Schied.− 2nd rev. and extended ed.
p. cm. − (De Gruyter studies in mathematics ; 27)
Includes bibliographical references and index.
ISBN 3-11-018346-3 (Cloth : alk. paper)
1. Finance − Statistical methods. 2. Stochastic analysis. 3. Probabilities. I. Schied, Alexander. II. Title. III. Series.
HG176.5.F65 2004
332'.01'519232−dc22 004021608

ISBN 3-11-018346-3

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Printed in Germany.
Cover design: Rudolf Hübler, Berlin.
Typeset using the authors' TEX files: I. Zimmermann, Freiburg.
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen.

Preface to the second edition

Since the publication of the first edition we have used it as the basis for several courses. These include courses for a whole semester on Mathematical Finance in Berlin and also short courses on special topics such as risk measures given at the Institut Henri Poincaré in Paris, at the Department of Operations Research at Cornell University, at the Academia Sinica in Taipei, and at the 8th Symposium on Probability and Stochastic Processes in Puebla. In the process we have made a large number of minor corrections, we have discovered many opportunities for simplification and clarification, and we have also learned more about several topics. As a result, major parts of this book have been improved or even entirely rewritten. Among them are those on robust representations of risk measures, arbitrage-free pricing of contingent claims, exotic derivatives in the CRR model, convergence to Black–Scholes prices, and stability under pasting with its connections to dynamically consistent coherent risk measures. In addition, this second edition contains several new sections, including a systematic discussion of law-invariant risk measures, of concave distortions, and of the relations between risk measures and Choquet integration.

It is a pleasure to express our thanks to all students and colleagues whose comments have helped us to prepare this second edition, in particular to Dirk Becherer, Hans Bühler, Rose-Anne Dana, Ulrich Horst, Mesrop Janunts, Christoph Kühn, Maren Liese, Harald Luschgy, Holger Pint, Philip Protter, Lothar Rogge, Stephan Sturm, Stefan Weber, Wiebke Wittmüß, and Ching-Tang Wu. Special thanks are due to Peter Bank and to Yuliya Mishura and Georgiy Shevchenko, our translators for the Russian edition. Finally, we thank Irene Zimmermann and Manfred Karbe of de Gruyter Verlag for urging us to write a second edition and for their efficient support.

Berlin, September 2004

Hans Föllmer
Alexander Schied

Preface to the first edition

This book is an introduction to probabilistic methods in Finance. It is intended for graduate students in mathematics, and it may also be useful for mathematicians in academia and in the financial industry. Our focus is on stochastic models in discrete time. This limitation has two immediate benefits. First, the probabilistic machinery is simpler, and we can discuss right away some of the key problems in the theory of pricing and hedging of financial derivatives. Second, the paradigm of a complete financial market, where all derivatives admit a perfect hedge, becomes the exception rather than the rule. Thus, the discrete-time setting provides a shortcut to some of the more recent literature on incomplete financial market models.

As a textbook for mathematicians, it is an introduction at an intermediate level, with special emphasis on martingale methods. Since it does not use the continuous-time methods of Itô calculus, it needs less preparation than more advanced texts such as [73], [74], [82], [129], [188]. On the other hand, it is technically more demanding than textbooks such as [160]: We work on general probability spaces, and so the text captures the interplay between probability theory and functional analysis which has been crucial for some of the recent advances in mathematical finance.

The book is based on our notes for first courses in Mathematical Finance which both of us are teaching in Berlin at Humboldt University and at Technical University. These courses are designed for students in mathematics with some background in probability. Sometimes, they are given in parallel to a systematic course on stochastic processes. At other times, martingale methods in discrete time are developed in the course, as they are in this book. Usually the course is followed by a second course on Mathematical Finance in continuous time. There it turns out to be useful that students are already familiar with some of the key ideas of Mathematical Finance.

The core of this book is the dynamic arbitrage theory in the first chapters of Part II. When teaching a course, we found it useful to explain some of the main arguments in the more transparent one-period model before using them in the dynamical setting. So one approach would be to start immediately in the multi-period framework of Chapter 5, and to go back to selected sections of Part I as the need arises. As an alternative, one could first focus on the one-period model, and then move on to Part II.

We include in Chapter 2 a brief introduction to the mathematical theory of expected utility, even though this is a classical topic, and there is no shortage of excellent expositions; see, for instance, [138] which happens to be our favorite. We have three reasons for including this chapter. Our focus in this book is on incompleteness, and incompleteness involves, in one form or another, preferences in the face of risk and uncertainty. We feel that mathematicians working in this area should be aware, at least to some extent, of the long line of thought which leads from Daniel Bernoulli via von Neumann–Morgenstern and Savage to some more recent developments which are motivated by shortcomings of the classical paradigm. This is our first reason. Second, the analysis of risk measures has emerged as a major topic in mathematical finance, and this is closely related to a robust version of the Savage theory. Third, but not least, our experience is that this part of the course was found particularly enjoyable, both by the students and by ourselves.

We acknowledge our debt and express our thanks to all colleagues who have contributed, directly or indirectly, through their publications and through informal discussions, to our understanding of the topics discussed in this book. Ideas and methods developed by Freddy Delbaen, Darrell Duffie, Nicole El Karoui, David Heath, Yuri Kabanov, Ioannis Karatzas, Dimitri Kramkov, David Kreps, Stanley Pliska, Chris Rogers, Steve Ross, Walter Schachermayer, Martin Schweizer, Dieter Sondermann and Christophe Stricker play a key role in our exposition. We are obliged to many others; for instance the textbooks [54], [73], [74], [116], and [143] were a great help when we started to teach courses on the subject.

We are grateful to all those who read parts of the manuscript and made useful suggestions, in particular to Dirk Becherer, Ulrich Horst, Steffen Krüger, Irina Penner, and to Alexander Giese who designed some of the figures. Special thanks are due to Peter Bank for a large number of constructive comments. We also express our thanks to Erhan Çinlar, Adam Monahan, and Philip Protter for improving some of the language, and to the Department of Operations Research and Financial Engineering at Princeton University for its hospitality during the weeks when we finished the manuscript.

Berlin, June 2002

Hans Föllmer
Alexander Schied

Contents

Part I

Mathematical finance in one period

Chapter 1

Arbitrage theory

In this chapter, we study the mathematical structure of a simple one-period model of a financial market. We consider a finite number of assets. Their initial prices at time $t = 0$ are known, their future prices at time $t = 1$ are described as random variables on some probability space. Trading takes place at time $t = 0$. Already in this simple model, some basic principles of mathematical finance appear very clearly. In Section 1.2, we single out those models which satisfy a condition of *market efficiency*: There are no trading opportunities which yield a profit without any downside risk. The absence of such *arbitrage opportunities* is characterized by the existence of an equivalent *martingale measure*. Under such a measure, discounted prices have the martingale property, that is, trading in the assets is the same as playing a fair game. As explained in Section 1.3, any equivalent martingale measure can be identified with a pricing rule: It extends the given prices of the primary assets to a larger space of *contingent claims*, or financial *derivatives*, without creating new arbitrage opportunities. In general, there will be several such extensions. A given contingent claim has a unique price if and only if it admits a *perfect hedge*. In our one-period model, this will be the exception rather than the rule. Thus, we are facing *market incompleteness*, unless our model satisfies the very restrictive conditions discussed in Section 1.4. The geometric structure of an arbitrage-free model is described in Section 1.5.

The one-period market model will be used throughout the first part of this book. On the one hand, its structure is rich enough to illustrate some of the key ideas of the field. On the other hand, it will provide an introduction to some of the mathematical methods which will be used in the dynamic hedging theory of the second part. In fact, the multi-period situation considered in Chapter 5 can be regarded as a sequence of one-period models whose initial conditions are contingent on the outcomes of previous periods. The techniques for dealing with such contingent initial data are introduced in Section 1.6.

1.1 Assets, portfolios, and arbitrage opportunities

Consider a financial market with $d + 1$ assets. The assets can consist, for instance, of equities, bonds, commodities, or currencies. In a simple one-period model, these assets are priced at the initial time $t = 0$ and at the final time $t = 1$. We assume that the i^{th} asset is available at time 0 for a price $\pi^i \geq 0$. The collection

$$\overline{\pi} = (\pi^0, \pi^1, \dots, \pi^d) \in \mathbb{R}_+^{d+1}$$

is called a *price system*. Prices at time 1 are usually not known beforehand at time 0. In order to model this uncertainty, we fix a probability space $(\Omega, \mathcal{F}, P)$ and describe the asset prices at time 1 as non-negative measurable functions

$$S^0, S^1, \dots, S^d$$

on $(\Omega, \mathcal{F})$ with values in $[0, \infty)$. Every $\omega \in \Omega$ corresponds to a particular scenario of market evolution, and $S^i(\omega)$ is the price of the i^{th} asset at time 1 if the scenario ω occurs.

However, not all asset prices in a market are necessarily uncertain. Usually there is a riskless bond which will pay a *sure* amount at time 1. In our simple model for one period, such a riskless investment opportunity will be included by assuming that

$$\pi^0 = 1 \quad \text{and} \quad S^0 \equiv 1 + r$$

for a constant r, the return of a unit investment into the riskless bond. In most situations it would be natural to assume $r \geq 0$, but for our purposes it is enough to require that $S^0 > 0$, or equivalently that

$$r > -1.$$

In order to distinguish S^0 from the risky assets $S^1, \dots, S^d$, it will be convenient to use the notation

$$\overline{S} = (S^0, S^1, \dots, S^d) = (S^0, S),$$

and in the same way we will write $\overline{\pi} = (1, \pi)$.

At time $t = 0$, an investor will choose a *portfolio*

$$\overline{\xi} = (\xi^0, \xi) = (\xi^0, \xi^1, \dots, \xi^d) \in \mathbb{R}^{d+1},$$

where ξ^i represents the number of shares of the i^{th} asset. The price for buying the portfolio $\overline{\xi}$ equals

$$\overline{\pi} \cdot \overline{\xi} = \sum_{i=0}^{d} \pi^i \xi^i.$$

At time $t = 1$, the portfolio will have the value

$$\overline{\xi} \cdot \overline{S}(\omega) = \sum_{i=0}^{d} \xi^i S^i(\omega) = \xi^0(1 + r) + \xi \cdot S(\omega),$$

depending on the scenario $\omega \in \Omega$. Here we assume implicitly that buying and selling assets does not create extra costs, an assumption which may not be valid for a small investor but which becomes more realistic for a large financial institution. Note our convention of writing $x \cdot y$ for the inner product of two vectors x and y in Euclidean space.

Our definition of a portfolio allows the components ξ^i to be negative. If $\xi^0 < 0$, this corresponds to taking out a loan such that we receive the amount $|\xi^0|$ at $t = 0$

and pay back the amount $(1+r)|\xi^0|$ at time $t=1$. If $\xi^i<0$ for $i\geq 1$, a quantity of $|\xi^i|$ shares of the i^{th} asset is sold without actually owning them. This corresponds to a *short sale* of the asset. In particular, an investor is allowed to take a short position $\xi^i<0$, and to use up the received amount $\pi^i|\xi^i|$ for buying quantities $\xi^j\geq 0$, $j\neq i$, of the other assets. In this case, the price of the portfolio $\overline{\xi}=(\xi^0,\xi)$ is given by $\overline{\xi}\cdot\overline{\pi}=0$.

Definition 1.1. A portfolio $\overline{\xi}\in\mathbb{R}^{d+1}$ is called an *arbitrage opportunity* if $\overline{\pi}\cdot\overline{\xi}\leq 0$ but $\overline{\xi}\cdot\overline{S}\geq 0$ P-a.s. and $P[\,\overline{\xi}\cdot\overline{S}>0\,]>0$.

Intuitively, an arbitrage opportunity is an investment strategy that yields with positive probability a positive profit and is not exposed to any downside risk. The existence of such an arbitrage opportunity may be regarded as a market inefficiency in the sense that certain assets are not priced in a reasonable way. In real-world markets, arbitrage opportunities are rather hard to find. If such an opportunity would show up, it would generate a large demand, prices would adjust, and the opportunity would disappear. Later on, the absence of such arbitrage opportunities will be our key assumption. Absence of arbitrage implies that S^i vanishes P-a.s. once $\pi^i=0$. Hence, there is no loss in generality if we assume from now on that

$$\pi^i>0 \quad \text{for } i=1,\dots,d.$$

Remark 1.2. Note that the probability measure P enters the definition of an arbitrage opportunity only through the null sets of P. In particular, the definition can be formulated without any explicit use of probabilities if Ω is countable. In this case there is no loss of generality in assuming that the underlying probability measure satisfies $P[\{\omega\}]>0$ for every $\omega\in\Omega$. Then an arbitrage opportunity is simply a portfolio $\overline{\xi}$ with $\overline{\pi}\cdot\overline{\xi}\leq 0$, with $\overline{\xi}\cdot\overline{S}(\omega)\geq 0$ for all $\omega\in\Omega$, and such that $\overline{\xi}\cdot\overline{S}(\omega_0)>0$ for at least one $\omega_0\in\Omega$. ◇

The following lemma shows that absence of arbitrage is equivalent to the following property of the market: Any investment in risky assets which yields with positive probability a better result than investing the same amount in the risk-free asset must be open to some downside risk.

Lemma 1.3. *The following statements are equivalent.*

(a) *The market model admits an arbitrage opportunity.*

(b) *There is a vector $\xi\in\mathbb{R}^d$ such that*

$$\xi\cdot S\geq(1+r)\xi\cdot\pi \;\; P\text{-a.s.} \quad \text{and} \quad P[\,\xi\cdot S>(1+r)\xi\cdot\pi\,]>0.$$

Proof. To see that (a) implies (b), let $\overline{\xi}$ be an arbitrage opportunity. Then $0\geq\overline{\xi}\cdot\overline{\pi}=\xi^0+\xi\cdot\pi$. Hence,

$$\xi\cdot S-(1+r)\xi\cdot\pi\geq\xi\cdot S+(1+r)\xi^0=\overline{\xi}\cdot\overline{S}.$$

Since $\overline{\xi} \cdot \overline{S}$ is P-a.s. non-negative and strictly positive with non-vanishing probability, the same must be true of $\xi \cdot S - (1+r)\xi \cdot \pi$.

Next let ξ be as in (b). We claim that the portfolio (ξ^0, ξ) with $\xi^0 := -\xi \cdot \pi$ is an arbitrage opportunity. Indeed, $\overline{\xi} \cdot \overline{\pi} = \xi^0 + \xi \cdot \pi = 0$ by definition. Moreover, $\overline{\xi} \cdot \overline{S} = -(1+r)\xi \cdot \pi + \xi \cdot S$, which is P-a.s. non-negative and strictly positive with non-vanishing probability. □

1.2 Absence of arbitrage and martingale measures

In this section, we are going to characterize those market models which do not admit any arbitrage opportunities. Such models will be called *arbitrage-free*.

Definition 1.4. A probability measure P^* is called a *risk-neutral measure*, or a *martingale measure*, if

$$\pi^i = E^*\left[\frac{S^i}{1+r} \right], \quad i = 0, 1, \dots, d. \tag{1.1}$$

Remark 1.5. In (1.1), the price of the i^{th} asset is identified as the expectation of the discounted payoff under the measure P^*. Thus, the pricing formula (1.1) can be seen as a classical valuation formula which does not take into account any risk aversion, in contrast to valuations in terms of expected utility which will be discussed in Section 2.3. This is why a measure P^* satisfying (1.1) is called risk-neutral. The connection to martingales will be made explicit in Section 1.6. ◇

The following basic result is sometimes called the "fundamental theorem of asset pricing" or, in short, FTAP. It characterizes arbitrage-free market models in terms of the set

$$\mathcal{P} := \left\{ P^* \mid P^* \text{ is a risk-neutral measure with } P^* \approx P \right\}$$

of risk-neutral measures which are *equivalent* to P. Recall that two probability measures P^* and P are said to be equivalent ($P^* \approx P$) if, for $A \in \mathcal{F}$, $P^*[\,A\,] = 0$ if and only if $P[\,A\,] = 0$. This holds if and only if P^* has a strictly positive density dP^*/dP with respect to P; see Appendix A.2. An equivalent risk-neutral measure is also called a *pricing measure* or an *equivalent martingale measure*.

Theorem 1.6. *A market model is arbitrage-free if and only if $\mathcal{P} \neq \emptyset$. In this case, there exists a $P^* \in \mathcal{P}$ which has a bounded density dP^*/dP.*

We show first that the existence of a risk-neutral measure implies the absence of arbitrage.

Proof of the implication $\Leftarrow$ of Theorem 1.6. Suppose that there exists a risk-neutral measure $P^* \in \mathcal{P}$. Take a portfolio $\overline{\xi} \in \mathbb{R}^{d+1}$ such that $\overline{\xi} \cdot \overline{S} \geq 0$ P-a.s. and

$E[\,\overline{\xi}\cdot\overline{S}\,] > 0$. Both properties remain valid if we replace P by the equivalent measure P^*. Hence,

$$\overline{\pi}\cdot\overline{\xi} = \sum_{i=0}^{d}\pi^i\xi^i = \sum_{i=0}^{d}E^*\Big[\frac{\xi^i S^i}{1+r}\Big] = E^*\Big[\frac{\overline{\xi}\cdot\overline{S}}{1+r}\Big] > 0.$$

Thus, $\overline{\xi}$ cannot be an arbitrage opportunity. □

For the proof of the implication $\Rightarrow$ of Theorem 1.6, it will be convenient to introduce the random vector $Y = (Y^1, \dots, Y^d)$ of *discounted net gains*:

$$Y^i := \frac{S^i}{1+r} - \pi^i, \quad i = 1, \dots, d. \tag{1.2}$$

With this notation, Lemma 1.3 implies that the absence of arbitrage is equivalent to the following condition:

$$\text{For } \xi \in \mathbb{R}^d\text{:} \quad \xi\cdot Y \geq 0 \ P\text{-a.s.} \implies \xi\cdot Y = 0 \ P\text{-a.s.} \tag{1.3}$$

Since Y^i is bounded from below by $-\pi^i$, the expectation $E^*[\,Y^i\,]$ of Y^i under any measure P^* is well-defined, and so P^* is a risk-neutral measure if and only if

$$E^*[\,Y\,] = 0. \tag{1.4}$$

Here, $E^*[\,Y\,]$ is a shorthand notation for the d-dimensional vector with components $E^*[\,Y^i\,]$, $i = 1, \dots, d$. The assertion of Theorem 1.6 can now be read as follows: Condition (1.3) holds if and only if there exists some $P^* \approx P$ such that $E^*[\,Y\,] = 0$, and in this case, P^* can be chosen such that the density dP^*/dP is bounded.

Proof of the implication $\Rightarrow$ of Theorem 1.6. We have to show that (1.3) implies the existence of some $P^* \approx P$ such that (1.4) holds and such that the density dP^*/dP is bounded. We will do this first in the case in which

$$E[\,|Y|\,] < \infty.$$

Let $\mathcal{Q}$ denote the convex set of all probability measures $Q \approx P$ with bounded densities dQ/dP, and denote by $E_Q[\,Y\,]$ the d-dimensional vector with components $E_Q[\,Y^i\,]$, $i = 1, \dots, d$. Due to our assumption that $|Y| \in \mathcal{L}^1(P)$, all these expectations are finite. Let

$$\mathcal{C} := \{\, E_Q[\,Y\,] \mid Q \in \mathcal{Q}\,\},$$

and note that $\mathcal{C}$ is a convex set in $\mathbb{R}^d$: If $Q_1, Q_0 \in \mathcal{Q}$ and $0 \leq \alpha \leq 1$, then $Q_\alpha := \alpha Q_1 + (1-\alpha)Q_0 \in \mathcal{Q}$ and

$$\alpha E_{Q_1}[\,Y\,] + (1-\alpha)E_{Q_0}[\,Y\,] = E_{Q_\alpha}[\,Y\,],$$

which lies in $\mathcal{C}$.

Our aim is to show that $\mathcal{C}$ contains the origin. To this end, we suppose by way of contradiction that $0 \notin \mathcal{C}$. Using the "separating hyperplane theorem" in the elementary form of Proposition A.1, we obtain a vector $\xi \in \mathbb{R}^d$ such that $\xi \cdot x \geq 0$ for all $x \in \mathcal{C}$, and such that $\xi \cdot x_0 > 0$ for some $x_0 \in \mathcal{C}$. Thus, ξ satisfies $E_Q[\,\xi \cdot Y\,] \geq 0$ for all $Q \in \mathcal{Q}$ and $E_{Q_0}[\,\xi \cdot Y\,] > 0$ for some $Q_0 \in \mathcal{Q}$. Clearly, the latter condition yields that $P[\,\xi \cdot Y > 0\,] > 0$. We claim that the first condition implies that $\xi \cdot Y$ is P-a.s. non-negative. This fact will be a contradiction to our assumption (1.3) and thus will prove that $0 \in \mathcal{C}$.

To prove the claim that $\xi \cdot Y \geq 0$ P-a.s., let $A := \{\,\xi \cdot Y < 0\}$, and define functions

$$\varphi_n := \left(1 - \frac{1}{n}\right) \cdot \mathrm{I}_A + \frac{1}{n} \cdot \mathrm{I}_{A^c}.$$

We take φ_n as densities for new probability measures Q_n:

$$\frac{dQ_n}{dP} := \frac{1}{E[\,\varphi_n\,]} \cdot \varphi_n, \quad n = 2, 3, \dots .$$

Since $0 < \varphi_n \leq 1$, it follows that $Q_n \in \mathcal{Q}$, and thus that

$$0 \leq \xi \cdot E_{Q_n}[\,Y\,] = \frac{1}{E[\,\varphi_n\,]} E[\,\xi \cdot Y\,\varphi_n\,].$$

Hence, Lebesgue's dominated convergence theorem yields that

$$E\big[\,\xi \cdot Y\,\mathrm{I}_{\{\,\xi \cdot Y < 0\,\}}\,\big] = \lim_{n \uparrow \infty} E[\,\xi \cdot Y\,\varphi_n\,] \geq 0.$$

This proves the claim that $\xi \cdot Y \geq 0$ P-a.s. and completes the proof of Theorem 1.6 in case $E[\,|Y|\,] < \infty$.

If Y is not P-integrable, then we simply replace the probability measure P by a suitable equivalent measure $\widetilde{P}$ whose density $d\widetilde{P}/dP$ is bounded and for which $\widetilde{E}[\,|Y|\,] < \infty$. For instance, one can define $\widetilde{P}$ by

$$\frac{d\widetilde{P}}{dP} = \frac{c}{1 + |Y|} \quad \text{for } c := \left(E\left[\,\frac{1}{1 + |Y|}\,\right]\right)^{-1}.$$

Recall from Remark 1.2 that replacing P with an equivalent probability measure does not affect the absence of arbitrage opportunities in our market model. Thus, the first part of this proof yields a risk-neutral measure P^* which is equivalent to $\widetilde{P}$ and whose density $dP^*/d\widetilde{P}$ is bounded. Then $P^* \in \mathcal{P}$, and

$$\frac{dP^*}{dP} = \frac{dP^*}{d\widetilde{P}} \cdot \frac{d\widetilde{P}}{dP}$$

is bounded. Hence, P^* is as desired, and the theorem is proved. □

Remark 1.7. Our assumption that asset prices are non-negative implies that the components of Y are bounded from below. Note however that this assumption was not needed in our proof. Thus, Theorem 1.6 also holds if we only assume that S is finite-valued and $\pi \in \mathbb{R}^d$. In this case, the definition of a risk-neutral measure P^* via (1.1) is meant to include the assumption that S^i is integrable with respect to P^* for $i = 1, \dots, d$. ◇

Example 1.8. Let P be any probability measure on the finite set $\Omega := \{ \omega_1, \dots, \omega_N \}$ that assigns strictly positive probability p_i to each singleton $\{ \omega_i \}$. Suppose that there is a single risky asset defined by its price $\pi = \pi^1$ at time 0 and by the random variable $S = S^1$. We may assume without loss of generality that the values $s_i := S(\omega_i)$ are distinct and arranged in increasing order: $s_1 < \dots < s_N$. According to Theorem 1.6, this model does not admit arbitrage opportunities if and only if

$$\pi(1+r) \in \{ \widetilde{E}[\, S \,] \mid \widetilde{P} \approx P \} = \Big\{ \sum_{i=1}^N s_i \widetilde{p}_i \;\Big|\; \widetilde{p}_i > 0,\ \sum_{i=1}^N \widetilde{p}_i = 1 \Big\} = (s_1, s_N),$$

and P^* is a risk-neutral measure if and only if the probabilities $p_i^* := P^*[\,\{\omega_i\}\,]$ solve the linear equations

$$\begin{aligned} s_1 p_1^* + \dots + s_N p_N^* &= \pi(1+r), \\ p_1^* + \dots + p_N^* &= 1. \end{aligned}$$

If a solution exists, it will be unique if and only if $N = 2$, and there will be infinitely many solutions for $N > 2$. ◇

Remark 1.9. The economic reason for working with the *discounted* asset prices

$$X^i := \frac{S^i}{1+r}, \quad i = 0, \dots, d, \tag{1.5}$$

is that one should distinguish between one unit of a currency (e.g. €) at time $t = 0$ and one unit at time $t = 1$. Usually people tend to prefer a certain amount today over the same amount which is promised to be paid at a later time. Such a preference is reflected in an interest $r > 0$ paid by the riskless bond: Only the amount $1/(1+r)$ € must be invested at time 0 to obtain 1 € at time 1. This effect is sometimes referred to as the *time value of money*. Similarly, the price S^i of the i^{th} asset is quoted in terms of € at time 1, while π^i corresponds to time-zero euros. Thus, in order to compare the two prices π^i and S^i, one should first convert them to a common standard. This is achieved by taking the riskless bond as a *numéraire* and by considering the *discounted* prices in (1.5). ◇

Remark 1.10. One can choose as numéraire any asset which is strictly positive. For instance, suppose that $\pi^1 > 0$ and $P[\, S^1 > 0 \,] = 1$. Then all asset prices can be expressed in units of the first asset by considering

$$\widetilde{\pi}^i := \frac{\pi^i}{\pi^1} \quad \text{and} \quad \frac{S^i}{S^1}, \quad i = 0, \dots, d.$$

Clearly, the definition of an arbitrage opportunity is independent of the choice of a particular numéraire. Thus, an arbitrage-free market model should admit a risk-neutral measure with respect to the new numéraire, i.e., a probability measure $\widetilde{P}^* \approx P$ such that

$$\widetilde{\pi}^i = \widetilde{E}^*\Big[\frac{S^i}{S^1}\Big], \quad i = 0, \dots, d.$$

Let us denote by $\widetilde{\mathcal{P}}$ the set of all such measures $\widetilde{P}^*$. Then

$$\widetilde{\mathcal{P}} = \Big\{ \widetilde{P}^* \,\Big|\, \frac{d\widetilde{P}^*}{dP} = \frac{S^1}{E^*[\,S^1\,]} \text{ for some } P^* \in \mathcal{P} \Big\}.$$

Indeed, if $\widetilde{P}^*$ lies in the set on the right, then

$$\widetilde{E}^*\Big[\frac{S^i}{S^1}\Big] = \frac{E^*[\,S^i\,]}{E^*[\,S^1\,]} = \frac{\pi^i}{\pi^1} = \widetilde{\pi}^i,$$

and so $\widetilde{P}^* \in \widetilde{\mathcal{P}}$. Reversing the roles of $\widetilde{\mathcal{P}}$ and $\mathcal{P}$ then yields the identity of the two sets. Note that

$$\mathcal{P} \cap \widetilde{\mathcal{P}} = \emptyset$$

as soon as S^1 is not P-a.s. constant, because Jensen's inequality then implies that

$$\frac{1}{\pi^1} = \widetilde{\pi}^0 = \widetilde{E}^*\Big[\frac{1+r}{S^1}\Big] > \frac{1+r}{\widetilde{E}^*[\,S^1\,]}$$

and hence $\widetilde{E}^*[\,S^1\,] > E^*[\,S^1\,]$ for all $\widetilde{P}^* \in \widetilde{\mathcal{P}}$ and $P^* \in \mathcal{P}$. ◇

Let

$$\mathcal{V} := \big\{ \overline{\xi} \cdot \overline{S} \mid \overline{\xi} \in \mathbb{R}^{d+1} \big\}$$

denote the linear space of all payoffs which can be generated by some portfolio. An element of $\mathcal{V}$ will be called an *attainable payoff*. The portfolio that generates $V \in \mathcal{V}$ is in general not unique, but we have the following *law of one price*.

Lemma 1.11. *Suppose that the market model is arbitrage-free and that $V \in \mathcal{V}$ can be written as $V = \overline{\xi} \cdot \overline{S} = \overline{\zeta} \cdot \overline{S}$ P-a.s. for two different portfolios $\overline{\xi}$ and $\overline{\zeta}$. Then $\overline{\pi} \cdot \overline{\xi} = \overline{\pi} \cdot \overline{\zeta}$.*

Proof. We have $(\overline{\xi} - \overline{\zeta}) \cdot \overline{S} = 0$ P^*-a.s. for any $P^* \in \mathcal{P}$. Hence,

$$\overline{\pi} \cdot \overline{\xi} - \overline{\pi} \cdot \overline{\zeta} = E^*\Big[\frac{(\overline{\xi} - \overline{\zeta}) \cdot \overline{S}}{1+r}\Big] = 0,$$

due to (1.1). □

By the preceding lemma, it makes sense to define the *price* of $V \in \mathcal{V}$ as

$$\pi(V) := \overline{\pi} \cdot \overline{\xi} \quad \text{if } V = \overline{\xi} \cdot \overline{S}, \tag{1.6}$$

whenever the market model is arbitrage-free.

Remark 1.12. Via (1.6), the price system π can be regarded as a linear form on the finite-dimensional vector space $\mathcal{V}$. For any $P^* \in \mathcal{P}$ we have

$$\pi(V) = E^*\Big[\frac{V}{1+r}\Big], \qquad V \in \mathcal{V}.$$

Thus, an equivalent risk-neutral measure P^* defines a linear extension of π onto the larger space $\mathcal{L}^1(P^*)$ of P^*-integrable random variables. Since this space is usually infinite-dimensional, one cannot expect that such a pricing measure is in general unique; see however Section 1.4. ◇

We have seen above that, in an arbitrage-free market model, the condition $\overline{\xi} \cdot \overline{S} = 0$ P-a.s. implies that $\overline{\pi} \cdot \overline{\xi} = 0$. In fact, one may assume without loss of generality that

$$\overline{\xi} \cdot \overline{S} = 0 \ P\text{-a.s.} \quad \Longrightarrow \quad \overline{\xi} = 0, \tag{1.7}$$

for otherwise we can find $i \in \{0, \dots, d\}$ such that $\xi^i \neq 0$ and represent the i^{th} asset as a linear combination of the remaining ones:

$$\pi^i = \frac{1}{\xi^i}\sum_{j \neq i} \xi^j \pi^j \quad \text{and} \quad S^i = \frac{1}{\xi^i}\sum_{j \neq i} \xi^j S^j.$$

In this sense, the i^{th} asset is redundant and can be omitted.

Definition 1.13. The market model is called *non-redundant* if (1.7) holds.

Remark 1.14. In any non-redundant market model, the components of the vector Y of discounted net gains are linearly independent in the sense that

$$\xi \cdot Y = 0 \ P\text{-a.s.} \quad \Longrightarrow \quad \xi = 0. \tag{1.8}$$

Conversely, via (1.3), condition (1.8) implies non-redundance if the market model is arbitrage-free. ◇

Definition 1.15. Suppose that the market model is arbitrage-free and that $V \in \mathcal{V}$ is an attainable payoff such that $\pi(V) \neq 0$. Then the *return* of V is defined by

$$R(V) := \frac{V - \pi(V)}{\pi(V)}.$$

Note that we have already seen the special case of the risk-free return

$$r = \frac{S^0 - \pi^0}{\pi^0} = R(S^0).$$

If an attainable payoff V is a linear combination $V = \sum_{k=1}^n \alpha_k V_k$ of non-zero attainable payoffs V_k, then

$$R(V) = \sum_{k=1}^n \beta_k R(V_k) \quad \text{for } \beta_k = \frac{\alpha_k \pi(V_k)}{\sum_{i=1}^n \alpha_i \pi(V_i)}.$$

The coefficient β_k can be interpreted as the proportion of the investment allocated to V_k. As a particular case of the formula above, we have that

$$R(V) = \sum_{i=0}^{d} \frac{\pi^i \, \xi^i}{\overline{\pi} \cdot \overline{\xi}} \cdot R(S^i)$$

for all non-zero attainable payoffs $V = \overline{\xi} \cdot \overline{S}$ (recall that we have assumed that all π^i are strictly positive).

Proposition 1.16. *Suppose that the market model is arbitrage-free, and let $V \in \mathcal{V}$ be an attainable payoff such that $\pi(V) \neq 0$.*

(a) *Under any risk-neutral measure P^*, the expected return of V is equal to the risk-free return r:*
$$E^*[\, R(V)\,] = r.$$

(b) *Under any measure $Q \approx P$ such that $E_Q[\, |\overline{S}|\,] < \infty$, the expected return of V is given by*
$$E_Q[\, R(V)\,] = r - \mathrm{cov}_Q\left(\frac{dP^*}{dQ}, R(V)\right),$$
where P^* is an arbitrary risk-neutral measure in $\mathcal{P}$ and cov_Q denotes the covariance with respect to Q.

Proof. (a): Since $E^*[\, V\,] = \pi(V)(1+r)$, we have

$$E^*[\, R(V)\,] = \frac{E^*[\, V\,] - \pi(V)}{\pi(V)} = r.$$

(b): Let $P^* \in \mathcal{P}$ and $\varphi^* := dP^*/dQ$. Then

$$\begin{aligned}\mathrm{cov}_Q\left(\varphi^*, R(V)\right) &= E_Q[\, \varphi^* R(V)\,] - E_Q[\, \varphi^*\,] \cdot E_Q[\, R(V)\,] \\ &= E^*[\, R(V)\,] - E_Q[\, R(V)\,].\end{aligned}$$

Using part (a) yields the assertion. □

Remark 1.17. Let us comment on the extension of the fundamental equivalence in Theorem 1.6 to market models with an infinity of tradable assets $S^0, S^1, S^2, \dots$. We assume that $S^0 \equiv 1 + r$ for some $r > -1$ and that the random vector

$$S(\omega) = \left(S^1(\omega), S^2(\omega), \dots\right)$$

takes values in the space ℓ^∞ of bounded real sequences. This space is a Banach space with respect to the norm

$$\|x\|_\infty := \sup_{i \geq 1} |x^i| \quad \text{for } x = (x^1, x^2, \dots) \in \ell^\infty.$$

A portfolio $\overline{\xi} = (\xi^0, \xi)$ is chosen in such a way that $\xi = (\xi^1, \xi^2, \dots)$ is a sequence in the space ℓ^1, i.e., $\sum_{i=1}^{\infty} |\xi^i| < \infty$. We assume that the corresponding price system $\overline{\pi} = (\pi^0, \pi)$ satisfies $\pi \in \ell^\infty$ and $\pi^0 = 1$. Clearly, this model class includes our model with $d+1$ traded assets as a special case.

Our first observation is that the implication $\Leftarrow$ of Theorem 1.6 remains valid, i.e., the existence of a measure $P^* \approx P$ with the properties

$$E^*[\, \|S\|_\infty \,] < \infty \quad \text{and} \quad E^*\left[\frac{S^i}{1+r} \right] = \pi^i$$

implies the absence of arbitrage opportunities. To this end, suppose that $\overline{\xi}$ is a portfolio strategy such that

$$\overline{\xi} \cdot \overline{S} \geq 0 \ \ P\text{-a.s.} \quad \text{and} \quad E[\, \overline{\xi} \cdot \overline{S} \,] > 0. \tag{1.9}$$

Then we can replace P in (1.9) by the equivalent measure P^*. Hence, $\overline{\xi}$ cannot be an arbitrage opportunity since

$$\overline{\xi} \cdot \overline{\pi} = \sum_{i=0}^{\infty} \xi^i E^*\left[\frac{S^i}{1+r} \right] = E^*\left[\frac{\overline{\xi} \cdot \overline{S}}{1+r} \right] > 0.$$

Note that interchanging summation and integration is justified by dominated convergence, because

$$\left|\xi^0\right| + \|S\|_\infty \sum_{i=0}^{\infty} |\xi^i| \in \mathcal{L}^1(P^*).$$

The following example shows that the implication $\Rightarrow$ of Theorem 1.6, namely that absence of arbitrage opportunities implies the existence of a risk-neutral measure, may no longer be true in an infinite market model. ◇

Example 1.18. Let $\Omega = \{1, 2, \dots\}$, and choose any probability measure P which assigns strictly positive probability to all singletons $\{\omega\}$. We take $r = 0$ and define a price system $\pi^i = 1$, for $i = 0, 1, \dots$. Prices at time 1 are given by $S^0 \equiv 1$ and, for $i = 1, 2, \dots$, by

$$S^i(\omega) = \begin{cases} 0 & \text{if } \omega = i, \\ 2 & \text{if } \omega = i+1, \\ 1 & \text{otherwise.} \end{cases}$$

Let us show that this market model is arbitrage-free. To this end, suppose that $\overline{\xi} = (\xi^0, \xi)$ is a portfolio such that $\xi \in \ell^1$ and such that $\overline{\xi} \cdot \overline{S}(\omega) \geq 0$ for each $\omega \in \Omega$, but such that $\overline{\pi} \cdot \overline{\xi} \leq 0$. Considering the case $\omega = 1$ yields

$$0 \leq \overline{\xi} \cdot \overline{S}(1) = \xi^0 + \sum_{k=2}^{\infty} \xi^k = \overline{\pi} \cdot \overline{\xi} - \xi^1 \leq -\xi^1.$$

Similarly, for $\omega = i > 1$,

$$0 \le \overline{\xi} \cdot \overline{S}(\omega) = \xi^0 + 2\xi^{i-1} + \sum_{\substack{k=1\\k\neq i,i-1}}^{\infty} \xi^k = \overline{\pi} \cdot \overline{\xi} + \xi^{i-1} - \xi^i \le \xi^{i-1} - \xi^i.$$

It follows that $0 \ge \xi^1 \ge \xi^2 \ge \cdots$. But this can only be true if all ξ^i vanish, since we have assumed that $\xi \in \ell^1$. Hence, there are no arbitrage opportunities.

However, there exists no probability measure $P^* \approx P$ such that $E^*[\, S^i \,] = \pi^i$ for all i. Such a measure P^* would have to satisfy

$$\begin{aligned} 1 = E^*[\, S^i \,] &= 2P^*[\{i+1\}] + \sum_{\substack{k=1\\k\neq i,i+1}}^{\infty} P^*[\{k\}] \\ &= 1 + P^*[\{i+1\}] - P^*[\{i\}] \end{aligned}$$

for $i > 1$. This relation implies that $P^*[\,\{i\}\,] = P^*[\,\{i+1\}\,]$ for all $i > 1$, contradicting the assumption that P^* is a probability measure and equivalent to P. ◇

1.3 Derivative securities

In real financial markets, not only the primary assets are traded. There is also a large variety of securities whose payoff depends in a non-linear way on the primary assets $S^0, S^1, \dots, S^d$, and sometimes also on other factors. Such financial instruments are usually called *derivative securities*, *options*, or *contingent claims*.

Example 1.19. Under a *forward contract*, one agent agrees to sell to another agent an asset at time 1 for a price K which is specified at time 0. Thus, the owner of a forward contract on the i^{th} asset gains the difference between the actual market price S^i and the *delivery price* K if S^i is larger than K at time 1. If $S^i < K$, the owner loses the amount $K - S^i$ to the issuer of the forward contract. Hence, a forward contract corresponds to the random payoff

$$C^{\text{fw}} = S^i - K.$$

◇

Example 1.20. The owner of a *call option* on the i^{th} asset has the right, but not the obligation, to buy the i^{th} asset at time 1 for a fixed price K, called the *strike price*. This corresponds to a payoff of the form

$$C^{\text{call}} = (S^i - K)^+ = \begin{cases} S^i - K & \text{if } S^i > K, \\ 0 & \text{otherwise.} \end{cases}$$

Conversely, a *put option* gives the right, but not the obligation, to sell the asset at time 1 for a strike price K. The corresponding random payoff is given by

$$C^{\text{put}} = (K - S^i)^+ = \begin{cases} K - S^i & \text{if } S^i < K, \\ 0 & \text{otherwise.} \end{cases}$$

Call and put options with the same strike K are related through the formula

$$C^{\text{call}} - C^{\text{put}} = S^i - K.$$

Hence, if the price $\pi(C^{\text{call}})$ of a call option has already been fixed, then the price $\pi(C^{\text{put}})$ of the corresponding put option is determined by linearity through the *put-call parity*

$$\pi(C^{\text{call}}) = \pi(C^{\text{put}}) + \pi^i - \frac{K}{1+r}. \tag{1.10}$$

◇

Example 1.21. An option on the value $V = \overline{\xi} \cdot \overline{S}$ of a portfolio of several risky assets is sometimes called a *basket* or *index option*. For instance, a basket call would be of the form $(V - K)^+$. ◇

Put and call options can be used as building blocks for a large class of derivatives.

Example 1.22. A *straddle* is a combination of "at-the-money" put and call options on a portfolio $V = \overline{\xi} \cdot \overline{S}$, i.e., on put and call options with strike $K = \pi(V)$:

$$C = (\pi(V) - V)^+ + (V - \pi(V))^+ = |V - \pi(V)|.$$

Thus, the payoff of the straddle increases proportionally to the change of the price of $\overline{\xi}$ between time 0 and time 1. In this sense, a straddle is a bet that the portfolio price will move, no matter in which direction. ◇

Example 1.23. The payoff of a *butterfly spread* is of the form

$$C = \big(K - |V - \pi(V)|\big)^+,$$

where $K > 0$ and where $V = \overline{\xi} \cdot \overline{S}$ is the price of a given portfolio or the value of a stock index. Clearly, the payoff of the butterfly spread is maximal if $V = \pi(V)$ and decreases if the price at time 1 of the portfolio $\overline{\xi}$ deviates from its price at time 0. Thus, the butterfly spread is a bet that the portfolio price will stay close to its present value. By letting $K_\pm := \pi(V) \pm K$, we can represent C as combinations of call or put options on V:

$$\begin{aligned} C &= (V - K_-)^+ - 2(V - \pi(V))^+ + (V - K_+)^+ \\ &= -(K_- - V)^+ + 2(\pi(V) - V)^+ - (K_+ - V)^+. \end{aligned}$$

◇

Example 1.24. The idea of *portfolio insurance* is to increase exposure to rising asset prices, and to reduce exposure to falling prices. This suggests to replace the payoff $V = \overline{\xi} \cdot \overline{S}$ of a given portfolio by a modified profile $h(V)$, where h is convex and increasing. Let us first consider the case where $V \geq 0$. Then the corresponding payoff $h(V)$ can be expressed as a combination of investments in bonds, in V itself, and in basket call options on V. To see this, recall that convexity implies that $h(x) = h(0) + \int_0^x h'(y)\,dy$ for the increasing right-hand derivative $h' := h'_+$ of h; see Appendix A.1. Note that h' can be represented as the distribution function of a positive Radon measure γ on $[0, \infty)$: $h'(x) = \gamma([0, x])$ for $x \geq 0$. Hence, Fubini's theorem implies that

$$\begin{aligned} h(x) &= h(0) + \int_0^x \int_{[0,y]} \gamma(dz)\,dy \\ &= h(0) + \gamma(\{0\})\,x + \int_{(0,\infty)} \int_{\{y \mid z \leq y \leq x\}} dy\,\gamma(dz). \end{aligned}$$

Since the inner integral equals $(x - z)^+$, we obtain

$$h(V) = h(0) + h'(0)\,V + \int_{(0,\infty)} (V - z)^+\,\gamma(dz). \tag{1.11}$$

The payoff $V = \overline{\xi} \cdot \overline{S}$ may take negative values if the portfolio $\overline{\xi}$ contains also short positions. In this case, the increasing convex function h must be defined on all of $\mathbb{R}$. Its right-hand derivative h' can be represented as

$$h'(y) - h'(x) = \gamma\big((x, y]\big), \quad x < y,$$

for a positive Radon measure γ on $\mathbb{R}$. Looking separately at the cases $x < 0$ and $x \geq 0$, we see that

$$h(x) = h(0) + h'(0)\,x + \int_{(0,\infty)} (x - z)^+\,\gamma(dz) + \int_{(-\infty,0]} (z - x)^+\,\gamma(dz).$$

Thus, the payoff $h(V)$ can be realized by holding bonds, shares in V, and a mixture of call and put options on V:

$$h(V) = h(0) + h'(0)\,V + \int_{(0,\infty)} (V - z)^+\,\gamma(dz) + \int_{(-\infty,0]} (z - V)^+\,\gamma(dz).$$

◇

Example 1.25. A *reverse convertible bond* pays interest which is higher than that earned by an investment into the riskless bond. But at maturity $t = 1$, the issuer may convert the bond into a predetermined number of shares of a given asset S^i instead of paying the nominal value in cash. The purchase of this contract is equivalent to the purchase of a standard bond and the sale of a certain put option. More precisely, suppose that 1 is the price of the reverse convertible bond at $t = 0$, that its nominal value at maturity is $1 + \widetilde{r}$, and that it can be converted into x shares of the i^{th} asset. This conversion will happen if the asset price S^i is below $K := (1 + \widetilde{r})/x$. Thus, the payoff of the reverse convertible bond is equal to

$$1 + \widetilde{r} - x(K - S^i)^+,$$

i.e., the purchase of this contract is equivalent to a risk-free investment of 1 with interest r and the sale of the put option $x(K - S^i)^+$ for the price $(\widetilde{r} - r)/(1 + r)$. ◇

Example 1.26. A *discount certificate* on $V = \overline{\xi} \cdot \overline{S}$ pays off the amount

$$C = V \wedge K,$$

where the number $K > 0$ is often called the *cap*. Since

$$C = V - (V - K)^+,$$

buying the discount certificate is the same as purchasing $\overline{\xi}$ and selling the basket call option $C^{\text{call}} := (V - K)^+$. If the price $\pi(C^{\text{call}})$ has already been fixed, then the price of C is given by $\pi(C) = \pi(V) - \pi(C^{\text{call}})$. Hence, the discount certificate is less expensive than the portfolio $\overline{\xi}$ itself, and this explains the name. On the other hand, it participates in gains of $\overline{\xi}$ only up to the cap K. ◇

Example 1.27. For an insurance company, it may be desirable to shift some of its insurance risk to the financial market. As an example of such an *alternative risk transfer*, consider a *catastrophe bond* issued by an insurance company. The interest paid by this security depends on the occurrence of certain special events. For instance, the contract may specify that no interest will be paid if more than a given number of insured cars are damaged by hail on a single day during the lifetime of the contract; as a compensation for taking this risk, the buyer will be paid an interest above the usual market rate if this event does not occur. ◇

Mathematically, it will be convenient to focus on contingent claims whose payoff is non-negative. Such a contingent claim will be interpreted as a contract which is sold at time 0 and which pays a random amount $C(\omega) \geq 0$ at time 1. A derivative security whose terminal value may also become negative can usually be reduced to a combination of a non-negative contingent claim and a short position in some of the primary assets $S^0, S^1, \dots, S^d$. For instance, the terminal value of a reverse convertible bond is bounded from below so that it can be decomposed into a short position in cash and into a contract with positive value. From now on, we will work with the following formal definition of the term "contingent claim".

Definition 1.28. A *contingent claim* is a random variable C on the underlying probability space $(\Omega, \mathcal{F}, P)$ such that

$$0 \leq C < \infty \quad P\text{-a.s.}$$

A contingent claim C is called a *derivative* of the primary assets $S^0, \dots, S^d$ if it is measurable with respect to the σ-field $\sigma(S^0, \dots, S^d)$ generated by the assets, i.e., if

$$C = f(S^0, \dots, S^d)$$

for a measurable function f on $\mathbb{R}^{d+1}$.

So far, we have only fixed the prices π^i of our primary assets S^i. Thus, it is not clear what the correct price should be for a general contingent claim C. Our main goal in this section is to identify those possible prices which are compatible with the given prices in the sense that they do not generate arbitrage. Our approach is based on the observation that trading C at time 0 for a price π^C corresponds to introducing a new asset with the prices

$$\pi^{d+1} := \pi^C \quad \text{and} \quad S^{d+1} := C. \tag{1.12}$$

Definition 1.29. A real number $\pi^C \geq 0$ is called an *arbitrage-free price* of a contingent claim C if the market model extended according to (1.12) is arbitrage-free. The set of all arbitrage-free prices for C is denoted $\Pi(C)$.

In the previous definition, we made the implicit assumption that the introduction of a contingent claim C as a new asset does not affect the prices of primary assets. This assumption is reasonable as long as the trading volume of C is small compared to that of the primary assets. In Section 3.4 we will discuss the equilibrium approach to asset pricing, where an extension of the market will typically change the prices of *all* traded assets.

The following result shows in particular that we can always find an arbitrage-free price for a given contingent claim C if the initial model is arbitrage-free.

Theorem 1.30. *Suppose that the set $\mathcal{P}$ of equivalent risk-neutral measures for the original market model is non-empty. Then the set of arbitrage-free prices of a contingent claim C is non-empty and given by*

$$\Pi(C) = \left\{ E^*\left[\frac{C}{1+r} \right] \,\middle|\, P^* \in \mathcal{P} \text{ such that } E^*[\,C\,] < \infty \right\}. \tag{1.13}$$

Proof. By Theorem 1.6, π^C is an arbitrage-free price for C if and only if there exists an equivalent risk-neutral measure $\hat{P}$ for the market model extended via (1.12), i.e.,

$$\pi^i = \hat{E}\left[\frac{S^i}{1+r} \right] \quad \text{for } i = 1, \dots, d+1.$$

In particular, $\hat{P}$ is necessarily contained in $\mathcal{P}$, and we obtain the inclusion $\subseteq$ in (1.13). Conversely, if $\pi^C = E^*[\,C/(1+r)\,]$ for some $P^* \in \mathcal{P}$, then this P^* is also an equivalent risk-neutral measure for the extended market model, and so the two sets in (1.13) are equal.

To show that $\Pi(C)$ is non-empty, we first fix some measure $\widetilde{P} \approx P$ such that $\widetilde{E}[\,C\,] < \infty$. For instance, we can take $d\widetilde{P} = c(1+C)^{-1}\,dP$, where c is the normalizing constant. Under $\widetilde{P}$, the market model is arbitrage-free. Hence, Theorem 1.6 yields $P^* \in \mathcal{P}$ such that $dP^*/d\widetilde{P}$ is bounded. In particular, $E^*[\,C\,] < \infty$ and $\pi^C = E^*[\,C/(1+r)\,] \in \Pi(C)$. □

The following theorem provides a dual characterization of the lower and upper bounds

$$\pi_{\inf}(C) := \inf \Pi(C) \quad \text{and} \quad \pi_{\sup}(C) := \sup \Pi(C),$$

which are often called *arbitrage bounds* for C.

Theorem 1.31. *In an arbitrage-free market model, the arbitrage bounds of a contingent claim C are given by*

$$\begin{aligned}\pi_{\inf}(C) &= \inf_{P^* \in \mathcal{P}} E^*\Big[\,\frac{C}{1+r}\,\Big]\\ &= \max\Big\{\, m \in [0,\infty) \,\Big|\, \exists\, \xi \in \mathbb{R}^d \textit{ with } m + \xi \cdot Y \le \frac{C}{1+r}\ P\text{-a.s.}\Big\}\end{aligned}$$

and

$$\begin{aligned}\pi_{\sup}(C) &= \sup_{P^* \in \mathcal{P}} E^*\Big[\,\frac{C}{1+r}\,\Big]\\ &= \min\Big\{\, m \in [0,\infty] \,\Big|\, \exists\, \xi \in \mathbb{R}^d \textit{ with } m + \xi \cdot Y \ge \frac{C}{1+r}\ P\text{-a.s.}\Big\}.\end{aligned}$$

Proof. We only prove the identities for the upper arbitrage bound. The ones for the lower bound are obtained in a similar manner. We take $m \in [0,\infty]$ and $\xi \in \mathbb{R}^d$ such that $m + \xi \cdot Y \ge C/(1+r)$ P-a.s., and we denote by M the set of all such m. Taking the expectation with $P^* \in \mathcal{P}$ yields $m \ge E^*[\,C/(1+r)\,]$, and we get

$$\begin{aligned}\inf M &\ge \sup_{P^* \in \mathcal{P}} E^*\Big[\,\frac{C}{1+r}\,\Big]\\ &\ge \sup\Big\{\, E^*\Big[\,\frac{C}{1+r}\,\Big] \,\Big|\, P^* \in \mathcal{P},\, E^*[\,C\,] < \infty \Big\} = \pi_{\sup}(C),\end{aligned} \tag{1.14}$$

where we have used Theorem 1.30 in the last identity.

Next we show that all inequalities in (1.14) are in fact identities. This is trivial if $\pi_{\sup}(C) = \infty$. For $\pi_{\sup}(C) < \infty$, we will show that $m > \pi_{\sup}(C)$ implies $m \ge \inf M$. By definition, $\pi_{\sup}(C) < m < \infty$ requires the existence of an arbitrage opportunity in the market model extended by $\pi^{d+1} := m$ and $S^{d+1} := C$. That is, there is $(\xi, \xi^{d+1}) \in \mathbb{R}^{d+1}$ such that $\xi \cdot Y + \xi^{d+1}(C/(1+r) - m)$ is almost-surely nonnegative and strictly positive with positive probability. Since the original market model is arbitrage-free, ξ^{d+1} must be non-zero. In fact, we have $\xi^{d+1} < 0$ as taking expectations with respect to $P^* \in \mathcal{P}$ for which $E^*[\,C\,] < \infty$ yields

$$\xi^{d+1}\Big(E^*\Big[\,\frac{C}{1+r}\,\Big] - m\Big) \geq 0,$$

and the term in parenthesis is negative since $m > \pi_{\sup}(C)$. Thus, we may define $\zeta := -\xi/\xi^{d+1} \in \mathbb{R}^d$ and obtain $m + \zeta \cdot Y \geq C/(1+r)$ P-a.s., hence $m \geq \inf M$.

We now prove that the infimum of M is in fact attained. To this end, we may assume without loss of generality that $\inf M < \infty$ and that the market model is non-redundant in the sense of Definition 1.13. For a sequence $m_n \in M$ that decreases towards $\inf M = \pi_{\sup}(C)$, we fix $\xi_n \in \mathbb{R}^d$ such that $m_n + \xi_n \cdot Y \geq C/(1+r)$ P-almost surely. If $\liminf_n |\xi_n| < \infty$, there exists a subsequence of (ξ_n) that converges to some $\xi \in \mathbb{R}^d$. Passing to the limit yields $\pi_{\sup}(C) + \xi \cdot Y \geq C/(1+r)$ P-a.s., which gives $\pi_{\sup}(C) \in M$. But this is already the desired result, since the following argument will show that the case $\liminf_n |\xi_n| = \infty$ cannot occur. Indeed, after passing to some subsequence if necessary, $\eta_n := \xi_n/|\xi_n|$ converges to some $\eta \in \mathbb{R}^d$ with $|\eta| = 1$. Under the assumption that $|\xi_n| \to \infty$, passing to the limit in

$$\frac{\pi_{\sup}(C)}{|\xi_n|} + \eta_n \cdot Y \geq \frac{C}{|\xi_n|(1+r)} \quad P\text{-a.s.}$$

yields $\eta \cdot Y \geq 0$. The absence of arbitrage opportunities thus implies $\eta \cdot Y = 0$ P-a.s., whence $\eta = 0$ by non-redundance of the model. But this contradicts the fact that $|\eta| = 1$. □

Remark 1.32. Theorem 1.31 shows that $\pi_{\sup}(C)$ is the lowest possible price of a portfolio $\overline{\xi}$ with

$$\overline{\xi} \cdot \overline{S} \geq C \quad P\text{-a.s.}$$

Such a portfolio is often called a "superhedging strategy" or "superreplication" of C, and the identities for $\pi_{\inf}(C)$ and $\pi_{\sup}(C)$ obtained in Theorem 1.31 are often called *superhedging duality relations*. When using $\overline{\xi}$, the *seller* of C would be protected against any possible future claims of the buyer of C. Thus, a natural goal for the seller would be to finance such a superhedging strategy from the proceeds of C. Conversely, the objective of the *buyer* would be to cover the price of C from the sale of a portfolio $\overline{\eta}$ with

$$\overline{\eta} \cdot \overline{S} \leq C \quad P\text{-a.s.},$$

which is possible if and only if $\overline{\pi} \cdot \overline{\eta} \leq \pi_{\inf}(C)$. Unless C is an attainable payoff, however, neither objective can be fulfilled by trading C at an arbitrage-free price, as shown in Corollary 1.34 below. Thus, any arbitrage-free price involves a trade-off between these two objectives. ◇

For a portfolio $\overline{\xi}$ the resulting payoff $V = \overline{\xi} \cdot \overline{S}$, if positive, may be viewed as a contingent claim, and in particular as a derivative. Those claims which can be replicated by a suitable portfolio will play a special role in the sequel.

Definition 1.33. A contingent claim C is called *attainable* (*replicable, redundant*) if $C = \overline{\xi} \cdot \overline{S}$ P-a.s. for some $\overline{\xi} \in \mathbb{R}^{d+1}$. Such a portfolio strategy $\overline{\xi}$ is then called a *replicating portfolio* for C.

If one can show that a given contingent claim C can be replicated by some portfolio $\overline{\xi}$, then the problem of determining a price for C has a straightforward solution: The price of C is unique and equal to the cost $\overline{\xi} \cdot \overline{\pi}$ of its replication, due to the law of one price. The following corollary also shows that the attainable contingent claims are in fact the only ones for which admit a unique arbitrage-free price.

Corollary 1.34. *Suppose the market model is arbitrage-free and C is a contingent claim.*

(a) *C is attainable if and only if it admits a unique arbitrage-free price.*

(b) *If C is not attainable, then $\pi_{\text{inf}}(C) < \pi_{\text{sup}}(C)$ and*

$$\Pi(C) = \big(\pi_{\text{inf}}(C), \pi_{\text{sup}}(C)\big).$$

Proof. Clearly $|\Pi(C)| = 1$ if C is attainable, and so assertion (a) is implied by (b).

In order to prove part (b), note first that $\Pi(C)$ is non-empty and convex due to the convexity of $\mathcal{P}$. Hence $\Pi(C)$ is an interval. To show that this interval is open, it suffices to exclude the possibility that it contains one of its boundary points $\pi_{\text{inf}}(C)$ and $\pi_{\text{sup}}(C)$. To this end, we use Theorem 1.31 to get $\xi \in \mathbb{R}^d$ such that

$$\pi_{\text{inf}}(C) + \xi \cdot Y \le \frac{C}{1+r} \quad P\text{-a.s.}$$

Since C is not attainable, this inequality cannot be an almost-sure identity. Hence, with $\xi^0 := -(1+r)\pi_{\text{inf}}(C)$, the strategy $(\xi^0, -\xi, 1) \in \mathbb{R}^{d+2}$ is an arbitrage opportunity in the market model extended by $\pi^{d+1} := \pi_{\text{inf}}(C)$ and $S^{d+1} := C$, so that $\pi_{\text{inf}}(C)$ is not an arbitrage-free price for C. The possibility $\pi_{\text{sup}}(C) \in \Pi(C)$ is excluded by a similar argument. □

Remark 1.35. In Theorem 1.31, the set $\mathcal{P}$ of *equivalent* risk-neutral measures can be replaced by the set $\widetilde{\mathcal{P}}$ of risk-neutral measures that are merely *absolutely continuous* with respect to P. That is,

$$\pi_{\text{inf}}(C) = \inf_{\widetilde{P} \in \widetilde{\mathcal{P}}} \widetilde{E}\Big[\frac{C}{1+r}\Big] \quad \text{and} \quad \pi_{\text{sup}}(C) = \sup_{\widetilde{P} \in \widetilde{\mathcal{P}}} \widetilde{E}\Big[\frac{C}{1+r}\Big], \tag{1.15}$$

for any contingent claim C. To prove this, note first that $\mathcal{P} \subset \widetilde{\mathcal{P}}$, so that we get the two inequalities "$\geq$" and "$\leq$" in (1.15). On the other hand, for any $\widetilde{P} \in \widetilde{\mathcal{P}}$, arbitrary $P^* \in \mathcal{P}$ and $\varepsilon \in (0, 1]$, the measure $P^*_\varepsilon := \varepsilon P^* + (1-\varepsilon)\widetilde{P}$ belongs to $\mathcal{P}$ and satisfies $E^*_\varepsilon[\, C\,] = \varepsilon E^*[\, C\,] + (1-\varepsilon)\widetilde{E}[\, C\,]$. Sending $\varepsilon \downarrow 0$ yields the converse inequalities. ◇

Remark 1.36. Consider any arbitrage-free market model, and let $C^{\text{call}} = (S^i - K)^+$ be a call option on the i^{th} asset with strike $K > 0$. Clearly, $C^{\text{call}} \le S^i$ so that

$$E^*\Big[\,\frac{C^{\text{call}}}{1+r}\,\Big] \le \pi^i$$

for any $P^* \in \mathcal{P}$. From Jensen's inequality, we obtain the following lower bound:

$$E^*\Big[\,\frac{C^{\text{call}}}{1+r}\,\Big] \ge \Big(E^*\Big[\,\frac{S^i}{1+r}\,\Big] - \frac{K}{1+r}\Big)^+ = \Big(\pi^i - \frac{K}{1+r}\Big)^+.$$

Thus, the following universal bounds hold for any arbitrage-free market model:

$$\Big(\pi^i - \frac{K}{1+r}\Big)^+ \le \pi_{\inf}\big(C^{\text{call}}\big) \le \pi_{\sup}\big(C^{\text{call}}\big) \le \pi^i. \tag{1.16}$$

For a put option $C^{\text{put}} = (K - S^i)^+$, one obtains the universal bounds

$$\Big(\frac{K}{1+r} - \pi^i\Big)^+ \le \pi_{\inf}\big(C^{\text{put}}\big) \le \pi_{\sup}\big(C^{\text{put}}\big) \le \frac{K}{1+r}. \tag{1.17}$$

If $r \ge 0$, then the lower bound in (1.16) can be further reduced to $\pi_{\inf}\big(C^{\text{call}}\big) \ge (\pi^i - K)^+$. Informally, this inequality states that the value of the right to buy the i^{th} asset at $t = 0$ for a price K is strictly less than any arbitrage-free price for C^{call}. This fact is sometimes expressed by saying that the *time value* of a call option is non-negative. The quantity $(\pi^i - K)^+$ is called the *intrinsic value* of the call option. Observe that an analogue of this relation usually fails for put options: The left-hand side of (1.17) can only be bounded by its intrinsic value $(K - \pi^i)^+$ if $r \le 0$. If the intrinsic value of a put or call option is positive, then one says that the option is "in the money". For $\pi^i = K$ one speaks of an "at-the-money" option. Otherwise, the option is "out of the money". ◇

In many situations, the universal arbitrage bounds (1.16) and (1.17) are in fact attained, as illustrated by the following example.

Example 1.37. Take any market model with a single risky asset $S = S^1$ such that the distribution of S under P is concentrated on $\{0, 1, \dots, \}$ with positive weights. Without loss of generality, we may assume that S has under P a Poisson distribution with parameter 1, i.e., S is P-a.s. integer-valued and

$$P[\,S = k\,] = \frac{e^{-1}}{k!} \quad \text{for } k = 0, 1, \dots.$$

If we take $r = 0$ and $\pi = 1$, then P is a risk-neutral measure and the market model is arbitrage-free. We are going to show that the upper and lower bounds in (1.16) are attained for this model by using Remark 1.35. To this end, consider the measure $\widetilde{P} \in \widetilde{\mathcal{P}}$ which is defined by its density

$$\frac{d\widetilde{P}}{dP} = e \cdot \mathrm{I}_{\{S=1\}}\,.$$

We get

$$\widetilde{E}[\,(S-K)^+\,] = (1-K)^+ = (\pi - K)^+,$$

so that the lower bound in (1.16) is attained, i.e., we have

$$\pi_{\inf}\big((S-K)^+\big) = (\pi - K)^+.$$

To see that also the upper bound is sharp, we define

$$g_n(k) := \left(e - \frac{e}{n}\right) \cdot \mathrm{I}_{\{0\}}(k) + (n-1)! \cdot e \cdot \mathrm{I}_{\{n\}}(k), \quad k = 0, 1, \dots$$

It is straightforward to check that

$$d\widetilde{P}_n := g_n(S)\, dP$$

defines a measure $\widetilde{P}_n \in \widetilde{\mathcal{P}}$ such that

$$\widetilde{E}_n[\,(S-K)^+\,] = \left(1 - \frac{K}{n}\right)^+.$$

By sending $n \uparrow \infty$, we see that also the upper bound in (1.16) is attained:

$$\pi_{\sup}\big((S-K)^+\big) = \pi.$$

Furthermore, the put-call parity (1.10) shows that the universal bounds (1.17) for put options are attained as well. ◇

1.4 Complete market models

Our goal in this section is to characterize the particularly transparent situation in which all contingent claims are attainable.

Definition 1.38. An arbitrage-free market model is called *complete* if every contingent claim is attainable.

In every market model, the following inclusion holds for each $P^* \in \mathcal{P}$:

$$\begin{aligned}\mathcal{V} = \left\{ \overline{\xi} \cdot \overline{S} \;\middle|\; \overline{\xi} \in \mathbb{R}^{d+1} \right\} &\subseteq L^1\big(\Omega, \sigma(S^1, \dots, S^d), P^*\big) \\ &\subseteq L^0(\Omega, \mathcal{F}, P^*) = L^0(\Omega, \mathcal{F}, P);\end{aligned}$$

see Appendix A.7 for the definition of L^p-spaces. If the market is complete then all of these inclusions are in fact equalities. In particular, $\mathcal{F}$ coincides with $\sigma(S^1, \dots, S^d)$ modulo P-null sets, and every contingent claim coincides P-a.s. with a derivative of the traded assets. Since the linear space $\mathcal{V}$ is finite-dimensional, it follows that the same must be true of $L^0(\Omega, \mathcal{F}, P)$. But this means that the model can be reduced to a finite number of relevant scenarios. This observation can be made precise by using the notion of an *atom* of the probability space $(\Omega, \mathcal{F}, P)$. Recall that a set $A \in \mathcal{F}$ is called an atom of $(\Omega, \mathcal{F}, P)$, if $P[\,A\,] > 0$ and if each $B \in \mathcal{F}$ with $B \subseteq A$ satisfies either $P[\,B\,] = 0$ or $P[\,B\,] = P[\,A\,]$.

Proposition 1.39. *For any $p \in [0, \infty]$, the dimension of the linear space $L^p(\Omega, \mathcal{F}, P)$ is given by*

$$\begin{aligned} &\dim L^p(\Omega, \mathcal{F}, P) \\ &\quad = \sup\left\{ n \in \mathbb{N} \mid \exists \text{ partition } A^1, \dots, A^n \text{ of } \Omega \text{ with } A^i \in \mathcal{F} \text{ and } P[\,A^i\,] > 0 \right\}. \end{aligned} \tag{1.18}$$

Moreover, $n := \dim L^p(\Omega, \mathcal{F}, P) < \infty$ if and only if there exists a partition of Ω into n atoms of $(\Omega, \mathcal{F}, P)$.

Proof. Suppose that there is a partition $A^1, \dots, A^n$ of Ω such that $A^i \in \mathcal{F}$ and $P[\,A^i\,] > 0$. The corresponding indicator functions $I_{A^1}, \dots, I_{A^n}$ can be regarded as linearly independent vectors in $L^p := L^p(\Omega, \mathcal{F}, P)$. Thus $\dim L^p \geq n$. Consequently, it suffices to consider only the case in which the right-hand side of (1.18) is a finite number, n_0. If $A^1, \dots, A^{n_0}$ is a corresponding partition, then each A^i is an atom because otherwise n_0 would not be maximal. Thus, any $Z \in L^p$ is P-a.s. constant on each A^i. If we denote the value of Z on A^i by z^i, then

$$Z = \sum_{i=1}^{n_0} z_i \, I_{A^i} \qquad P\text{-a.s.}$$

Hence, the indicator functions $I_{A^1}, \dots, I_{A^{n_0}}$ form a basis of L^p, and this implies $\dim L^p = n_0$. □

Theorem 1.40. *An arbitrage-free market model is complete if and only if there exists exactly one risk-neutral probability measure, i.e., if $|\mathcal{P}| = 1$. In this case, $\dim L^0(\Omega, \mathcal{F}, P) \leq d + 1$.*

Proof. If the model is complete, then the indicator I_A of each set $A \in \mathcal{F}$ is an attainable contingent claim. Hence, Corollary 1.34 implies that $P^*[\,A\,] = E^*[\,I_A\,]$ is independent of $P^* \in \mathcal{P}$. Consequently, there is just one risk-neutral probability measure.

Conversely, suppose that $\mathcal{P} = \{P^*\}$, and let C be a bounded contingent claim, so that $E^*[\,C\,] < \infty$. Then C has the unique arbitrage-free price $E^*[\,C/(1 + r)\,]$, and

Corollary 1.34 implies that C is attainable. It follows that $L^\infty(\Omega, \mathcal{F}, P)$ is contained in the linear space $\mathcal{V}$ of all possible portfolio values. This implies that

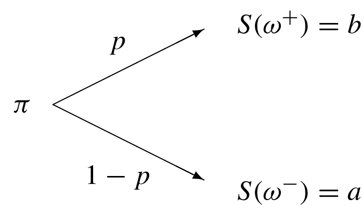
$$\dim L^\infty(\Omega, \mathcal{F}, P) \le \dim \mathcal{V} \le d + 1.$$

Hence, we conclude from Proposition 1.39 that $(\Omega, \mathcal{F}, P)$ has at most $d + 1$ atoms. But then every contingent claim must be bounded and, in turn, attainable. □

Example 1.41. Consider the simple situation where the sample space Ω consists of two elements ω^+ and ω^-, and where the measure P is such that

$$p := P[\,\{\omega^+\}\,] \in (0, 1).$$

We assume that there is one single risky asset, which takes at time $t = 1$ the two values b and a with the respective probabilities p and $1 - p$, where a and b are such that $0 \le a < b$:



This model does not admit arbitrage if and only if

$$\pi(1 + r) \in \big\{\widetilde{E}[\,S\,] \mid \widetilde{P} \approx P\big\} = \big\{\widetilde{p}\, b + (1 - \widetilde{p})a \mid \widetilde{p} \in (0, 1)\big\} = (a, b); \quad (1.19)$$

see also Example 1.8. In this case, the model is also complete: Any risk-neutral measure P^* must satisfy

$$\pi(1 + r) = E^*[\,S\,] = p^* b + (1 - p^*)a,$$

and this condition uniquely determines the parameter $p^* = P^*[\,\{\omega^+\}\,]$ as

$$p^* = \frac{\pi(1 + r) - a}{b - a} \in (0, 1).$$

Hence $|\mathcal{P}| = 1$, and completeness follows from Theorem 1.40. Alternatively, we can directly verify completeness by showing that a given contingent claim C is attainable if (1.19) holds. Observe that the condition

$$C(\omega) = \xi^0\, S^0(\omega) + \xi\, S(\omega) = \xi^0(1 + r) + \xi\, S(\omega) \quad \text{for all } \omega \in \Omega$$

is a system of two linear equations for the two real variables ξ^0 and ξ. The solution is given by

$$\xi = \frac{C(\omega^+) - C(\omega^-)}{b - a} \quad \text{and} \quad \xi^0 = \frac{C(\omega^-)b - C(\omega^+)a}{(b - a)(1 + r)}.$$

Therefore, the unique arbitrage-free price of C is

$$\pi(C) = \overline{\pi} \cdot \overline{\xi} = \frac{C(\omega^+)}{1+r} \cdot \frac{\pi(1+r)-a}{b-a} + \frac{C(\omega^-)}{1+r} \cdot \frac{b-\pi(1+r)}{b-a}.$$

For a call option $C = (S-K)^+$ with strike $K \in [a,b]$, we have

$$\pi\big((S-K)^+\big) = \frac{b-K}{b-a} \cdot \pi - \frac{(b-K)a}{b-a} \cdot \frac{1}{1+r}. \tag{1.20}$$

Note that this price is *independent* of p and *increasing* in r, while the classical discounted expectation with respect to the "objective" measure P,

$$E\left[\frac{C}{1+r}\right] = \frac{p(b-K)}{1+r},$$

is *decreasing* in r and *increasing* in p.

In this example, one can illustrate how options can be used to modify the risk of a position. Consider the particular case in which the risky asset can be bought at time $t=0$ for the price $\pi = 100$. At time $t=1$, the price is either $S(\omega^+) = b = 120$ or $S(\omega^-) = a = 90$, both with positive probability. If we invest in the risky asset, the corresponding returns are given by

$$R(S)(\omega^+) = +20\% \quad \text{or} \quad R(S)(\omega^-) = -10\%.$$

Now consider a call option $C := (S-K)^+$ with strike $K = 100$. Choosing $r = 0$, the price of the call option is

$$\pi(C) = \frac{20}{3} \approx 6.67$$

from formula (1.20). Hence the return

$$R(C) = \frac{(S-K)^+ - \pi(C)}{\pi(C)}$$

on the initial investment $\pi(C)$ equals

$$R(C)(\omega^+) = \frac{20-\pi(C)}{\pi(C)} = +200\%$$

or

$$R(C)(\omega^-) = \frac{0-\pi(C)}{\pi(C)} = -100\%,$$

according to the outcome of the market at time $t=1$. Here we see a dramatic increase of both profit opportunity and risk; this is sometimes referred to as the *leverage effect* of options.

On the other hand, we could reduce the risk of holding the asset by holding a combination

$$\widetilde{C} := (K - S)^+ + S$$

of a put option and the asset itself. This "portfolio insurance" will of course involve an additional cost. If we choose our parameters as above, then the put-call parity (1.10) yields that the price of the put option $(K - S)^+$ is equal to $20/3$. Thus, in order to hold both S and a put, we must invest the capital $100 + 20/3$ at time $t = 0$. At time $t = 1$, we have an outcome of either 120 or of 100 so that the return of $\widetilde{C}$ is given by

$$R(\widetilde{C})(\omega^+) = +12.5\% \quad \text{and} \quad R(\widetilde{C})(\omega^-) = -6.25\%. \qquad \diamond$$

1.5 Geometric characterization of arbitrage-free models

The "fundamental theorem of asset pricing" in the form of Theorem 1.6 states that a market model is arbitrage-free if and only if the origin is contained in the set

$$\mathcal{M}_b(Y, P) := \left\{ E_Q[\, Y\,] \;\middle|\; Q \approx P,\ \frac{dQ}{dP} \text{ is bounded, } E_Q[\,|Y|\,] < \infty \right\} \subset \mathbb{R}^d,$$

where $Y = (Y^1, \dots, Y^d)$ is the random vector of discounted net gains defined in (1.2). The aim of this section is to give a geometric description of the set $\mathcal{M}_b(Y, P)$ as well as of the larger set

$$\mathcal{M}(Y, P) := \left\{ E_Q[\, Y\,] \mid Q \approx P,\ E_Q[\,|Y|\,] < \infty \right\}.$$

To this end, it will be convenient to work with the distribution

$$\mu := P \circ Y^{-1}$$

of Y with respect to P. That is, μ is a Borel probability measure on $\mathbb{R}^d$ such that

$$\mu(A) = P[\, Y \in A\,] \quad \text{for each Borel set } A \subset \mathbb{R}^d.$$

If ν is a Borel probability measure on $\mathbb{R}^d$ such that $\int |y|\,\nu(dy) < \infty$, we will call $\int y\,\nu(dy)$ its *barycenter*.

Lemma 1.42. *We have*

$$\mathcal{M}_b(Y, P) = \mathcal{M}_b(\mu) := \left\{ \int y\,\nu(dy) \;\middle|\; \nu \approx \mu,\ \frac{d\nu}{d\mu} \text{ is bounded, } \int |y|\,\nu(dy) < \infty \right\},$$

and

$$\mathcal{M}(Y, P) = \mathcal{M}(\mu) := \left\{ \int y\,\nu(dy) \;\middle|\; \nu \approx \mu,\ \int |y|\,\nu(dy) < \infty \right\}.$$

Proof. If $\nu \approx \mu$ is a Borel probability measure on $\mathbb{R}^d$, then the Radon–Nikodym derivative of ν with respect to μ evaluated at the random variable Y defines a probability measure $Q \approx P$ on $(\Omega, \mathcal{F})$:

$$\frac{dQ}{dP}(\omega) := \frac{d\nu}{d\mu}\big(Y(\omega)\big).$$

Clearly, $E_Q[\,Y\,] = \int y\,\nu(dy)$. This shows that $\mathcal{M}(\mu) \subseteq \mathcal{M}(Y, P)$ and $\mathcal{M}_b(\mu) \subseteq \mathcal{M}_b(Y, P)$.

Conversely, if $\widetilde{Q}$ is a given probability measure on $(\Omega, \mathcal{F})$ which is equivalent to P, then the Radon–Nikodym theorem in Appendix A.2 shows that the distribution $\widetilde{\nu} := \widetilde{Q} \circ Y^{-1}$ must be equivalent to μ, whence $\mathcal{M}(Y, P) \subseteq \mathcal{M}(\mu)$. Moreover, it follows from Proposition A.11 that the density $d\widetilde{\nu}/d\mu$ is bounded if $d\widetilde{Q}/dP$ is bounded, and so $\mathcal{M}_b(Y, P) \subseteq \mathcal{M}_b(\mu)$ also follows. □

By the above lemma, the characterization of the two sets $\mathcal{M}_b(Y, P)$ and $\mathcal{M}(Y, P)$ is reduced to a problem for Borel probability measures on $\mathbb{R}^d$. Here and in the sequel, we do not need the fact that μ is the distribution of the lower bounded random vector Y of discounted net gains; our results are true for arbitrary μ such that $\int |y|\,\mu(dy) < \infty$; see also Remark 1.7.

Definition 1.43. The *support* of a Borel probability measure ν on $\mathbb{R}^d$ is the smallest closed set $A \subset \mathbb{R}^d$ such that $\nu(A^c) = 0$, and it will be denoted by $\operatorname{supp}\nu$.

The support of a measure ν can be obtained as the intersection of all closed sets A with $\nu(A^c) = 0$, i.e.,

$$\operatorname{supp}\nu = \bigcap_{\substack{A \text{ closed} \\ \nu(A^c)=0}} A.$$

We denote by

$$\begin{aligned}\Gamma(\mu) &:= \operatorname{conv}(\operatorname{supp}\mu) \\ &= \left\{ \sum_{k=1}^{n} \alpha_k y_k \;\middle|\; \alpha_k \geq 0,\ \sum_{k=1}^{n} \alpha_k = 1,\ y_k \in \operatorname{supp}\mu,\ n \in \mathbb{N} \right\}\end{aligned}$$

the *convex hull of the support of* μ. Thus, $\Gamma(\mu)$ is the smallest convex set which contains $\operatorname{supp}\mu$; see also Appendix A.1.

Example 1.44. Take $d = 1$, and consider the measure

$$\mu = \frac{1}{2}\,(\delta_{-1} + \delta_{+1}).$$

Clearly, the support of μ is equal to $\{-1, +1\}$ and so $\Gamma(\mu) = [-1, +1]$. A measure ν is equivalent to μ if and only if

$$\nu = \alpha\delta_{-1} + (1-\alpha)\delta_{+1}$$

for some $\alpha \in (-1, +1)$. Hence, $\mathcal{M}_b(\mu) = \mathcal{M}(\mu) = (-1, +1)$. ◇

The previous example gives the correct intuition, namely that one always has the inclusions

$$M_b(\mu) \subset M(\mu) \subset \Gamma(\mu).$$

But while the first inclusion will turn out to be an identity, the second inclusion is usually strict. Characterizing $M(\mu)$ in terms of $\Gamma(\mu)$ will involve the following concept:

Definition 1.45. The *relative interior* of a convex set $C \subset \mathbb{R}^d$ is the set of all points $x \in C$ such that for all $y \in C$ there exists some $\varepsilon > 0$ with

$$x - \varepsilon(y - x) \in C.$$

The relative interior of C is denoted $\operatorname{ri} C$.

If the convex set C has non-empty topological interior $\operatorname{int} C$, then $\operatorname{ri} C = \operatorname{int} C$, and the elementary properties of the relative interior collected in the following remarks become obvious. This applies in particular to the set $\Gamma(\mu)$ if the non-redundance condition (1.8) is satisfied. For the general case, proofs of these statements can be found, for instance, in §6 of [166].

Remark 1.46. Let C be a non-empty convex subset of $\mathbb{R}^d$, and consider the *affine hull* $\operatorname{aff} C$ spanned by C, i.e., the smallest affine set which contains C. If we identify $\operatorname{aff} C$ with some $\mathbb{R}^n$, then the relative interior of C is equal to the topological interior of C, considered as a subset of $\operatorname{aff} C \cong \mathbb{R}^n$. In particular, each non-empty convex set has non-empty relative interior. ◇

Remark 1.47. Let C be a non-empty convex subset of $\mathbb{R}^d$ and denote by $\overline{C}$ its closure. Then, if $x \in \operatorname{ri} C$,

$$\alpha x + (1 - \alpha) y \in \operatorname{ri} C \quad \text{for all } y \in \overline{C} \text{ and all } \alpha \in (0, 1]. \tag{1.21}$$

In particular, $\operatorname{ri} C$ is convex. Moreover, the operations of taking the closure or the relative interior of a convex set C are consistent with each other:

$$\operatorname{ri} \overline{C} = \operatorname{ri} C \quad \text{and} \quad \overline{\operatorname{ri} C} = \overline{C}. \tag{1.22}$$

◇

After these preparations, we can now state the announced geometric characterization of the set $M_b(\mu)$. Note that the proof of this characterization relies on the "fundamental theorem of asset pricing" in the form of Theorem 1.6.

Theorem 1.48. *The set of all barycenters of probability measures $\nu \approx \mu$ coincides with the relative interior of the convex hull of the support of μ. More precisely,*

$$M_b(\mu) = M(\mu) = \operatorname{ri} \Gamma(\mu).$$

Proof. In a first step, we show the inclusion $\operatorname{ri}\Gamma(\mu) \subseteq M_b(\mu)$. Suppose we are given $m \in \operatorname{ri}\Gamma(\mu)$. Let $\widetilde{\mu}$ denote the translated measure

$$\widetilde{\mu}(A) := \mu(A + m) \quad \text{for Borel sets } A \subset \mathbb{R}^d$$

where $A + m := \{x + m \mid x \in A\}$. Then $M_b(\widetilde{\mu}) = M_b(\mu) - m$, and analogous identities hold for $M(\widetilde{\mu})$ and $\Gamma(\widetilde{\mu})$. It follows that there is no loss of generality in assuming that $m = 0$, i.e., we must show that $0 \in M_b(\mu)$ if $0 \in \operatorname{ri}\Gamma(\mu)$.

We claim that $0 \in \operatorname{ri}\Gamma(\mu)$ implies the following "no-arbitrage" condition:

$$\text{If } \xi \in \mathbb{R}^d \text{ is such that } \xi \cdot y \geq 0 \text{ for } \mu\text{-a.e. } y, \text{ then } \xi \cdot y = 0 \text{ for } \mu\text{-a.e. } y. \tag{1.23}$$

If (1.23) is false, then we can find some $\xi \in \mathbb{R}^d$ such that $\xi \cdot y \geq 0$ for μ-a.e. y but $\mu(\{y \mid \xi \cdot y > \delta\}) > 0$ for some $\delta > 0$. In this case, the support of μ is contained in the closed set $\{ y \mid \xi \cdot y \geq 0 \}$ but not in the hyperplane $\{ y \mid \xi \cdot y = 0 \}$. We conclude that $\xi \cdot y \geq 0$ for all $y \in \operatorname{supp}\mu$ and that there exists at least one $y^* \in \operatorname{supp}\mu$ such that $\xi \cdot y^* > 0$. In particular, $y^* \in \Gamma(\mu)$ so that our assumption $m = 0 \in \operatorname{ri}\Gamma(\mu)$ implies the existence of some $\varepsilon > 0$ such that $-\varepsilon y^* \in \Gamma(\mu)$. Consequently, $-\varepsilon y^*$ can be represented as a convex combination

$$-\varepsilon y^* = \alpha_1 y_1 + \cdots + \alpha_n y_n$$

of certain $y_1, \ldots, y_n \in \operatorname{supp}\mu$. It follows that

$$0 > -\varepsilon \xi \cdot y^* = \alpha_1 \xi \cdot y_1 + \cdots + \alpha_n \xi \cdot y_n,$$

in contradiction to our assumption that $\xi \cdot y \geq 0$ for all $y \in \operatorname{supp}\mu$. Hence, (1.23) must be true.

Applying the "fundamental theorem of asset pricing" in the form of Theorem 1.6 to $\Omega := \mathbb{R}^d$, $P := \mu$, and to the random variable $Y(y) := y$, yields a probability measure $\mu^* \approx \mu$ whose density $d\mu^*/d\mu$ is bounded and which satisfies $\int |y|\, \mu^*(dy) < \infty$ and $\int y\, \mu^*(dy) = 0$. This proves the inclusion $\operatorname{ri}\Gamma(\mu) \subseteq M_b(\mu)$.

Clearly, $M_b(\mu) \subset M(\mu)$. So the theorem will be proved if we can show the inclusion $M(\mu) \subset \operatorname{ri}\Gamma(\mu)$. To this end, suppose by way of contradiction that $\nu \approx \mu$ is such that

$$\int |y|\, \nu(dy) < \infty \quad \text{and} \quad m := \int y\, \nu(dy) \notin \operatorname{ri}\Gamma(\mu).$$

Again, we may assume without loss of generality that $m = 0$. Applying the separating hyperplane theorem in the form of Proposition A.1 with $\mathcal{C} := \operatorname{ri}\Gamma(\mu)$ yields some $\xi \in \mathbb{R}^d$ such that $\xi \cdot y \geq 0$ for all $y \in \operatorname{ri}\Gamma(\mu)$ and $\xi \cdot y^* > 0$ for at least one $y^* \in \operatorname{ri}\Gamma(\mu)$. We deduce from (1.21) that $\xi \cdot y \geq 0$ holds also for all $y \in \Gamma(\mu)$. Moreover, $\xi \cdot y_0$ must be strictly positive for at least one $y_0 \in \operatorname{supp}\mu$. Hence,

$$\xi \cdot y \geq 0 \text{ for } \mu\text{-a.e. } y \in \mathbb{R}^d \quad \text{and} \quad \mu\big(\{y \mid \xi \cdot y > 0\}\big) > 0. \tag{1.24}$$

By the equivalence of μ and ν, (1.24) is also true for ν instead of μ, and so

$$\xi \cdot m = \xi \cdot \int y \, \nu(dy) = \int \xi \cdot y \, \nu(dy) > 0,$$

in contradiction to our assumption that $m = 0$. We conclude that $M(\mu) \subset \operatorname{ri} \Gamma(\mu)$. $\square$

Remark 1.49. Note that Theorem 1.48 does *not* extend to the set

$$\widetilde{M}(\mu) := \left\{ \int y \, \nu(dy) \;\middle|\; \nu \ll \mu \text{ and } \int |y| \, \nu(dy) < \infty \right\}.$$

Already the simple case $\mu := \frac{1}{2}(\delta_{-1} + \delta_{+1})$ serves as a counterexample, because here $\widetilde{M}(\mu) = [-1, +1]$ while $\operatorname{ri} \Gamma(\mu) = (-1, +1)$. In this case, we have an identity between $\widetilde{M}(\mu)$ and $\Gamma(\mu)$. However, also this identity fails in general as can be seen by considering the normalized Lebesgue measure λ on $[-1, +1]$. For this choice one finds $\widetilde{M}(\lambda) = (-1, +1)$ but $\Gamma(\lambda) = [-1, +1]$. $\diamond$

From Theorem 1.48 we obtain the following geometric characterization of the absence of arbitrage.

Corollary 1.50. *Let μ be the distribution of the discounted price vector $S/(1+r)$ of the risky assets. Then the market model is arbitrage-free if and only if the price system π belongs to the relative interior* $\operatorname{ri} \Gamma(\mu)$ *of the convex hull of the support of μ.*

1.6 Contingent initial data

The idea of hedging contingent claims develops its full power only in a dynamic setting in which trading may occur at several times. The corresponding discrete-time theory constitutes the core of these notes and is presented in Chapter 5. The introduction of additional trading periods requires more sophisticated techniques than those we have used so far. In this section we will introduce some of these techniques in an extended version of our previous market model in which initial prices, and hence strategies, are contingent on scenarios. In this context, we are going to characterize the absence of arbitrage strategies. The results will be used as building blocks in the multiperiod setting of Part II; their study can be postponed until Chapter 5.

Suppose that we are given a σ-algebra $\mathcal{F}_0 \subset \mathcal{F}$ which specifies the information that is available to an investor at time $t = 0$. The prices for our $d + 1$ assets at time 0 will be modelled as non-negative $\mathcal{F}_0$-measurable random variables $S_0^0, S_0^1, \dots, S_0^d$. Thus, the price system $\overline{\pi} = (\pi^0, \pi^1, \dots, \pi^d)$ of our previous discussion is replaced by the vector

$$\overline{S}_0 = (S_0^0, \dots, S_0^d).$$

The portfolio $\overline{\xi}$ chosen by an investor at time $t = 0$ will also depend on the information available at time 0. Thus, we assume that

$$\overline{\xi} = (\xi^0, \xi^1, \dots, \xi^d)$$

is an $\mathcal{F}_0$-measurable random vector. The asset prices observed at time $t = 1$ will be denoted by

$$\overline{S}_1 = (S_1^0, S_1^1, \dots, S_1^d).$$

They are modelled as non-negative random variables which are measurable with respect to a σ-algebra $\mathcal{F}_1$ such that $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}$. The σ-algebra $\mathcal{F}_1$ describes the information available at time 1, and in this section we can assume that $\mathcal{F} = \mathcal{F}_1$.

A riskless bond could be included by taking $S_0^0 \equiv 1$ and by assuming S_1^0 to be $\mathcal{F}_0$-measurable and P-a.s. strictly positive. However, in the sequel it will be sufficient to assume that S_0^0 is $\mathcal{F}_0$-measurable, S_1^0 is $\mathcal{F}_1$-measurable, and that

$$P\big[\, S_0^0 > 0 \text{ and } S_1^0 > 0 \,\big] = 1. \tag{1.25}$$

Thus, we can take the 0th asset as *numéraire*, and we denote by

$$X_t^i := \frac{S_t^i}{S_t^0}, \quad i = 1, \dots, d, \; t = 0, 1,$$

the discounted asset prices and by

$$Y = X_1 - X_0$$

the vector of the discounted net gains.

Definition 1.51. An *arbitrage opportunity* is a portfolio $\overline{\xi}$ such that $\overline{\xi} \cdot \overline{S}_0 \le 0$, $\overline{\xi} \cdot \overline{S}_1 \ge 0$ P-a.s., and $P[\, \overline{\xi} \cdot \overline{S}_1 > 0\,] > 0$.

By our assumption (1.25), any arbitrage opportunity $\overline{\xi} = (\xi^0, \xi)$ satisfies

$$\xi \cdot Y \ge 0 \; P\text{-a.s.} \quad \text{and} \quad P[\, \xi \cdot Y > 0\,] > 0. \tag{1.26}$$

In fact, the existence of a d-dimensional $\mathcal{F}_0$-measurable random vector ξ with (1.26) is equivalent to the existence of an arbitrage opportunity. This can be seen as in Lemma 1.3.

The space of discounted net gains which can be generated by some portfolio is given by

$$\mathcal{K} := \big\{\, \xi \cdot Y \mid \xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d) \,\big\}.$$

Here, $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ denotes the space of $\mathbb{R}^d$-valued random variables which are P-a.s. finite and $\mathcal{F}_0$-measurable modulo the equivalence relation (A.21) of coincidence up to P-null sets. The spaces $L^p(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ for $p > 0$ are defined in the same manner. We denote by $L^p_+ := L^p_+(\Omega, \mathcal{F}_1, P)$ the cone of all non-negative elements in the space $L^p := L^p(\Omega, \mathcal{F}_1, P)$. With this notation, writing

$$\mathcal{K} \cap L^0_+ = \{0\}$$

is shorthand for the absence of arbitrage opportunities. We will denote by

$$\mathcal{K} - L^0_+$$

the convex cone of all $Z \in L^0$ which can be written as the difference of some $\xi \cdot Y \in \mathcal{K}$ and some $U \in L^0_+$.

The following definition involves the notion of the *conditional expectation*

$$E_Q[\, Z \mid \mathcal{F}_0 \,]$$

of a random variable Z with respect to a probability measure Q, given the σ-algebra $\mathcal{F}_0 \subset \mathcal{F}$; see Appendix A.2 and the references therein. If $Z = (Z^1, \dots, Z^n)$ is a random vector, then $E_Q[\, Z \mid \mathcal{F}_0 \,]$ is shorthand for the random vector with components $E_Q[\, Z^i \mid \mathcal{F}_0 \,]$, $i = 1, \dots, n$.

Definition 1.52. A probability measure Q satisfying

$$E_Q[\, X^i_t \,] < \infty \qquad \text{for } i = 1, \dots, d \text{ and } t = 0, 1$$

and

$$X_0 = E_Q[\, X_1 \mid \mathcal{F}_0 \,] \qquad Q\text{-a.s.}$$

is called a *risk-neutral measure* or *martingale measure*. We denote by $\mathcal{P}$ the set of all risk-neutral measures P^* which are equivalent to P.

Remark 1.53. The definition of a martingale measure Q means that for each asset $i = 0, \dots, d$, the discounted price process $(X^i_t)_{t=0,1}$ is a *martingale* under Q with respect to the σ-fields $(\mathcal{F}_t)_{t=0,1}$. The systematic discussion of martingales in a multi-period setting will begin in Section 5.2. The martingale aspect will be crucial for the theory of dynamic hedging in Part II. ◇

As the main result of this section, we can now state an extension of the "fundamental theorem of asset pricing" in Theorem 1.6 to our present setting.

Theorem 1.54. *The following conditions are equivalent:*

(a) $\mathcal{K} \cap L^0_+ = \{0\}$.

(b) $(\mathcal{K} - L^0_+) \cap L^0_+ = \{0\}$.

(c) *There exists a measure $P^* \in \mathcal{P}$ with a bounded density dP^*/dP.*

(d) $\mathcal{P} \neq \emptyset$.

Proof. (d) ⇒(a): Suppose by way of contradiction that there exist both a $P^* \in \mathcal{P}$ and some $\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ with non-zero payoff $\xi \cdot Y \in \mathcal{K} \cap L^0_+$. For large enough $c > 0$, $\xi^{(c)} := I_{\{|\xi| \le c\}}\xi$ will be bounded, and the payoff $\xi^{(c)} \cdot Y$ will still be non-zero and in $\mathcal{K} \cap L^0_+$. However,

$$E^*[\,\xi^{(c)} \cdot Y\,] = E^*\big[\,\xi^{(c)} \cdot E^*[\,Y \mid \mathcal{F}_0\,]\,\big] = 0,$$

which is the desired contradiction.

(a)⇔(b): It is obvious that (a) is necessary for (b). In order to prove sufficiency, suppose that we are given some $Z \in (\mathcal{K} - L^0_+) \cap L^0_+$. Then there exists a random variable $U \ge 0$ and a random vector $\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ such that

$$0 \le Z = \xi \cdot Y - U.$$

This implies that $\xi \cdot Y \ge U \ge 0$, which, according to condition (a), can only happen if $\xi \cdot Y = 0$. Hence, also $U = 0$ and in turn $Z = 0$.

(b) ⇒(c): This is the difficult part of the proof. The assertion will follow by combining Lemmas 1.56, 1.57, 1.59, and 1.67. □

Remark 1.55. If Ω is discrete, or if there exists a decomposition of Ω in countable many atoms of $(\Omega, \mathcal{F}_0, P)$, then the martingale measure P^* can be constructed by applying the result of Theorem 1.6 separately on each atom. In the general case, the idea of patching together conditional martingale measures would involve subtle arguments of measurable selection; see [51]. Here we present a different approach which is based on separation arguments in $L^1(P)$. It is essentially due to W. Schachermayer [175]; our version uses in addition arguments by Y. Kabanov and C. Stricker [122]. ◇

We start with the following simple lemma, which takes care of the integrability condition in Definition 1.52.

Lemma 1.56. *For the proof of the implication* (b) ⇒ (c) *in Theorem* 1.54*, we may assume without loss of generality that*

$$E[\,|X_t|\,] < \infty \quad \textit{for } t = 0, 1. \tag{1.27}$$

Proof. Define a probability measure $\widetilde{P}$ by

$$\frac{d\widetilde{P}}{dP} := c\big(1 + |X_0| + |X_1|\big)^{-1}$$

where c is chosen such that the right-hand side integrates to 1. Clearly, (1.27) holds for $\widetilde{P}$. Moreover, condition (b) of Theorem 1.54 is satisfied by P if and only if it is satisfied by the equivalent measure $\widetilde{P}$. If $P^* \in \mathcal{P}$ is such that the density $dP^*/d\widetilde{P}$ is bounded, then so is the density

$$\frac{dP^*}{dP} = \frac{dP^*}{d\widetilde{P}} \cdot \frac{d\widetilde{P}}{dP}.$$

Therefore, the implication (b) ⇒ (c) holds for P if and only if it holds for $\widetilde{P}$. □

From now on, we will always assume (1.27). Our goal is to construct a suitable $Z \in L^\infty$ such that

$$\frac{dP^*}{dP} := \frac{Z}{E[\,Z\,]}$$

defines an equivalent risk-neutral measure P^*. The following simple lemma gives a criterion for this purpose, involving the convex cone

$$\mathcal{C} := (\mathcal{K} - L^0_+) \cap L^1.$$

Lemma 1.57. *Suppose $c \geq 0$ and $Z \in L^\infty$ are such that*

$$E[\,Z\,W\,] \leq c \quad \textit{for all } W \in \mathcal{C}.$$

Then:

(a) $E[\,Z\,W\,] \leq 0$ *for all $W \in \mathcal{C}$, i.e., we can take $c = 0$.*

(b) $Z \geq 0$ *P-a.s.*

(c) *If Z does not vanish P-a.s., then*

$$\frac{dQ}{dP} := \frac{Z}{E[\,Z\,]}$$

defines a risk-neutral measure $Q \ll P$.

Proof. (a): Note that $\mathcal{C}$ is a cone, i.e., $W \in \mathcal{C}$ implies that $\alpha W \in \mathcal{C}$ for all $\alpha \geq 0$. This property excludes the possibility that $E[\,ZW\,] > 0$ for some $W \in \mathcal{C}$.

(b): $\mathcal{C}$ contains the function $W := -I_{\{Z<0\}}$. Hence, by part (a),

$$E[\,Z^-\,] = E[\,Z\,W\,] \leq 0.$$

(c): For all $\xi \in L^\infty(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ and $\alpha \in \mathbb{R}$ we have $\alpha\xi \cdot Y \in \mathcal{C}$ by our integrability assumption (1.27). Thus, a similar argument as in the proof of (a) yields $E[\,Z\xi \cdot Y\,] = 0$. Since ξ is bounded, we may conclude that

$$0 = E[\,Z\xi \cdot Y\,] = E\big[\,\xi \cdot E[\,ZY \mid \mathcal{F}_0\,]\,\big].$$

As ξ is arbitrary, this yields $E[\,ZY \mid \mathcal{F}_0\,] = 0$ P-almost surely. Proposition A.12 now implies

$$E_Q[\,Y \mid \mathcal{F}_0\,] = \frac{1}{E[\,Z \mid \mathcal{F}_0\,]} E[\,ZY \mid \mathcal{F}_0\,] = 0 \quad Q\text{-a.s.},$$

which concludes the proof. □

In view of the preceding lemma, the construction of risk-neutral measures is reduced to the construction of elements of the set

$$\mathcal{Z} := \left\{ Z \in L^\infty \mid 0 \le Z \le 1,\ P[\, Z > 0\,] > 0,\ \text{and } E[\, Z\, W\,] \le 0 \text{ for all } W \in \mathcal{C} \right\}.$$

In the following lemma, we will construct such elements by applying a separation argument suggested by the condition

$$\mathcal{C} \cap L^1_+ = \{0\},$$

which follows from condition (b) of Theorem 1.54. This separation argument needs the additional assumption that $\mathcal{C}$ is closed in L^1. Showing that this assumption is indeed satisfied in our situation will be one of the key steps in our proof; see Lemma 1.67 below.

Lemma 1.58. *Assume that $\mathcal{C}$ is closed in L^1 and satisfies $\mathcal{C} \cap L^1_+ = \{0\}$. Then for each non-zero $F \in L^1_+$ there exists some $Z \in \mathcal{Z}$ such that $E[\, FZ\,] > 0$.*

Proof. Let $\mathcal{B} := \{F\}$ so that $\mathcal{B} \cap \mathcal{C} = \emptyset$, and note that the set $\mathcal{C}$ is non-empty, convex and closed. Thus we may apply the Hahn–Banach separation theorem in the form of Theorem A.56 to obtain a continuous linear functional ℓ on L^1 such that

$$\sup_{W \in \mathcal{C}} \ell(W) < \ell(F).$$

Since the dual space of L^1 can be identified with L^∞, there exists some $Z \in L^\infty$ such that $\ell(F) = E[\, FZ\,]$ for all $F \in L^1$. We may assume without loss of generality that $\|Z\|_\infty \le 1$. By construction, Z satisfies the assumptions of Lemma 1.57, and so $Z \in \mathcal{Z}$. Moreover, $E[\, FZ\,] = \ell(F) > 0$ since the constant function $W \equiv 0$ is contained in $\mathcal{C}$. □

We will now use an exhaustion argument to conclude that $\mathcal{Z}$ contains a strictly positive element Z^* under the assumptions of Lemma 1.58. After normalization, Z^* will serve as the density of our desired risk-neutral measure $P^* \in \mathcal{P}$.

Lemma 1.59. *Under the assumptions of Lemma* 1.58*, there exists $Z^* \in \mathcal{Z}$ with $Z^* > 0$ P-a.s.*

Proof. As a first step, we claim that $\mathcal{Z}$ is *countably convex*: If $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of non-negative real numbers summing up to 1, and if $Z^{(k)} \in \mathcal{Z}$ for all k, then

$$Z := \sum_{k=1}^{\infty} \alpha_k Z^{(k)} \in \mathcal{Z}.$$

Indeed, for $W \in \mathcal{C}$

$$\sum_{k=1}^{\infty} |\alpha_k Z^{(k)} W| \le |W| \in L^1,$$

and so Lebesgue's dominated convergence theorem implies that

$$E[\,ZW\,] = \sum_{k=1}^{\infty} \alpha_k E[\,Z^{(k)} W\,] \le 0.$$

For the second step, let

$$c := \sup\{P[\,Z > 0\,] \mid Z \in \mathcal{Z}\}.$$

We choose $Z^{(n)} \in \mathcal{Z}$ such that $P[\,Z^{(n)} > 0\,] \to c$. Then

$$Z^* := \sum_{n=1}^{\infty} 2^{-n} Z^{(n)} \in \mathcal{Z}$$

by step one, and

$$\{\,Z^* > 0\,\} = \bigcup_{n=1}^{\infty} \{\,Z^{(n)} > 0\,\}.$$

Hence $P[Z^* > 0] = c$.

In the final step, we show that $c = 1$. Then Z^* will be as desired. Suppose by way of contradiction that $P[Z^* = 0] > 0$, so that $W := \mathrm{I}_{\{Z^*=0\}}$ is a non-zero element of L^1_+. Lemma 1.58 yields $Z \in \mathcal{Z}$ with $E[\,WZ\,] > 0$. Hence,

$$P[\,\{Z > 0\} \cap \{Z^* = 0\}\,] > 0,$$

and so

$$P\left[\,\frac{1}{2}(Z + Z^*) > 0\,\right] > P[\,Z^* > 0\,] = c,$$

in contradiction to the maximality of $P[Z^* > 0]$. □

Thus, we have completed the proof of the implication (b) ⇒(c) of Theorem 1.54 up to the requirement that $\mathcal{C}$ is closed in L^1. Let us pause here in order to state general versions of two of the arguments we have used so far. The first is known as the *Halmos–Savage theorem*. It can be proved by a straightforward modification of the exhaustion argument used in the proof of Lemma 1.59.

Theorem 1.60. *Let $\mathcal{Q}$ be a set of probability measures which are all absolutely continuous with respect to a given measure P. Suppose that $\mathcal{Q} \approx P$ in the sense that $Q[\,A\,] = 0$ for all $Q \in \mathcal{Q}$ implies that $P[\,A\,] = 0$. Then there exists a countable subfamily $\widetilde{\mathcal{Q}} \subset \mathcal{Q}$ which satisfies $\widetilde{\mathcal{Q}} \approx P$.*

An inspection of Lemmas 1.57, 1.58, and 1.59 shows that the particular structure of $\mathcal{C} = (\mathcal{K} - L^0_+) \cap L^1$ was only used for part (c) of Lemma 1.57. All other arguments relied only on the fact that $\mathcal{C}$ is a closed convex cone in L^1 that contains all bounded negative functions and no non-trivial positive function. Thus, we have in fact proved the following *Kreps–Yan theorem*, which was obtained independently in [199] and [137].

Theorem 1.61. *Suppose $\mathcal{C}$ is a closed convex cone in L^1 satisfying*

$$\mathcal{C} \supset -L^\infty_+ \quad \text{and} \quad \mathcal{C} \cap L^1_+ = \{0\}.$$

Then there exists $Z \in L^\infty$ such that $Z > 0$ P-a.s. and $E[\, W Z\,] \le 0$ for all $W \in \mathcal{C}$.

Let us now turn to the closedness of our set $\mathcal{C} = (\mathcal{K} - L^0_+) \cap L^1$. The following example illustrates that we cannot expect $\mathcal{C}$ to be closed without assuming the absence of arbitrage opportunities.

Example 1.62. Let P be the Lebesgue measure on the Borel field $\mathcal{F}_1$ of $\Omega = [0, 1]$, and take $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $Y(\omega) = \omega$. This choice clearly violates the no-arbitrage condition, i.e., we have $\mathcal{K} \cap L^0_+ \neq \{0\}$. The convex set $\mathcal{C} = (\mathcal{K} - L^0_+) \cap L^1$ is a proper subset of L^1. More precisely, $\mathcal{C}$ does not contain any function $F \in L^1$ with $F \ge 1$: If we could represent F as $\xi \cdot Y - U$ for a non-negative function U, then it would follow that

$$\xi \cdot Y = F + U \ge 1,$$

which is impossible for any ξ. However, as we show next, the closure of $\mathcal{C}$ in L^1 coincides with the *full* space L^1. In particular, $\mathcal{C}$ cannot be closed. Let $F \in L^1$ be arbitrary, and observe that

$$F_n := (F^+ \wedge n)\, \mathrm{I}_{[\frac{1}{n},1]} - F^-$$

converges to F in L^1 as $n \uparrow \infty$. Moreover, each F_n belongs to $\mathcal{C}$ as

$$(F^+ \wedge n)\, \mathrm{I}_{[\frac{1}{n},1]} \le n^2 \cdot Y.$$

Consequently, F is contained in the L^1-closure of $\mathcal{C}$. ◇

In the special case $\mathcal{F}_0 = \{\emptyset, \Omega\}$, we can directly go on to the proof that $\mathcal{C}$ is closed, using a simplified version of Lemma 1.67 below. In this way, we obtain an alternative proof of Theorem 1.6. In the general case we need some preparation. Let us first prove a "randomized" version of the Bolzano–Weierstraß theorem. It yields a simple construction of a *measurable* selection of a convergent subsequence of a given sequence in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$.

Lemma 1.63. *Let (ξ_n) be a sequence in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ with $\liminf_n |\xi_n| < \infty$. Then there exists $\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ and a strictly increasing sequence (σ_m) of $\mathcal{F}_0$-measurable integer-valued random variables such that*

$$\xi_{\sigma_m(\omega)}(\omega) \to \xi(\omega) \quad \text{for } P\text{-a.e. } \omega \in \Omega.$$

Proof. Let $\Lambda(\omega) := \liminf_n |\xi_n(\omega)|$, and define $\sigma_m := m$ on the P-null set $\{\Lambda = \infty\}$. On $\{\Lambda < \infty\}$ we let $\sigma^0_1 := 1$, and we define $\mathcal{F}_0$-measurable random indices σ^0_m by

$$\sigma^0_m := \inf\left\{ n > \sigma^0_{m-1} \;\middle|\; \big||\xi_n| - \Lambda\big| \le \frac{1}{m} \right\}, \quad m = 2, 3, \dots.$$

We use recursion on $i = 1, \dots, d$ to define the i^{th} component ξ^i of the limit ξ and to extract a new subsequence σ^i_m of random indices. Let

$$\xi^i = \liminf_{m\uparrow\infty} \xi^i_{\sigma^{i-1}_m},$$

which is already defined if $i = 1$. This ξ^i can be used in the construction of σ^i_m: Let $\sigma^i_1 := 1$ and, for $m = 2, 3, \dots,$

$$\sigma^i_m(\omega) := \inf\left\{\sigma^{i-1}_n(\omega) \;\middle|\; \sigma^{i-1}_n(\omega) > \sigma^i_{m-1}(\omega) \text{ and } |\xi^i_{\sigma^{i-1}_n}(\omega) - \xi^i(\omega)| \le \frac{1}{m}\right\}.$$

Then $\sigma_m := \sigma^d_m$ yields the desired sequence of random indices. □

It may happen that

$$\xi \cdot Y = \widetilde{\xi} \cdot Y \quad P\text{-a.s.},$$

although ξ and $\widetilde{\xi}$ are two different portfolios in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$.

Remark 1.64. We could exclude this possibility by the following assumption of *non-redundance*:

$$\xi \cdot Y = \widetilde{\xi} \cdot Y \ P\text{-a.s.} \quad \Longrightarrow \quad \xi = \widetilde{\xi} \ P\text{-a.s.} \tag{1.28}$$

Under this assumption, we could immediately move on to the final step in Lemma 1.67. ◇

Without assumption (1.28), it will be convenient to have a suitable linear space $N^\perp$ of "reference portfolios" which are uniquely determined by their payoff. The construction of $N^\perp$ is the purpose of the following lemma. We will assume that the spaces L^0 and $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ are endowed with the topology of convergence in P-measure, which is generated by the metric d of (A.22).

Lemma 1.65. *Define two linear subspaces N and $N^\perp$ of $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ by*

$$N := \{\eta \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d) \mid \eta \cdot Y = 0 \ P\text{-a.s.}\},$$
$$N^\perp := \{\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d) \mid \xi \cdot \eta = 0 \ P\text{-a.s. for all } \eta \in N\}.$$

(a) *Both N and $N^\perp$ are closed in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ and, in the following sense, invariant under the multiplication with scalar functions $g \in L^0(\Omega, \mathcal{F}_0, P)$: If $\eta \in N$ and $\xi \in N^\perp$, then $g\eta \in N$ and $g\xi \in N^\perp$.*

(b) *If $\xi \in N^\perp$ and $\xi \cdot Y = 0$ P-a.s., then $\xi = 0$, i.e., $N \cap N^\perp = \{0\}$.*

(c) *Every $\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ has a unique decomposition $\xi = \eta + \xi^\perp$, where $\eta \in N$ and $\xi^\perp \in N^\perp$.*

Remark 1.66. For the proof of this lemma, we will use a projection argument in Hilbert space. Let us sketch a more probabilistic construction of the decomposition $\xi = \eta + \xi^\perp$. Take a regular conditional distribution of Y given $\mathcal{F}_0$, i.e., a stochastic kernel K from $(\Omega, \mathcal{F}_0)$ to $\mathbb{R}^d$ such that $K(\omega, A) = P[\, Y \in A \mid \mathcal{F}_0\,](\omega)$ for all Borel sets $A \subset \mathbb{R}^d$ and P-a.e. ω (see, e.g., §44 of [19]). If one defines $\xi^\perp(\omega)$ as the orthogonal projection of $\xi(\omega)$ onto the linear hull $L(\omega)$ of the support of the measure $K(\omega, \cdot)$, then $\eta := \xi - \xi^\perp$ satisfies $\eta \cdot Y = 0$ P-a.s., and any $\widetilde{\eta}$ with the same property must be P-a.s. perpendicular to $L(\omega)$. However, carrying out the details of this construction involves certain measurability problems; this is why we use the projection argument below. ◇

Proof. (a): The closedness of N and $N^\perp$ follows immediately from the metrizability of $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ (see Appendix A.7) and the fact that every sequence which converges in measure has an almost-surely converging subsequence. The invariance under the multiplication with $\mathcal{F}_0$-measurable scalar functions is obvious.

(b): Suppose that $\xi \in N \cap N^\perp$. Then taking $\eta := \xi$ in the definition of $N^\perp$ yields $\xi \cdot \xi = |\xi|^2 = 0$ P-a.s.

(c): Any given $\xi \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ can be written as

$$\xi(\omega) = \xi^1(\omega)\, e_1 + \cdots + \xi^d(\omega)\, e_d,$$

where e_i denotes the i^{th} Euclidean unit vector, and where $\xi^i(\omega)$ is the i^{th} component of $\xi(\omega)$. Consider e_i as a constant element of $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$, and suppose that we can decompose e_i as

$$e_i = n_i + e_i^\perp \quad \text{where } n_i \in N \text{ and } e_i^\perp \in N^\perp. \tag{1.29}$$

Since by part (a) both N and $N^\perp$ are invariant under the multiplication with $\mathcal{F}_0$-measurable functions, we can then obtain the desired decomposition of ξ by letting

$$\eta(\omega) := \sum_{i=1}^d \xi^i(\omega)\, n_i(\omega) \quad \text{and} \quad \xi^\perp(\omega) := \sum_{i=1}^d \xi^i(\omega)\, e_i^\perp(\omega).$$

Uniqueness of the decomposition follows from $N \cap N^\perp = \{0\}$.

It remains to construct the decomposition (1.29) of e_i. The constant e_i is an element of the space $H := L^2(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$, which becomes a Hilbert space if endowed with the natural inner product

$$(\eta, \xi)_H := E[\, \eta \cdot \xi\,], \quad \eta, \xi \in L^2(\Omega, \mathcal{F}_0, P; \mathbb{R}^d).$$

Observe that both $N \cap H$ and $N^\perp \cap H$ are closed subspaces of H, because convergence in H implies convergence in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$. Therefore, we can define the corresponding orthogonal projections

$$\pi^0 : H \to N \cap H \quad \text{and} \quad \pi^\perp : H \to N^\perp \cap H.$$

Thus, letting $n_i := \pi^0(e_i)$ and $e_i^\perp := \pi^\perp(e_i)$ will be the desired decomposition (1.29), once we know that $e_i = \pi^0(e_i) + \pi^\perp(e_i)$. To prove this, we need only show that $\zeta := e_i - \pi^0(e_i)$ is contained in $N^\perp$. We assume by way of contradiction that ζ is not contained in $N^\perp \cap H$. Then there exists some $\eta \in N$ such that $P[\,\zeta \cdot \eta > 0\,] > 0$. Clearly,

$$\widetilde{\eta} := \eta\, \mathrm{I}_{\{\zeta\cdot\eta>0,\ |\eta|\le c\}}$$

is contained in $N \cap H$ for each $c > 0$. But if c is large enough, then $0 < E[\,\widetilde{\eta} \cdot \zeta\,] = (\widetilde{\eta}, \zeta)_H$, which contradicts the fact that ζ is by construction orthogonal to $N \cap H$. □

After these preparations, we can now complete the proof of Theorem 1.54 by showing the closedness of $\mathcal{C} = (\mathcal{K} - L^0_+) \cap L^1$, which is an immediate consequence of the following lemma. Recall that we have already proved the equivalence of the conditions (a) and (b) in Theorem 1.54.

Lemma 1.67. *If $\mathcal{K} \cap L^0_+ = \{0\}$, then $\mathcal{K} - L^0_+$ is closed in L^0.*

Proof. Suppose $W_n \in (\mathcal{K} - L^0_+)$ converges in L^0 to some W as $n \uparrow \infty$. By passing to a suitable subsequence, we may assume without loss of generality that $W_n \to W$ P-almost surely. We can write $W_n = \xi_n \cdot Y - U_n$ for $\xi_n \in N^\perp$ and $U_n \in L^0_+$.

In a first step, we will prove the assertion given the fact that

$$\liminf_{n\uparrow\infty} |\xi_n| < \infty \quad P\text{-a.s.}, \tag{1.30}$$

which will be established afterwards. Assuming (1.30), Lemma 1.63 yields $\mathcal{F}_0$-measurable integer-valued random variables $\sigma_1 < \sigma_2 < \cdots$ and some $\xi \in L^0(\omega, \mathcal{F}_0, P; \mathbb{R}^d)$ such that P-a.s. $\xi_{\sigma_n} \to \xi$. It follows that

$$U_{\sigma_n} = \xi_{\sigma_n} \cdot Y - W_{\sigma_n} \longrightarrow \xi \cdot Y - W =: U \quad P\text{-a.s.}, \tag{1.31}$$

so that $W = \xi \cdot Y - U \in \mathcal{K} - L^0_+$.

Let us now show that $A := \{\liminf_n |\xi_n| = +\infty\}$ satisfies $P[\,A\,] = 0$ as claimed in (1.30). Using Lemma 1.63 on $\zeta_n := \xi_n/|\xi_n|$ yields $\mathcal{F}_0$-measurable integer-valued random variables $\tau_1 < \tau_2 < \cdots$ and some $\zeta \in L^0(\omega, \mathcal{F}_0, P; \mathbb{R}^d)$ such that P-a.s. $\zeta_{\tau_n} \to \zeta$. The convergence of (W_n) implies that

$$0 \le \mathrm{I}_A \frac{U_{\tau_n}}{|\xi_{\tau_n}|} = \mathrm{I}_A \Big(\zeta_{\tau_n} \cdot Y - \frac{W_{\tau_n}}{|\xi_{\tau_n}|}\Big) \longrightarrow \mathrm{I}_A\, \zeta \cdot Y \quad P\text{-a.s.}$$

Hence, our assumption $\mathcal{K} \cap L^0_+ = \{0\}$ yields $(\mathrm{I}_A \zeta) \cdot Y = 0$. Below we will show that $\mathrm{I}_A \zeta \in N^\perp$, so that

$$\zeta = 0 \quad P\text{-a.s. on } A. \tag{1.32}$$

On the other hand, the fact that $|\zeta_n| = 1$ P-a.s. implies that $|\zeta| = 1$ P-a.s., which can only be consistent with (1.32) if $P[\,A\,] = 0$.

It remains to show that $I_A \zeta \in N^{\perp}$. To this end, we first observe that each ζ_{τ_n} belongs to $N^{\perp}$ since, for each $\eta \in N$,

$$\zeta_{\tau_n} \cdot \eta = \sum_{k=1}^{\infty} I_{\{\tau_n = k\}} \frac{1}{|\xi_k|} \xi_k \cdot \eta = 0 \quad P\text{-a.s.}$$

The closedness of $N^{\perp}$ implies $\zeta \in N^{\perp}$, and $A \in \mathcal{F}_0$ yields $I_A \zeta \in N^{\perp}$. □

If in the proof of Lemma 1.67 $W_n = \xi_n \cdot Y$ for all n, then $U = 0$ in (1.31), and $W = \lim_n W_n$ is itself contained in $\mathcal{K}$. We thus get the following lemma, which will be useful in Chapter 5.

Lemma 1.68. *Suppose that* $\mathcal{K} \cap L^0_+ = \{0\}$. *Then* $\mathcal{K}$ *is closed in* L^0.

In fact, it is possible to show that $\mathcal{K}$ is *always* closed in L^0; see [193], [175]. But this stronger result will not be needed here.

As an alternative to the randomized Bolzano–Weierstraß theorem in Lemma 1.63, we can use the following variant of Komlos' *principle of subsequences*. It yields a convergent sequence of convex combinations of a sequence in $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$, and this will be needed later on. Recall from Appendix A.1 the notion of the *convex hull*

$$\text{conv } A = \left\{ \sum_{i=1}^{n} \alpha_i x_i \;\middle|\; x_i \in A,\ \alpha_i \geq 0,\ \sum_{i=1}^{n} \alpha_i = 1,\ n \in \mathbb{N} \right\}$$

of a subset A of a linear space, which in our case will be $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$.

Lemma 1.69. *Let* (ξ_n) *be a sequence in* $L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$ *such that* $\sup_n |\xi_n| < \infty$ *P-almost surely. Then there exists a sequence of convex combinations*

$$\eta_n \in \text{conv}\{\, \xi_n, \xi_{n+1}, \dots \}$$

which converges P-almost surely to some $\eta \in L^0(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$.

Proof. We can assume without loss of generality that $\sup_n |\xi_n| \leq 1$ P-a.s.; otherwise we consider the sequence $\widetilde{\xi}_n := \xi_n / \sup_n |\xi_n|$. Then (ξ_n) is a bounded sequence in the Hilbert space $H := L^2(\Omega, \mathcal{F}_0, P; \mathbb{R}^d)$. Since the closed unit ball in H is weakly compact, the sequence (ξ_n) has an accumulation point $\eta \in H$; note that weak sequential compactness follows from the Banach–Alaoglu theorem in the form of Theorem A.62 and the fact that the dual H' of the Hilbert space H is isomorphic to H itself. For each n, the accumulation point η belongs to the L^2-closure $\mathcal{C}_n$ of conv$\{\xi_n, \xi_{n+1}, \dots\}$, due to the fact that a closed convex set in H is also weakly closed; see Theorem A.59. Thus, we can find $\eta_n \in \text{conv}\{\xi_n, \xi_{n+1}, \dots\}$ such that

$$E[\, |\eta_n - \eta|^2 \,] \leq \frac{1}{n^2}.$$

This sequence (η_n) converges P-a.s. to η. □

Remark 1.70. The original result by Komlos [133] is more precise: It states that for any bounded sequence (ξ_n) in $L^1(\Omega, \mathcal{F}, P; \mathbb{R}^d)$ there is a subsequence (ξ_{n_k}) which satisfies a strong law of large numbers, i.e.,

$$\lim_{N\uparrow\infty} \frac{1}{N} \sum_{k=1}^{N} \xi_{n_k}$$

exists P-almost surely; see also [195]. ◇

Chapter 2
Preferences

In a complete financial market model, the price of a contingent claim is determined by arbitrage arguments, without involving the preferences of economic agents. In an incomplete model, such claims may carry an intrinsic risk which cannot be hedged away. In order to determine desirable strategies in view of such risks, the preferences of an investor should be made explicit, and this is usually done in terms of an expected utility criterion.

The paradigm of expected utility is the theme of this chapter. We begin with a general discussion of preference relations on a set $\mathcal{X}$ of alternative choices and their numerical representation by some functional U on $\mathcal{X}$. In the financial context, such choices can usually be described as payoff profiles. These are defined as functions X on an underlying set of scenarios with values in some set of payoffs. Thus we are facing risk or even uncertainty. In the case of *risk*, a probability measure is given on the set of scenarios. In this case, we can focus on the resulting payoff distributions. We are then dealing with preferences on "lotteries", i.e., on probability measures on the set of payoffs.

In Sections 2.2 and 2.3 we discuss the conditions – or *axioms* – under which such a preference relation on lotteries μ can be represented by a functional of the form

$$\int u(x)\,\mu(dx),$$

where u is a utility function on the set of payoffs. This formulation of preferences on lotteries in terms of expected utility goes back to D. Bernoulli [22]; the axiomatic theory was initiated by J. von Neumann and O. Morgenstern [155]. Section 2.4 characterizes uniform preference relations which are shared by a given class of functions u. This involves the general theory of probability measures on product spaces with given marginals which will be discussed in Section 2.6.

In Section 2.5 we return to the more fundamental level where preferences are defined on payoff profiles, and where we are facing *uncertainty* in the sense that no probability measure is given a priori. L. Savage [174] clarified the conditions under which such preferences on a space of functions X admit a representation of the form

$$U(X) = E_Q[\,u(X)\,]$$

where Q is a "subjective" probability measure on the set of scenarios. We are going to concentrate on a robust extension of the Savage representation which was introduced by I. Gilboa and D. Schmeidler [104]. Here the utility functional is of the form

$$U(X) = \inf_{Q\in\mathcal{Q}} E_Q[\,u(X)\,],$$

and it involves a whole class $\mathcal{Q}$ of probability measures on the set of scenarios. The axiomatic approach to the robust Savage representation is closely related to the construction of risk measures, which will be the topic of Chapter 4.

2.1 Preference relations and their numerical representation

Let $\mathcal{X}$ be some non-empty set. An element $x \in \mathcal{X}$ will be interpreted as a possible choice of an economic agent. If presented with two choices $x,\ y \in \mathcal{X}$, the agent might prefer one over the other. This will be formalized as follows.

Definition 2.1. A *preference order* (or *preference relation*) on $\mathcal{X}$ is a binary relation $\succ$ with the following two properties.

- *Asymmetry:* If $x \succ y$, then $y \nsucc x$.
- *Negative transitivity*: If $x \succ y$ and $z \in \mathcal{X}$, then either $x \succ z$ or $z \succ y$ or both must hold.

Negative transitivity states that if a clear preference exists between two choices x and y, and if a third choice z is added, then there is still a choice which is least preferable (y if $z \succ y$) or most preferable (x if $x \succ z$).

Definition 2.2. A preference order $\succ$ on $\mathcal{X}$ induces a corresponding *weak preference order* $\succeq$ defined by

$$x \succeq y \;:\Longleftrightarrow\; y \nsucc x,$$

and an *indifference relation* $\sim$ given by

$$x \sim y \;:\Longleftrightarrow\; x \succeq y \text{ and } y \succeq x.$$

Thus, $x \succeq y$ means that either x is preferred to y or there is no clear preference between the two.

Remark 2.3. It is easy to check that the asymmetry and the negative transitivity of $\succ$ are equivalent to the following two respective properties of $\succeq$:

(a) *Completeness:* For all $x, y \in \mathcal{X}$, either $y \succeq x$ or $x \succeq y$ or both are true.

(b) *Transitivity:* If $x \succeq y$ and $y \succeq z$, then also $x \succeq z$.

Conversely, any complete and transitive relation $\succeq$ induces a preference order $\succ$ via the negation of $\succeq$, i.e.,

$$y \succ x \quad :\Longleftrightarrow \quad x \not\succeq y.$$

The indifference relation $\sim$ is an equivalence relation, i.e., it is reflexive, symmetric and transitive. ◇

Definition 2.4. A *numerical representation* of a preference order $\succ$ is a function $U : \mathcal{X} \to \mathbb{R}$ such that

$$y \succ x \iff U(y) > U(x). \tag{2.1}$$

Clearly, (2.1) is equivalent to

$$y \succeq x \iff U(y) \geq U(x).$$

Note that such a numerical representation U is not unique: If f is any strictly increasing function, then $\widetilde{U}(x) := f(U(x))$ is again a numerical representation.

Definition 2.5. Let $\succ$ be a preference relation on $\mathcal{X}$. A subset $\mathcal{Z}$ of $\mathcal{X}$ is called *order dense* if for any pair $x, y \in \mathcal{X}$ such that $x \succ y$ there exists some $z \in \mathcal{Z}$ with $x \succeq z \succeq y$.

The following theorem characterizes those preference relations for which there exists a numerical representation.

Theorem 2.6. *For the existence of a numerical representation of a preference relation $\succ$ it is necessary and sufficient that $\mathcal{X}$ contains a countable, order dense subset $\mathcal{Z}$. In particular, any preference order admits a numerical representation if $\mathcal{X}$ is countable.*

Proof. Suppose first that we are given a countable order dense subset $\mathcal{Z}$ of $\mathcal{X}$. For $x \in \mathcal{X}$, let

$$\overline{\mathcal{Z}}(x) := \{z \in \mathcal{Z} \mid z \succ x\} \quad \text{and} \quad \underline{\mathcal{Z}}(x) := \{z \in \mathcal{Z} \mid x \succ z\}.$$

The relation $x \succeq y$ implies that $\overline{\mathcal{Z}}(x) \subseteq \overline{\mathcal{Z}}(y)$ and $\underline{\mathcal{Z}}(x) \supseteq \underline{\mathcal{Z}}(y)$. If the strict relation $x \succ y$ holds, then at least one of these inclusions is also strict. To see this, pick $z \in \mathcal{Z}$ with $x \succeq z \succeq y$, so that either $x \succ z \succeq y$ or $x \succeq z \succ y$. In the first case, $z \in \underline{\mathcal{Z}}(x) \backslash \underline{\mathcal{Z}}(y)$, while $z \in \overline{\mathcal{Z}}(y) \backslash \overline{\mathcal{Z}}(x)$ in the second case.

Next, take any strictly positive probability distribution μ on $\mathcal{Z}$, and let

$$U(x) := \sum_{z \in \underline{\mathcal{Z}}(x)} \mu(z) - \sum_{z \in \overline{\mathcal{Z}}(x)} \mu(z).$$

By the above, $U(x) > U(y)$ if and only if $x \succ y$ so that U is the desired numerical representation.

For the proof of the converse assertion take a numerical representation U and let $\mathcal{J}$ denote the countable set

$$\mathcal{J} := \Big\{ [a, b] \mid a, b \in \mathbb{Q},\ a < b,\ U^{-1}([a, b]) \neq \emptyset \Big\}.$$

For every interval $I \in \mathcal{J}$ we can choose some $z_I \in \mathcal{X}$ with $U(z_I) \in I$ and thus define the countable set

$$A := \{ z_I \mid I \in \mathcal{J} \}.$$

At first glance it may seem that A is a good candidate for an order dense set. However, it may happen that there are $x, y \in \mathcal{X}$ such that $U(x) < U(y)$ and for which there is no $z \in \mathcal{X}$ with $U(x) < U(z) < U(y)$. In this case, an order dense set must contain at least one z with $U(z) = U(x)$ or $U(z) = U(y)$, a condition which cannot be guaranteed by A.

Let us define the set C of all pairs (x, y) which do not admit any $z \in A$ with $y \succ z \succ x$:

$$C := \Big\{ (x, y) \mid x, y \in \mathcal{X}\backslash A,\ y \succ x \text{ and } \nexists z \in A \text{ with } y \succ z \succ x \Big\}.$$

Then $(x, y) \in C$ implies the apparently stronger fact that we cannot find any $z \in \mathcal{X}$ such that $y \succ z \succ x$: Otherwise we could find $a, b \in \mathbb{Q}$ such that

$$U(x) < a < U(z) < b < U(y),$$

so $I := [a, b]$ would belong to $\mathcal{J}$, and the corresponding z_I would be an element of A with $y \succ z_I \succ x$, contradicting the assumption that $(x, y) \in C$.

It follows that all intervals $\big(U(x), U(y)\big)$ with $(x, y) \in C$ are disjoint and non-empty. Hence, there can be only countably many of them. For each such interval J we pick now exactly one pair $(x^J, y^J) \in C$ such that $U(x^J)$ and $U(y^J)$ are the endpoints of J, and we denote by B the countable set containing all x^J and all y^J.

Finally, we claim that $\mathcal{Z} := A \cup B$ is an order dense subset of $\mathcal{X}$. Indeed, if x, $y \in \mathcal{X}\backslash\mathcal{Z}$ with $y \succ x$, then either there is some $z \in A$ such that $y \succ z \succ x$, or $(x, y) \in C$. In the latter case, there will be some $z \in B$ with $U(y) = U(z) > U(x)$ and, consequently, $y \succeq z \succ x$. □

The following example shows that even in a seemingly straightforward situation, a given preference order may *not* admit a numerical representation.

Example 2.7. Let $\succ$ be the usual lexicographical order on $\mathcal{X} := [0, 1] \times [0, 1]$, i.e., $(x_1, x_2) \succ (y_1, y_2)$ if and only if either $x_1 > y_1$, or if $x_1 = y_1$ and simultaneously $x_2 > y_2$. In order to show that there cannot be a numerical representation for this preference order, suppose on the contrary that U is such a numerical representation. Then

$$d(\alpha) := U(\alpha, 1) - U(\alpha, 0)$$

is strictly positive for all $\alpha \in [0, 1]$. Hence,

$$[0, 1] = \bigcup_{n=1}^{\infty} \Big\{ \alpha \in [0, 1] \ \Big|\ d(\alpha) > \frac{1}{n} \Big\}.$$

Denote $A_n := \{\alpha \mid d(\alpha) > 1/n\}$. There must be at least one set A_{n_0} having infinitely many elements, and we can pick an arbitrary number N of elements $\alpha_1, \dots, \alpha_N \in A_{n_0}$ such that $\alpha_1 < \cdots < \alpha_N$. Since $(\alpha_{i+1}, 0) \succ (\alpha_i, 1)$, it follows that

$$U(\alpha_{i+1}, 0) - U(\alpha_i, 0) > U(\alpha_i, 1) - U(\alpha_i, 0) = d(\alpha_i) > \frac{1}{n_0}.$$

Hence

$$
\begin{aligned}
&U(1,1) - U(0,0)\\
&\quad = U(1,1) - U(\alpha_N, 0) + \sum_{i=1}^{N-1} [\, U(\alpha_{i+1}, 0) - U(\alpha_i, 0)\,] + U(\alpha_1, 0) - U(0,0)\\
&\quad > \frac{N-1}{n_0}.
\end{aligned}
$$

But N can be chosen arbitrarily large while n_0 remains fixed. Thus $U(1,1) - U(0,0)$ must be infinite, which is impossible. ◇

Definition 2.8. Let $\mathcal{X}$ be a topological space. A preference relation $\succ$ is called *continuous* if for all $x \in \mathcal{X}$

$$\overline{\mathcal{B}}(x) := \{y \in \mathcal{X} \mid y \succ x\} \quad \text{and} \quad \underline{\mathcal{B}}(x) := \{y \in \mathcal{X} \mid x \succ y\} \tag{2.2}$$

are open subsets of $\mathcal{X}$.

Remark 2.9. Every preference order that admits a continuous numerical representation is itself continuous. Under some mild conditions on the underlying space $\mathcal{X}$, the converse statement is also true; see Theorem 2.15 below. ◇

Example 2.10. The lexicographical order of Example 2.7 is not continuous: If $(x_1, x_2) \in [0,1] \times [0,1]$ is given, then

$$\big\{\,(y_1, y_2) \mid (y_1, y_2) \succ (x_1, x_2)\,\big\} = (x_1, 1] \times [0,1] \cup \{x_1\} \times (x_2, 1],$$

which is typically not an open subset of $[0,1] \times [0,1]$. ◇

Recall that a topological space $\mathcal{X}$ is called a *topological Hausdorff space* if any two distinct points in $\mathcal{X}$ have disjoint open neighborhoods. In this case, all singletons $\{x\}$ are closed. Clearly, every metric space is a topological Hausdorff space.

Proposition 2.11. *Let $\succ$ be a preference order on a topological Hausdorff space $\mathcal{X}$. Then the following properties are equivalent.*

(a) *$\succ$ is continuous.*

(b) *The set $\{\,(x, y) \mid y \succ x\,\}$ is open in $\mathcal{X} \times \mathcal{X}$.*

(c) *The set $\{\,(x, y) \mid y \succeq x\,\}$ is closed in $\mathcal{X} \times \mathcal{X}$.*

Proof. (a) $\Rightarrow$(b): We have to show that for any pair

$$(x_0, y_0) \in M := \{\,(x, y) \mid y \succ x\,\}$$

there exist open sets $U, V \subset \mathcal{X}$ such that $x_0 \in U$, $y_0 \in V$, and $U \times V \subset M$. Consider first the case in which there exists some $z \in \overline{\mathcal{B}}(x_0) \cap \underline{\mathcal{B}}(y_0)$ for the notation $\overline{\mathcal{B}}(x_0)$

and $\underline{\mathcal{B}}(y_0)$ introduced in (2.2). Then $y_0 \succ z \succ x_0$, so that $U := \underline{\mathcal{B}}(z)$ and $V := \overline{\mathcal{B}}(z)$ are open neighborhoods of x_0 and y_0, respectively. Moreover, if $x \in U$ and $y \in V$, then $y \succ z \succ x$, and thus $U \times V \subset M$.

If $\overline{\mathcal{B}}(x_0) \cap \underline{\mathcal{B}}(y_0) = \emptyset$, we let $U := \underline{\mathcal{B}}(y_0)$ and $V := \overline{\mathcal{B}}(x_0)$. If $(x, y) \in U \times V$, then $y_0 \succ x$ and $y \succ x_0$ by definition. We want to show that $y \succ x$ in order to conclude that $U \times V \subset M$. To this end, suppose that $x \succeq y$. Then $y_0 \succ y$ by negative transitivity, hence $y_0 \succ y \succ x_0$. But then $y \in \overline{\mathcal{B}}(x_0) \cap \underline{\mathcal{B}}(y_0) \neq \emptyset$, and we have a contradiction.

(b)$\Rightarrow$(c): First note that the mapping $\phi(x, y) := (y, x)$ is a homeomorphism of $\mathcal{X} \times \mathcal{X}$. Then observe that the set $\{ (x, y) \mid y \succeq x \}$ is just the complement of the open set $\phi(\{ (x, y) \mid y \succ x \})$.

(c)$\Rightarrow$(a): Since $\mathcal{X}$ is a topological Hausdorff space, $\{x\} \times \mathcal{X}$ is closed in $\mathcal{X} \times \mathcal{X}$, and so is the set

$$\{x\} \times \mathcal{X} \cap \{ (x, y) \mid y \succeq x \} = \{x\} \times \{ y \mid y \succeq x \}.$$

Hence $\{ y \mid y \succeq x \}$ is closed in $\mathcal{X}$, and its complement $\{ y \mid x \succ y \}$ is open. The same argument applies to $\{ y \mid y \succ x \}$. □

Example 2.12. For $x_0 < y_0$ consider the set $\mathcal{X} := (-\infty, x_0] \cup [y_0, \infty)$ endowed with the usual order $>$ on $\mathbb{R}$. Then, with the notation introduced in (2.2), $\underline{\mathcal{B}}(y_0) = (-\infty, x_0]$ and $\overline{\mathcal{B}}(x_0) = [y_0, \infty)$. Hence,

$$\overline{\mathcal{B}}(x_0) \cap \underline{\mathcal{B}}(y_0) = \emptyset$$

despite $y_0 \succ x_0$, a situation we had to consider in the preceding proof. ◇

Recall that the topological space $\mathcal{X}$ is called *connected* if $\mathcal{X}$ cannot be written as the union of two disjoint and non-empty open sets. Assuming that $\mathcal{X}$ is connected will rule out the situation occurring in Example 2.12.

Proposition 2.13. *Let $\mathcal{X}$ be a connected topological space with a continuous preference order $\succ$. Then every dense subset $\mathcal{Z}$ of $\mathcal{X}$ is also order dense in $\mathcal{X}$. In particular, there exists a numerical representation of $\succ$ if $\mathcal{X}$ is separable.*

Proof. Take $x, y \in \mathcal{X}$ with $y \succ x$, and consider $\overline{\mathcal{B}}(x)$ and $\underline{\mathcal{B}}(y)$ as defined in (2.2). Since $y \in \overline{\mathcal{B}}(x)$ and $x \in \underline{\mathcal{B}}(y)$, neither $\overline{\mathcal{B}}(x)$ nor $\underline{\mathcal{B}}(y)$ are empty sets. Moreover, negative transitivity implies that $\mathcal{X} = \overline{\mathcal{B}}(x) \cup \underline{\mathcal{B}}(y)$. Hence, the open sets $\overline{\mathcal{B}}(x)$ and $\underline{\mathcal{B}}(y)$ cannot be disjoint, as $\mathcal{X}$ is connected. Thus, the open set $\overline{\mathcal{B}}(x) \cap \underline{\mathcal{B}}(y)$ must contain some element z of the dense subset $\mathcal{Z}$, which then satisfies $y \succ z \succ x$. Therefore $\mathcal{Z}$ is an order dense subset of $\mathcal{X}$.

Separability of $\mathcal{X}$ means that there exists a countable dense subset $\mathcal{Z}$ of $\mathcal{X}$, which then is order dense. Hence, the existence of a numerical representation follows from Theorem 2.6. □

Remark 2.14. Consider the situation of Example 2.12, where $\mathcal{X} := (-\infty, x_0] \cup [y_0, \infty)$, and suppose that x_0 and y_0 are both irrational. Then $\mathcal{Z} := \mathbb{Q} \cap \mathcal{X}$ is dense in $\mathcal{X}$, but there exists no $z \in \mathcal{Z}$ such that $y_0 \succeq z \succeq x_0$. This example shows that the assumption of topological connectedness is essential for Proposition 2.13. ◇

Theorem 2.15. *Let $\mathcal{X}$ be a topological space which satisfies at least one of the following two properties:*

- *$\mathcal{X}$ has a countable base of open sets.*
- *$\mathcal{X}$ is separable and connected.*

Then every continuous preference order on $\mathcal{X}$ admits a continuous numerical representation.

For a proof we refer to [57], Propositions 3 and 4. For our purposes, namely for the proof of the von Neumann–Morgenstern representation in the next section and for the proof of the robust Savage representation in Section 2.5, the following lemma will be sufficient.

Lemma 2.16. *Let $\mathcal{X}$ be a connected metric space with a continuous preference order $\succ$. If $U : \mathcal{X} \to \mathbb{R}$ is a continuous function, and if its restriction to some dense subset $\mathcal{Z}$ is a numerical representation for the restriction of $\succ$ to $\mathcal{Z}$, then U is also a numerical representation for $\succ$ on $\mathcal{X}$.*

Proof. We have to show that $y \succ x$ if and only if $U(y) > U(x)$. In order to verify the "only if" part, take $x, y \in \mathcal{X}$ with $y \succ x$. As in the proof of Proposition 2.13, we obtain the existence of some $z_0 \in \mathcal{Z}$ with $y \succ z_0 \succ x$. Repeating this argument yields $z_0' \in \mathcal{Z}$ such that $z_0 \succ z_0' \succ x$. Now we take two sequences (z_n) and (z_n') in $\mathcal{Z}$ with $z_n \to y$ and $z_n' \to x$. By continuity of $\succ$, eventually

$$z_n \succ z_0 \succ z_0' \succ z_n',$$

and thus

$$U(z_n) > U(z_0) > U(z_0') > U(z_n').$$

The continuity of U implies that $U(z_n) \to U(y)$ and $U(z_n') \to U(x)$, whence

$$U(y) \geq U(z_0) > U(z_0') \geq U(x).$$

For the proof of the converse implication, suppose that $x, y \in \mathcal{X}$ are such that $U(y) > U(x)$. Since U is continuous,

$$\overline{\mathcal{U}}(x) := \{ z \in \mathcal{X} \mid U(z) > U(x) \}$$

and

$$\underline{\mathcal{U}}(y) := \{ z \in \mathcal{X} \mid U(z) < U(y) \}$$

are both non-empty open subsets of $\mathcal{X}$. Moreover, $\underline{\mathcal{U}}(y) \cup \overline{\mathcal{U}}(x) = \mathcal{X}$. Connectedness of $\mathcal{X}$ implies that $\underline{\mathcal{U}}(y) \cap \overline{\mathcal{U}}(x) \neq \emptyset$. As above, a repeated application of the preceding argument yields $z_0, z_0' \in \mathcal{Z}$ such that

$$U(y) > U(z_0) > U(z_0') > U(x).$$

Since $\mathcal{Z}$ is a dense subset of $\mathcal{X}$, we can find sequences (z_n) and (z_n') in $\mathcal{Z}$ with $z_n \to y$ and $z_n' \to x$ as well as with $U(z_n) > U(z_0)$ and $U(z_n') < U(z_0')$. Since U is a numerical representation of $\succ$ on $\mathcal{Z}$, we have

$$z_n \succ z_0 \succ z_0' \succ z_n'.$$

Hence, by the continuity of $\succ$, neither $z_0 \succ y$ nor $x \succ z_0'$ can be true, and negative transitivity yields $y \succ x$. □

2.2 Von Neumann–Morgenstern representation

Suppose that each possible choice for our economic agent corresponds to a probability distribution on a given set of scenarios. Thus, the set $\mathcal{X}$ can be identified with a subset $\mathcal{M}$ of the set $\mathcal{M}_1(S, \mathcal{S})$ of all probability distributions on a measurable space $(S, \mathcal{S})$. In the context of the theory of choice, the elements of $\mathcal{M}$ are sometimes called *lotteries*. We will assume in the sequel that $\mathcal{M}$ is convex. The aim of this section is to characterize those preference orders $\succ$ on $\mathcal{M}$ which allow for a numerical representation U of the form

$$U(\mu) = \int u(x)\, \mu(dx) \quad \text{for all } \mu \in \mathcal{M}, \tag{2.3}$$

where u is a real function on S.

Definition 2.17. A numerical representation U of a preference order $\succ$ on $\mathcal{M}$ is called a *von Neumann–Morgenstern representation* if it is of the form (2.3).

Any von Neumann–Morgenstern representation U is *affine* on $\mathcal{M}$ in the sense that

$$U(\alpha\mu + (1-\alpha)\nu) = \alpha U(\mu) + (1-\alpha)U(\nu)$$

for all $\mu, \nu \in \mathcal{M}$ and $\alpha \in [0, 1]$. It is easy to check that affinity of U implies the following two properties, or axioms, for a preference order $\succ$ on $\mathcal{M}$. The first property says that a preference $\mu \succ \nu$ is preserved in any convex combination, independent of the context described by another lottery λ.

Definition 2.18. A preference relation $\succ$ on $\mathcal{M}$ satisfies the *independence axiom* if, for all $\mu, \nu \in \mathcal{M}$, the relation $\mu \succ \nu$ implies

$$\alpha\mu + (1-\alpha)\lambda \succ \alpha\nu + (1-\alpha)\lambda$$

for all $\lambda \in \mathcal{M}$ and all $\alpha \in (0, 1]$.

The independence axiom is also called the *substitution axiom*. It can be illustrated by introducing a *compound lottery*, which represents the distribution $\alpha\mu + (1-\alpha)\lambda$ as a two-step procedure. First, we sample either lottery μ or λ with probability α and $1-\alpha$, respectively. Then the lottery drawn in this first step is realized. Clearly, this is equivalent to playing directly the lottery $\alpha\mu + (1-\alpha)\lambda$. With probability $1-\alpha$, the distribution λ is drawn and in this case there is no difference to the compound lottery where ν is replaced by μ. The only difference occurs when μ is drawn, and this happens with probability α. Thus, if $\mu \succ \nu$ then it seems reasonable to prefer the compound lottery with μ over the one with ν.

Definition 2.19. A preference relation $\succ$ on $\mathcal{M}$ satisfies the *Archimedean axiom* if for any triple $\mu \succ \lambda \succ \nu$ there are $\alpha, \beta \in (0, 1)$ such that

$$\alpha\mu + (1-\alpha)\nu \succ \lambda \succ \beta\mu + (1-\beta)\nu.$$

The Archimedean axiom derives its name from its similarity to the Archimedean principle in real analysis: For every small $\varepsilon > 0$ and each large x, there is some $n \in \mathbb{N}$ such that $n\,\varepsilon > x$. Sometimes it is also called the *continuity axiom*, because it can act as a substitute for the continuity of $\succ$ in a suitable topology on $\mathcal{M}$. More precisely, suppose that $\mathcal{M}$ is endowed with a topology for which convex combinations are continuous curves, i.e., $\alpha\mu + (1-\alpha)\nu$ converges to ν or μ as $\alpha \downarrow 0$ or $\alpha \uparrow 1$, respectively. Then continuity of our preference order $\succ$ in this topology automatically implies the Archimedean axiom.

Remark 2.20. As an axiom for consistent behavior in the face of risk, the Archimedean axiom is less intuitive than the independence axiom. Consider the following three deterministic distributions: ν yields 1000 €, λ yields 10 €, and μ is the lottery where one dies for sure. Even for small $\alpha \in (0, 1)$ it is not clear that someone would prefer the gamble $\alpha\mu + (1-\alpha)\nu$, which involves the probability α of dying, over the conservative 10 € yielded by λ. Note, however, that most people would not hesitate to drive a car for a distance of 50 km in order to receive a premium of 1000 €, even though this might involve the risk of a deadly accident. ◇

Our first goal is to show that the Archimedean axiom and the independence axiom imply the existence of an affine numerical representation.

Theorem 2.21. *Suppose that $\succ$ is a preference relation on $\mathcal{M}$ satisfying both the Archimedean and the independence axiom. Then there exists an affine numerical representation U of $\succ$. Moreover, U is unique up to positive affine transformations, i.e., any other affine numerical representation $\widetilde{U}$ with these properties is of the form $\widetilde{U} = a\,U + b$ for some $a > 0$ and $b \in \mathbb{R}$.*

In two important cases, such an affine numerical representation will already be of von Neumann–Morgenstern form. This is the content of the following two corollaries, which we state before proving Theorem 2.21. For the first corollary, we need the notion of a *simple probability distribution*. This is a probability measure μ on S which can be written as a finite convex combination of Dirac masses, i.e., there exist $x_1, \dots, x_N \in S$ and $\alpha_1, \dots, \alpha_N \in (0, 1]$ such that

$$\mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}.$$

Corollary 2.22. *Suppose that $\mathcal{M}$ is the set of all simple probability distributions on S and that $\succ$ is a preference order on $\mathcal{M}$ that satisfies both the Archimedean and the independence axiom. Then there exists a von Neumann–Morgenstern representation U. Moreover, both U and u are unique up to positive affine transformations.*

Proof. Let U be an affine numerical representation, which exists by Theorem 2.21. We define $u(x) := U(\delta_x)$, for $x \in S$. If $\mu \in \mathcal{M}$ is of the form $\mu = \alpha_1\delta_{x_1} + \cdots + \alpha_N \delta_{x_N}$, then affinity of U implies

$$U(\mu) = \sum_{i=1}^{N} \alpha_i U(\delta_{x_i}) = \int u(x)\, \mu(dx).$$

This is the desired von Neumann–Morgenstern representation. □

On a finite set S, every probability measure is simple. Thus, we obtain the following result as a special case.

Corollary 2.23. *Suppose that $\mathcal{M}$ is the set of all probability distributions on a finite set S and that $\succ$ is a preference order on $\mathcal{M}$ that satisfies both the Archimedean and the independence axiom. Then there exists a von Neumann–Morgenstern representation, and it is unique up to positive affine transformations.*

For the proof of Theorem 2.21, we need the following auxiliary lemma. Its first assertion states that taking convex combination is monotone with respect to a preference order $\succ$ satisfying our two axioms. Its second part can be regarded as an "intermediate value theorem" for straight lines in $\mathcal{M}$, and (c) is the analogue of the independence axiom for the indifference relation $\sim$.

Lemma 2.24. *Under the assumptions of Theorem* 2.21*, the following assertions are true.*

(a) *If $\mu \succ \nu$, then $\alpha \mapsto \alpha\mu + (1-\alpha)\nu$ is strictly increasing with respect to $\succ$. More precisely, $\beta\mu + (1-\beta)\nu \succ \alpha\mu + (1-\alpha)\nu$ for $0 \le \alpha < \beta \le 1$.*

(b) *If $\mu \succ \nu$ and $\mu \succeq \lambda \succeq \nu$, then there exists a unique $\alpha \in [0, 1]$ with $\lambda \sim \alpha\mu + (1-\alpha)\nu$.*

(c) *If $\mu \sim \nu$, then $\alpha\mu + (1-\alpha)\lambda \sim \alpha\nu + (1-\alpha)\lambda$ for all $\alpha \in [0, 1]$ and all $\lambda \in \mathcal{M}$.*

Proof. (a): Let $\lambda := \beta\mu + (1-\beta)\nu$. The independence axiom implies that $\lambda \succ \beta\nu + (1-\beta)\nu = \nu$. Hence, for $\gamma := \alpha/\beta$,

$$\beta\mu + (1-\beta)\nu = (1-\gamma)\lambda + \gamma\lambda \succ (1-\gamma)\nu + \gamma\lambda = \alpha\mu + (1-\alpha)\nu.$$

(b): Part (a) guarantees that α is unique if it exists. To show existence, we need only to consider the case $\mu \succ \lambda \succ \nu$, for otherwise we can take either $\alpha = 0$ or $\alpha = 1$. The natural candidate is

$$\alpha := \sup\{\, \gamma \in [0,1] \mid \lambda \succeq \gamma\mu + (1-\gamma)\nu \,\}.$$

If $\lambda \sim \alpha\mu + (1-\alpha)\nu$ is not true, then one of the following two possibilities must occur:

$$\lambda \succ \alpha\mu + (1-\alpha)\nu, \quad \text{or} \quad \lambda \prec \alpha\mu + (1-\alpha)\nu. \tag{2.4}$$

In the first case, we apply the Archimedean axiom to obtain some $\beta \in (0,1)$ such that

$$\lambda \succ \beta\big[\,\alpha\mu + (1-\alpha)\nu\,\big] + (1-\beta)\mu = \gamma\mu + (1-\gamma)\nu \tag{2.5}$$

for $\gamma = 1 - \beta(1-\alpha)$. Since $\gamma > \alpha$, it follows from the definition of α that $\gamma\mu + (1-\gamma)\nu \succ \lambda$, which contradicts (2.5). If the second case in (2.4) occurs, the Archimedean axiom yields some $\beta \in (0,1)$ such that

$$\beta\big(\alpha\mu + (1-\alpha)\nu\big) + (1-\beta)\nu = \beta\alpha\mu + (1-\beta\alpha)\nu \succ \lambda. \tag{2.6}$$

Clearly $\beta\alpha < \alpha$, so that the definition of α yields some $\gamma \in (\beta\alpha, \alpha]$ with $\lambda \succeq \gamma\mu + (1-\gamma)\nu$. Part (a) and the fact that $\beta\alpha < \gamma$ imply that

$$\lambda \succeq \gamma\mu + (1-\gamma)\nu \succ \beta\alpha\mu + (1-\beta\alpha)\nu,$$

which contradicts (2.6).

(c): We must exclude both of the following two possibilities

$$\alpha\mu + (1-\alpha)\lambda \succ \alpha\nu + (1-\alpha)\lambda \quad \text{and} \quad \alpha\nu + (1-\alpha)\lambda \succ \alpha\mu + (1-\alpha)\lambda. \tag{2.7}$$

To this end, we may assume that there exists some $\rho \in \mathcal{M}$ with $\rho \not\sim \mu \sim \nu$; otherwise the result is trivial. Let us assume that $\rho \succ \mu \sim \nu$; the case in which $\mu \sim \nu \succ \rho$ is similar. Suppose that the first possibility in (2.7) would occur. The independence axiom yields

$$\beta\rho + (1-\beta)\nu \succ \beta\nu + (1-\beta)\nu = \nu \sim \mu$$

for all $\beta \in (0,1)$. Therefore,

$$\alpha\big[\,\beta\rho + (1-\beta)\nu\,\big] + (1-\alpha)\lambda \succ \alpha\mu + (1-\alpha)\lambda \quad \text{for all } \beta \in (0,1). \tag{2.8}$$

Using our assumption that the first possibilities in (2.7) is occurring, we obtain from part (b) a unique $\gamma \in (0,1)$ such that, for any fixed β,

$$\begin{aligned}\alpha\mu + (1-\alpha)\lambda &\sim \gamma\big(\alpha\big[\,\beta\rho + (1-\beta)\nu\,\big] + (1-\alpha)\lambda\big) + (1-\gamma)[\,\alpha\nu + (1-\alpha)\lambda\,]\\ &= \alpha\big[\,\beta\gamma\rho + (1-\beta\gamma)\nu\,\big] + (1-\alpha)\lambda\\ &\succ \alpha\mu + (1-\alpha)\lambda,\end{aligned}$$

where we have used (2.8) for β replaced by $\beta\gamma$ in the last step. This is a contradiction. The second possibility in (2.7) is excluded by an analogous argument. □

Proof of Theorem 2.21. For the construction of U, we first fix two lotteries λ and ρ with $\lambda \succ \rho$ and define

$$\mathcal{M}(\lambda,\rho) := \{\mu \in \mathcal{M} \mid \lambda \succeq \mu \succeq \rho\};$$

the assertion is trivial if no such pair $\lambda \succ \rho$ exists. If $\mu \in \mathcal{M}(\lambda,\rho)$, part (b) of Lemma 2.24 yields a unique $\alpha \in [0,1]$ such that $\mu \sim \alpha\lambda + (1-\alpha)\rho$, and we put $U(\mu) := \alpha$. To prove that U is a numerical representation of $\succ$ on $\mathcal{M}(\lambda,\rho)$, we must show that for $\nu, \mu \in \mathcal{M}(\lambda,\rho)$ we have $U(\mu) > U(\nu)$ if and only if $\mu \succ \nu$. To prove sufficiency, we apply part (a) of Lemma 2.24 to conclude that

$$\mu \sim U(\mu)\lambda + \big(1 - U(\mu)\big)\rho \succ U(\nu)\lambda + \big(1 - U(\nu)\big)\rho \sim \nu,$$

Hence $\mu \succ \nu$. Conversely, if $\mu \succ \nu$ then the preceding arguments already imply that we cannot have $U(\nu) > U(\mu)$. Thus, it suffices to rule out the case $U(\mu) = U(\nu)$. But if $U(\mu) = U(\nu)$, then the definition of U yields $\mu \sim \nu$, which contradicts $\mu \succ \nu$. We conclude that U is indeed a numerical representation of $\succ$ restricted to $\mathcal{M}(\lambda,\rho)$.

Let us now show that $\mathcal{M}(\lambda,\rho)$ is a convex set. Take $\mu, \nu \in \mathcal{M}(\lambda,\rho)$ and $\alpha \in [0,1]$. Then

$$\lambda \succeq \alpha\lambda + (1-\alpha)\nu \succeq \alpha\mu + (1-\alpha)\nu,$$

using the independence axiom to handle the cases $\lambda \succ \nu$ and $\lambda \succ \mu$, and part (c) of Lemma 2.24 for $\lambda \sim \nu$ and for $\lambda \sim \mu$. By the same argument it follows that $\alpha\mu + (1-\alpha)\nu \succeq \rho$, which implies the convexity of the set $\mathcal{M}(\lambda,\rho)$.

Therefore, $U(\alpha\mu + (1-\alpha)\nu)$ is well defined; we proceed to show that it equals $\alpha U(\mu) + (1-\alpha)U(\nu)$. To this end, we apply part (c) of Lemma 2.24 twice:

$$\begin{aligned}\alpha\mu + (1-\alpha)\nu &\sim \alpha\big(U(\mu)\lambda + \big(1-U(\mu)\big)\rho\big) + (1-\alpha)\big(U(\nu)\lambda + \big(1-U(\nu)\big)\rho\big)\\ &= \big[\alpha U(\mu) + (1-\alpha)U(\nu)\big]\lambda + \big[1 - \alpha U(\mu) - (1-\alpha)U(\nu)\big]\rho.\end{aligned}$$

The definition of U and the uniqueness in part (b) of Lemma 2.24 imply that

$$U(\alpha\mu + (1-\alpha)\nu) = \alpha U(\mu) + (1-\alpha)U(\nu).$$

So U is indeed an affine numerical representation of $\succ$ on $\mathcal{M}(\lambda,\rho)$.

In a further step, we now show that the affine numerical representation U on $\mathcal{M}(\lambda,\rho)$ is unique up to positive affine transformations. So let $\widetilde{U}$ be another affine numerical representation of $\succ$ on $\mathcal{M}(\lambda,\rho)$, and define

$$\hat{U}(\mu) := \frac{\widetilde{U}(\mu) - \widetilde{U}(\rho)}{\widetilde{U}(\lambda) - \widetilde{U}(\rho)}, \quad \mu \in \mathcal{M}(\lambda,\rho).$$

Then $\hat{U}$ is a positive affine transformation of $\widetilde{U}$, and $\hat{U}(\rho) = 0 = U(\rho)$ as well as $\hat{U}(\lambda) = 1 = U(\lambda)$. Hence, affinity of $\hat{U}$ and the definition of U imply

$$\hat{U}(\mu) = \hat{U}\big(U(\mu)\lambda + \big(1 - U(\mu)\big)\rho\big) = U(\mu)\hat{U}(\lambda) + \big(1 - U(\mu)\big)\hat{U}(\rho) = U(\mu)$$

for all $\mu \in \mathcal{M}(\lambda, \rho)$. Thus $\hat{U} = U$.

Finally, we have to show that U can be extended as a numerical representation to the full space $\mathcal{M}$. To this end, we first take $\widetilde{\lambda}, \widetilde{\rho} \in \mathcal{M}$ such that $\mathcal{M}(\widetilde{\lambda}, \widetilde{\rho}) \supset \mathcal{M}(\lambda, \rho)$. By the arguments in the first part of this proof, there exists an affine numerical representation $\widetilde{U}$ of $\succ$ on $\mathcal{M}(\widetilde{\lambda}, \widetilde{\rho})$, and we may assume that $\widetilde{U}(\lambda) = 1$ and $\widetilde{U}(\rho) = 0$; otherwise we apply a positive affine transformation to $\widetilde{U}$. By the previous step of the proof, $\widetilde{U}$ coincides with U on $\mathcal{M}(\lambda, \rho)$, and so $\widetilde{U}$ is the unique consistent extension of U. Since each lottery belongs to some set $\mathcal{M}(\widetilde{\lambda}, \widetilde{\rho})$, the affine numerical representation U can be uniquely extended to all of $\mathcal{M}$. □

Remark 2.25. In the proof of the preceding theorem, we did not use the fact that the elements of $\mathcal{M}$ are probability measures. All that was needed was convexity of the set $\mathcal{M}$, the Archimedean, and the independence axiom. Yet, even the concept of convexity can be generalized by introducing the notion of a *mixture space*; see, e.g., [138], [84], or [112]. ◇

Let us now return to the problem of constructing a von Neumann–Morgenstern representation for preference relations on distributions. If $\mathcal{M}$ is the set of all probability measures on a finite set S, any affine numerical representation is already of this form, as we saw in the proof of Corollary 2.23. However, the situation becomes more involved if we take an infinite set S. In fact, the following examples show that in this case a von Neumann–Morgenstern representation may not exist.

Example 2.26. Let $\mathcal{M}$ be the set of probability measures μ on $S := \{1, 2, \dots\}$ for which $U(\mu) := \limsup_{k\uparrow\infty} k\,\mu(k)$ is finite. Clearly, U is affine and induces a preference order on $\mathcal{M}$ which satisfies both the Archimedean and the independence axiom. However, U obviously does not admit a von Neumann–Morgenstern representation. ◇

Example 2.27. Let $\mathcal{M}$ be set the of all Borel probability measures on $S = [0, 1]$, and denote by λ the Lebesgue measure on S. According to the Lebesgue decomposition theorem, which is recalled in Theorem A.13, every $\mu \in \mathcal{M}$ can be decomposed as

$$\mu = \mu_s + \mu_a,$$

where μ_s is singular with respect to λ, and μ_a is absolutely continuous. We define a function $U : \mathcal{M} \to [0, 1]$ by

$$U(\mu) := \int x\mu_a(dx).$$

It is easily seen that U is an affine function on $\mathcal{M}$. Hence, U induces a preference order $\succ$ on $\mathcal{M}$ which satisfies both the Archimedean and the independence axioms. But $\succ$ cannot have a von Neumann–Morgenstern representation: Since $U(\delta_x) = 0$ for all x, the only possible choice for u in (2.3) would be $u \equiv 0$. So the preference relation would be trivial in the sense that $\mu \sim \lambda$ for all $\mu \in \mathcal{M}$, in contradiction for instance to $U(\lambda) = \frac{1}{2}$ and $U(\delta_{\frac{1}{2}}) = 0$. ◇

One way to obtain a von Neumann–Morgenstern representation is to assume additional continuity properties of $\succ$, where continuity is understood in the sense of Definition 2.8. As we have already remarked, the Archimedean axiom holds automatically if taking convex combinations is continuous for the topology on $\mathcal{M}$. This is indeed the case for the weak topology on the set $\mathcal{M}_1(S, \mathcal{S})$ of all probability measures on a separable metric space S, endowed with the σ-field $\mathcal{S}$ of Borel sets. The space S will be fixed for the rest of this section, and we will simply write $\mathcal{M}_1(S) = \mathcal{M}_1(S, \mathcal{S})$.

Theorem 2.28. *Let $\mathcal{M} := \mathcal{M}_1(S)$ be the space of all probability measures on S endowed with the weak topology, and let $\succ$ be a continuous preference order on $\mathcal{M}$ satisfying the independence axiom. Then there exists a von Neumann–Morgenstern representation*

$$U(\mu) = \int u(x)\, \mu(dx)$$

for which the function $u : S \to \mathbb{R}$ is bounded and continuous. Moreover, U and u are unique up to positive affine transformations.

Proof. Let $\mathcal{M}_s$ denote the set of all simple probability distributions on S. Since continuity of $\succ$ implies the Archimedean axiom, we deduce from Corollary 2.22 that $\succ$ restricted to $\mathcal{M}_s$ has a von Neumann–Morgenstern representation.

Let us show that the function u in this representation is bounded. For instance, if u is not bounded from above, then there are $x_0, x_1, \ldots \in S$ such that $u(x_0) < u(x_1)$ and $u(x_n) > n$. Now let

$$\mu_n := \left(1 - \frac{1}{\sqrt{n}}\right)\delta_{x_0} + \frac{1}{\sqrt{n}}\delta_{x_n}.$$

Clearly, $\mu_n \to \delta_{x_0}$ weakly as $n \uparrow \infty$. The continuity of $\succ$ together with the assumption that $\delta_{x_1} \succ \delta_{x_0}$ imply that $\delta_{x_1} \succ \mu_n$ for all large n. However, $U(\mu_n) > \sqrt{n}$ for all n, in contradiction to $\delta_{x_1} \succ \mu_n$.

Suppose that the function u is not continuous. Then there exists some $x \in S$ and a sequence $(x_n)_{n\in\mathbb{N}} \subset S$ such that $x_n \to x$ but $u(x_n) \not\to u(x)$. By taking a subsequence if necessary, we can assume that $u(x_n)$ converges to some number $a \neq u(x)$. Suppose that $u(x) - a =: \varepsilon > 0$. Then there exists some m such that $|u(x_n) - a| < \varepsilon/3$ for all $n \geq m$. Let $\mu := \frac{1}{2}(\delta_x + \delta_{x_m})$. For all $n \geq m$

$$U(\delta_x) = a + \varepsilon > a + \frac{2\varepsilon}{3} > \frac{1}{2}\big(u(x) + u(x_m)\big) = U(\mu) > a + \frac{\varepsilon}{3} > U(\delta_{x_n}).$$

Therefore $\delta_x \succ \mu \succ \delta_{x_n}$, although δ_{x_n} converges weakly to δ_x, in contradiction to the continuity of $\succ$. The case $u(x) < a$ is excluded in the same manner.

Let us finally show that

$$U(\mu) := \int u(x)\,\mu(dx) \quad \text{for } \mu \in \mathcal{M}$$

defines a numerical representation of $\succ$ on all of $\mathcal{M}$. Since u is bounded and continuous, U is continuous with respect to the weak topology on $\mathcal{M}$. Moreover, Theorem A.37 states that $\mathcal{M}_s$ is a dense subset of the connected metrizable space $\mathcal{M}$. So the proof is completed by an application of Lemma 2.16. □

The scope of the preceding theorem is limited insofar as it involves only *bounded* functions u. This will not be flexible enough for our purposes. In the next section, for instance, we will consider risk-averse preferences which are defined in terms of *concave* functions u on the space $S = \mathbb{R}$. Such a function cannot be bounded unless it is constant. Thus, we must relax the conditions of the previous theorem. We will present two approaches. In our first approach, we fix some point $x_0 \in S$ and denote by $\overline{B}_r(x_0)$ the closed metric ball of radius r around x_0. The space of *boundedly supported measures* on S is given by

$$\begin{aligned} \mathcal{M}_b(S) &:= \bigcup_{r>0} \mathcal{M}_1\big(\overline{B}_r(x_0)\big) \\ &= \big\{\, \mu \in \mathcal{M}_1(S) \mid \mu\big(\overline{B}_r(x_0)\big) = 1 \text{ for some } r \ge 0 \,\big\}. \end{aligned}$$

Clearly, this definition does not depend on the particular choice of x_0.

Corollary 2.29. *Let $\succ$ be a preference order on $\mathcal{M}_b(S)$ whose restriction to each space $\mathcal{M}_1\big(\overline{B}_r(x_0)\big)$ is continuous with respect to the weak topology. If $\succ$ satisfies the independence axiom, then there exists a von Neumann–Morgenstern representation*

$$U(\mu) = \int u(x)\,\mu(dx)$$

with a continuous function $u : S \to \mathbb{R}$. Moreover, U and u are unique up to positive affine transformations.

Proof. Theorem 2.28 yields a von Neumann–Morgenstern representation of the restriction of $\succ$ to $\mathcal{M}_1\big(\overline{B}_r(x_0)\big)$ in terms of some continuous function $u_r : \overline{B}_r(x_0) \to \mathbb{R}$. The uniqueness part of the theorem implies that the restriction of u_r to some smaller ball $\overline{B}_{r'}(x_0)$ must be a equal to $u_{r'}$ up to a positive affine transformation. Thus, it is possible to find a unique continuous extension $u : S \to \mathbb{R}$ of $u_{r'}$ which defines a von Neumann–Morgenstern representation of $\succ$ on each set $\mathcal{M}_1\big(\overline{B}_r(x_0)\big)$. □

Our second variant of Theorem 2.28 includes measures with *unbounded* support, but we need stronger continuity assumptions. Let ψ be a continuous function with values in $[1, \infty)$ on the separable metric space S. We use ψ as a *gauge function* and define

$$\mathcal{M}_1^{\psi}(S) := \Big\{ \mu \in \mathcal{M}_1(S) \;\Big|\; \int \psi(x)\,\mu(dx) < \infty \Big\}.$$

A suitable space of continuous test functions for measures in $\mathcal{M}_1^{\psi}(S)$ is provided by

$$C_{\psi}(S) := \big\{ f \in C(S) \mid \exists\, c : \; |f(x)| \le c \cdot \psi(x) \text{ for all } x \in S \big\}.$$

These test functions can now be used to define a topology on $\mathcal{M}_1^{\psi}(S)$ in precisely the same way one uses the set of bounded continuous function to define the weak topology: A sequence (μ_n) in $\mathcal{M}_1^{\psi}(S)$ converges to some $\mu \in \mathcal{M}_1^{\psi}(S)$ if and only if

$$\int f\, d\mu_n \longrightarrow \int f\, d\mu \quad \text{for all } f \in C_{\psi}(S).$$

To be rigorous, one should first define a neighborhood base for the topology and then check that this topology is metrizable, so that it suffices indeed to consider the convergence of sequences; the reader will find all necessary details in Appendix A.6. We will call this topology the *ψ-weak topology* on $\mathcal{M}_1^{\psi}(S)$. If we take the trivial case $\psi \equiv 1$, $C_{\psi}(S)$ consists of all bounded continuous functions, and we recover the standard weak topology on $\mathcal{M}_1^{1}(S) = \mathcal{M}_1(S)$. However, by taking ψ as some non-bounded function, we can also include von Neumann–Morgenstern representations in terms of unbounded functions u. The following theorem is a version of Theorem 2.28 for the ψ-weak topology. Its proof is analogous to that of Theorem 2.28, and we leave it to the reader to fill in the details.

Theorem 2.30. *Let $\succ$ be a preference order on $\mathcal{M}_1^{\psi}(S)$ that is continuous in the ψ-weak topology and satisfies the independence axiom. Then there exists a numerical representation U of von Neumann–Morgenstern form*

$$U(\mu) = \int u(x)\,\mu(dx)$$

with a function $u \in C_{\psi}(S)$. Moreover, U and u are unique up to positive affine transformations.

Remark 2.31. Instead of making topological assumptions on $\succ$, one can introduce, in addition to the Archimedean and the independence axiom, the so-called *sure-thing principle*: For $\mu, \nu \in \mathcal{M}$ and $A \in \mathcal{S}$ such that $\mu(A) = 1$:

$$\delta_x \succ \nu \text{ for all } x \in A \quad \Longrightarrow \quad \mu \succ \nu,$$

and

$$\nu \succ \delta_x \text{ for all } x \in A \quad \Longrightarrow \quad \nu \succ \mu.$$

This axiom, together with a couple of technical assumptions, guarantees the existence of a von Neumann–Morgenstern representation; see [84]. Conversely, it is easy to see that the sure-thing principle is automatically implied by the existence of a von Neumann–Morgenstern representation. Note that the sure-thing principle is violated in both Examples 2.26 and 2.27. ◇

So far, we have presented the classical theory of expected utility, starting with the independence axiom and the Archimedean axiom. However, it is well known that in reality people may not behave according to this paradigm.

Example 2.32 (Allais Paradox). The so-called *Allais paradox* questions the descriptive aspect of expected utility by considering the following lotteries. Lottery

$$\nu_1 = 0.33\,\delta_{2500} + 0.66\,\delta_{2400} + 0.01\,\delta_0$$

yields 2500 € with a probability of 0.33, 2400 € with probability 0.66, and draws a blank with the remaining probability of 0.01. Lottery

$$\mu_1 := \delta_{2400}$$

yields 2400 € for sure. When asked, most people prefer the sure amount – even though lottery ν_1 has the larger expected value, namely 2409 €.

Next, consider the following two lotteries μ_2 and ν_2:

$$\mu_2 := 0.34\,\delta_{2400} + 0.66\,\delta_0 \quad \text{and} \quad \nu_2 := 0.33\,\delta_{2500} + 0.67\,\delta_0.$$

Here people tend to prefer the slightly riskier lottery ν_2 over μ_2, in accordance with the expectations of ν_2 and μ_2, which are 825 € and 816 €, respectively.

This observation is due to M. Allais [4]. It was confirmed by D. Kahnemann and A. Tversky [123] in empirical tests where 82 % of interviewees preferred μ_1 over ν_1 while 83 % chose ν_2 rather than μ_2. This means that at least 65 % chose both $\mu_1 \succ \nu_1$ and $\nu_2 \succ \mu_2$. As pointed out by M. Allais, this simultaneous choice leads to a "paradox" in the sense that it is inconsistent with the von Neumann–Morgenstern paradigm. More precisely, any preference relation $\succ$ for which $\mu_1 \succ \nu_1$ and $\nu_2 \succ \mu_2$ are both valid violates the independence axiom, as we will show now. If the independence axiom were satisfied, then necessarily

$$\alpha\mu_1 + (1-\alpha)\nu_2 \succ \alpha\nu_1 + (1-\alpha)\nu_2 \succ \alpha\nu_1 + (1-\alpha)\mu_2$$

for all $\alpha \in (0, 1)$. By taking $\alpha = 1/2$ we would arrive at

$$\frac{1}{2}(\mu_1 + \nu_2) \succ \frac{1}{2}(\nu_1 + \mu_2)$$

which is a contradiction to the fact that

$$\frac{1}{2}(\mu_1 + \nu_2) = \frac{1}{2}(\nu_1 + \mu_2).$$

Therefore, the independence axiom was violated by at least 65 % of the people who were interviewed. This effect is empirical evidence against the von Neumann–Morgenstern theory as a *descriptive* theory. Even from a *normative* point of view, there are good reasons to go beyond our present setting, and this will be done in Section 2.5. In particular, we will take a second look at the Allais paradox in Remark 2.74. ◇

2.3 Expected utility

In this section, we focus on individual financial assets under the assumption that their payoff distributions at a fixed time are known, and without any regard to hedging opportunities in the context of a financial market model. Such asset distributions may be viewed as lotteries with monetary outcomes in some interval $S \subset \mathbb{R}$. Thus, we take $\mathcal{M}$ as a fixed set of Borel probability measures on S. In this setting, we discuss the paradigm of expected utility in its standard form, where the function u appearing in the von Neumann–Morgenstern representation has additional properties suggested by the monetary interpretation. We introduce risk aversion and certainty equivalents, and illustrate these notions with a number of examples.

Throughout this section, we assume that $\mathcal{M}$ is convex and contains all point masses δ_x for $x \in S$. We assume also that each $\mu \in \mathcal{M}$ has a well-defined expectation

$$m(\mu) := \int x \, \mu(dx) \in \mathbb{R}.$$

Remark 2.33. For an asset whose (discounted) random payoff has a known distribution μ, the expected value $m(\mu)$ is often called the *fair price* of the asset. For an insurance contract where μ is the distribution of payments to be received by the insured party in dependence of some random damage within a given period, the expected value $m(\mu)$ is also called the *fair premium*. Typically, actual asset prices and actual insurance premiums will be different from these values. In many situations, such differences can be explained within the conceptual framework of expected utility, and in particular in terms of *risk aversion*. ◇

Definition 2.34. A preference relation $\succ$ on $\mathcal{M}$ is called *monotone* if

$$x > y \text{ implies } \delta_x \succ \delta_y.$$

The preference relation is called *risk averse* if for $\mu \in \mathcal{M}$

$$\delta_{m(\mu)} \succ \mu \quad \text{unless } \mu = \delta_{m(\mu)}.$$

It is easy to characterize these properties within the class of preference relations which admit a von Neumann–Morgenstern representation.

Proposition 2.35. *Suppose the preference relation $\succ$ has a von Neumann–Morgenstern representation*

$$U(\mu) = \int u \, d\mu.$$

Then:

(a) *$\succ$ is monotone if and only if u is strictly increasing.*

(b) *$\succ$ is risk averse if and only if u is strictly concave.*

Proof. (a): Monotonicity is equivalent to

$$u(x) = U(\delta_x) > U(\delta_y) = u(y) \quad \text{for } x > y.$$

(b): If $\succ$ is risk-averse, then

$$\delta_{\alpha x+(1-\alpha)y} \succ \alpha\,\delta_x + (1-\alpha)\delta_y$$

holds for all distinct $x,\ y \in S$ and $\alpha \in (0,1)$. Hence,

$$u\big(\alpha x + (1-\alpha)y\big) > \alpha\, u(x) + (1-\alpha)u(y),$$

i.e., u is strictly concave. Conversely, if u is strictly concave, then Jensen's inequality implies risk aversion:

$$U(\delta_{m(\mu)}) = u\Big(\int x\,\mu(dx)\Big) \geq \int u(x)\,\mu(dx) = U(\mu)$$

with equality if and only if $\mu = \delta_{m(\mu)}$. □

Remark 2.36. In view of the monetary interpretation of the state space S, it is natural to assume that the preference relation $\succ$ is *monotone*. The assumption of *risk aversion* is more debatable, at least from a *descriptive* point of view. In fact, there is considerable empirical evidence that agents tend to switch between risk aversion and risk seeking behavior, depending on the context. In particular, they may be risk averse after prior gains, and they may become risk seeking if they see an opportunity to compensate prior losses. Tversky and Kahneman [194] propose to describe such a behavioral pattern by a function u of the form

$$u(x) = \begin{cases} (x-c)^\gamma & \text{for } x \geq c, \\ -\lambda(c-x)^\gamma & \text{for } x < c, \end{cases}$$

where c is a given benchmark level, and their experiments suggest parameter values λ around 2 and γ slightly less than 1. Nevertheless, one can insist on risk aversion from a *normative* point of view, and this is the approach we will take for the purposes in this book. ◇

Definition 2.37. A function $u : S \to \mathbb{R}$ is called a *utility function* if it is strictly concave, strictly increasing, and continuous on S.

Any increasing concave function $u : S \to \mathbb{R}$ is necessarily continuous on every interval $(a, b] \subset S$; see Proposition A.4. Hence, the condition of continuity in the preceding definition is only relevant if S contains its lower boundary point. Note that any utility function $u(x)$ decreases at least linearly as $x \downarrow \inf S$. Therefore, u cannot be bounded from below unless $\inf S > -\infty$.

From now on, we will consider a fixed preference relation $\succ$ on $\mathcal{M}$ which admits an *expected utility representation*, that is, a von Neumann–Morgenstern representation

$$U(\mu) = \int u \, d\mu$$

in terms of a utility function $u : S \to \mathbb{R}$. The intermediate value theorem applied to the strictly increasing continuous function u yields for any $\mu \in \mathcal{M}$ a unique real number $c(\mu)$ for which

$$u\big(c(\mu)\big) = U(\mu) = \int u \, d\mu. \tag{2.9}$$

It follows that

$$\delta_{c(\mu)} \sim \mu,$$

i.e., there is indifference between the lottery μ and the sure amount of money $c(\mu)$.

Definition 2.38. The *certainty equivalent* of the lottery $\mu \in \mathcal{M}$ is defined as the number $c(\mu)$ of (2.9), and

$$\rho(\mu) := m(\mu) - c(\mu)$$

is called the *risk premium* of μ.

Risk aversion implies via Jensen's inequality that $c(\mu) \leq m(\mu)$, and

$$c(\mu) < m(\mu) \quad \iff \quad \mu \neq \delta_{m(\mu)}.$$

In particular, the risk premium $\rho(\mu)$ is strictly positive as soon as the distribution μ carries any risk.

Remark 2.39. The certainty equivalent $c(\mu)$ can be viewed as an upper bound for any price of μ which would be acceptable to an economic agent with utility function u. Thus, the fair price $m(\mu)$ must be reduced at least by the risk premium $\rho(\mu)$ if one wants the agent to buy the asset distribution μ. Alternatively, suppose that the agent holds an asset with distribution μ. Then the risk premium may be viewed as the amount that the agent would be ready to pay for replacing the asset by its expected value $m(\mu)$. ◇

Example 2.40 ("St. Petersburg Paradox"). Consider the lottery

$$\mu = \sum_{n=1}^{\infty} 2^{-n} \delta_{2^{n-1}}$$

which may be viewed as the payoff distribution of the following game. A fair coin is tossed until a head appears. If the head appears on the n^{th} toss, the payoff will be 2^{n-1} €. Up to the early 18[th] century, it was commonly accepted that the price of a lottery should be computed as the *fair price*, i.e., as the expected value $m(\mu)$.

In the present example, the fair price is given by $m(\mu) = \infty$, but it is hard to find someone who is ready to pay even 20 €. In view of this "paradox", posed by Nicholas Bernoulli in 1713, Gabriel Cramer and Daniel Bernoulli [22] independently introduced the idea of determining the actual price as a certainty equivalent with respect to a utility function. For the two utility functions

$$u_1(x) = \sqrt{x} \quad \text{and} \quad u_2(x) = \log x$$

proposed, respectively, by G. Cramer and by D. Bernoulli, these certainty equivalents are given by

$$c_1(\mu) = \big(2 - \sqrt{2}\big)^{-2} \approx 2.91 \quad \text{and} \quad c_2(\mu) = 2,$$

and this is within the range of prices people are usually ready to pay. Note, however, that for any utility function which is unbounded from above we could modify the payoff in such a way that the paradox reappears. For example, we could replace the payoff 2^n by $u^{-1}(2^n)$ for $n \geq 1000$, so that $\int u \, d\mu = +\infty$. The choice of a bounded utility function would remove this difficulty, but would create others; see the discussion on pp. 69–72. ◇

Given the preference order $\succ$ on $\mathcal{M}$, we can now try to determine those distributions in $\mathcal{M}$ which are maximal with respect to $\succ$. As a first illustration, consider the following simple *optimization problem*. Let X be an integrable random variable on some probability space $(\Omega, \mathcal{F}, P)$ with non-degenerate distribution μ. We assume that X is bounded from below by some number a in the interior of S. Which is the best mix

$$X_\lambda := (1 - \lambda)X + \lambda c$$

of the risky payoff X and the certain amount c, that also belongs to the interior of S? If we evaluate X_λ by its expected utility $E[\,u(X_\lambda)\,]$ and denote by μ_λ the distribution of X_λ under P, then we are looking for a maximum of the function f on $[0, 1]$ defined by

$$f(\lambda) := U(\mu_\lambda) = \int u \, d\mu_\lambda.$$

Since f is strictly concave, it attains its maximum in a unique point $\lambda^* \in [0, 1]$.

Proposition 2.41. (a) *We have $\lambda^* = 1$ if $E[\,X\,] \leq c$, and $\lambda^* > 0$ if $c \geq c(\mu)$.*

(b) *If u is differentiable, then*

$$\lambda^* = 1 \quad \Longleftrightarrow \quad E[\,X\,] \leq c$$

and

$$\lambda^* = 0 \quad \Longleftrightarrow \quad c \leq \frac{E[\,Xu'(X)\,]}{E[\,u'(X)\,]}.$$

Proof. (a): Jensen's inequality yields that

$$f(\lambda) \le u\big(E[\, X_\lambda \,]\big) = u\big((1-\lambda)E[\, X \,] + \lambda c\big),$$

with equality if and only if $\lambda = 1$. It follows that $\lambda^* = 1$ if the right-hand side is increasing in λ, i.e., if $E[\, X \,] \le c$.

Strict concavity of u implies

$$\begin{aligned} f(\lambda) &\ge E\big[\, (1-\lambda)u(X) + \lambda u(c) \,\big] \\ &= (1-\lambda)u\big(c(\mu)\big) + \lambda u(c), \end{aligned}$$

with equality if and only if $\lambda \in \{0, 1\}$. The right-hand side is increasing in λ if $c \ge c(\mu)$, and this implies $\lambda^* > 0$.

(b): Clearly, we have $\lambda^* = 0$ if and only if the right-hand derivative f'_+ of f satisfies $f'_+(0) \le 0$; see Appendix A.1 for the definition of f'_+ and f'_-. Note that the difference quotients

$$\frac{u(X_\lambda) - u(X)}{\lambda} = \frac{u(X_\lambda) - u(X)}{X_\lambda - X} \cdot (c - X)$$

are P-a.s. bounded by

$$u'_+(a \wedge c)|c - X| \in \mathcal{L}^1(P)$$

and that they converge to

$$u'_+(X)(c - X)^+ - u'_-(X)(c - X)^-$$

as $\lambda \downarrow 0$. By Lebesgue's theorem, this implies

$$f'_+(0) = E[\, u'_+(X)(c - X)^+ \,] - E[\, u'_-(X)(c - X)^- \,].$$

If u is differentiable, or if the countable set $\{\, x \mid u'_+(x) \neq u'_-(x) \,\}$ has μ-measure 0, then we can conclude

$$f'_+(0) = E[\, u'(X)(c - X) \,],$$

i.e., $f'_+(0) \le 0$ if and only if

$$c \le \frac{E[\, X u'(X) \,]}{E[\, u'(X) \,]}.$$

In the same way, we obtain

$$f'_-(1) = u'_-(c)E[\, (X - c)^- \,] - u'_+(c)E[\, (X - c)^+ \,].$$

If u is differentiable at c, then we can conclude

$$f'_-(1) = u'(c)\big(c - E[\, X \,]\big).$$

This implies $f'_-(1) < 0$, and hence $\lambda^* < 1$, if and only if $E[\, X \,] > c$. □

Remark 2.42. Note that for a differentiable utility function u we have

$$m(\mu) > c(\mu) \geq \frac{E[\,u'(X)X\,]}{E[\,u'(X)\,]}. \tag{2.10}$$

Indeed, concavity of $u \in C^1(\mathbb{R})$ implies

$$E[\,u(X)\,] + E[\,u'(X)(c - X)\,] \geq u(c),$$

hence $E[\,u'(X)(c - X)\,] \geq 0$ for $c = c(\mu)$. ◇

Example 2.43 (Demand for a risky asset). Let $S = S^1$ be a risky asset with price $\pi = \pi^1$. Given an initial wealth w, an agent with utility function $u \in C^1$ can invest a fraction $(1 - \lambda)w$ into the asset and the remaining part λw into a risk-free bond with interest rate r. The resulting payoff is

$$X_\lambda = \frac{(1-\lambda)w}{\pi}(S - \pi) + \lambda w \cdot r.$$

The preceding proposition implies that there will be no investment into the risky asset if and only if

$$E\left[\frac{S}{1+r}\right] \leq \pi.$$

In other words, the price of the risky asset must be below its expected discounted payoff in order to attract any risk averse investor, and in that case it will indeed be optimal for the investor to invest at least some amount. Instead of the simple linear profiles X_λ, the investor may wish to consider alternative forms of investment. For example, this may involve derivatives such as $\max(S, K) = K + (S - K)^+$ for some threshold K. In order to discuss such non-linear payoff profiles, we need an extended formulation of the optimization problem; see Section 3.3 below. ◇

Example 2.44 (Demand for insurance). Suppose an agent with utility function $u \in C^1$ considers taking at least some partial insurance against a random loss Y, with $0 \leq Y \leq w$ and $P[\,Y \neq E[\,Y\,]\,] > 0$, where w is a given initial wealth. If insurance of λY is available at the insurance premium $\lambda\pi$, the resulting final payoff is given by

$$X_\lambda := w - Y + \lambda(Y - \pi) = (1-\lambda)(w - Y) + \lambda(w - \pi).$$

By Proposition 2.41, full insurance is optimal if and only if $\pi \leq E[\,Y\,]$. In reality, however, the insurance premium π will exceed the "fair premium" $E[\,Y\,]$. In this case, it will be optimal to insure only a fraction $\lambda^* Y$ of the loss, with $\lambda^* \in [0, 1)$. This fraction will be strictly positive as long as

$$\pi < \frac{E[\,Yu'(w - Y)\,]}{E[\,u'(w - Y)\,]} = w - \frac{E[\,(w - Y)u'(w - Y)\,]}{E[\,u'(w - Y)\,]}.$$

Since the right-hand side is strictly larger than $E[\,Y\,]$ due to (2.10), risk aversion may create a demand for insurance even if the insurance premium π lies above the "fair" price $E[\,Y\,]$. As in the previous example, the agent may wish to consider alternative forms of insurance such as a stop-loss contract whose payoff has the non-linear structure $(Y-K)^+$ of a call option. ◇

Let us take another look at the risk premium $\rho(\mu)$ of a lottery μ. For an approximate calculation, we consider the Taylor expansion of a sufficiently smooth utility function $u(x)$ at $x=c(\mu)$ around $m:=m(\mu)$, and we assume that μ has finite variance $\mathrm{var}(\mu)$. On the one hand,

$$u\big(c(\mu)\big)\approx u(m)+u'(m)(c(\mu)-m)=u(m)-u'(m)\rho(\mu).$$

On the other hand,

$$\begin{aligned}u\big(c(\mu)\big)&=\int u(x)\,\mu(dx)\\&=\int\Big[u(m)+u'(m)(x-m)+\frac{1}{2}u''(m)(x-m)^2+r(x)\Big]\,\mu(dx)\\&\approx u(m)+\frac{1}{2}u''(m)\,\mathrm{var}(\mu),\end{aligned}$$

where $r(x)$ denotes the remainder term in the Taylor expansion of u. It follows that

$$\rho(\mu)\approx-\frac{u''(m)}{2\cdot u'(m)}\,\mathrm{var}(\mu)=:\frac{1}{2}\,\alpha(m)\,\mathrm{var}(\mu). \tag{2.11}$$

Thus, $\alpha(m(\mu))$ is the factor by which an economic agent with utility function u weighs the risk, measured by $\frac{1}{2}\,\mathrm{var}(\mu)$, in order to determine the risk premium he or she is ready to pay.

Definition 2.45. Suppose that u is a twice continuously differentiable utility function on S. Then

$$\alpha(x):=-\frac{u''(x)}{u'(x)}$$

is called the *Arrow–Pratt coefficient of absolute risk aversion* of u at level x.

Example 2.46. The following classes of utility functions u and their corresponding coefficients of risk aversion are standard examples.

(a) **Constant absolute risk aversion (CARA):** $\alpha(x)$ equals some constant $\alpha>0$. Since $\alpha(x)=-(\log u')'(x)$, it follows that $u(x)=a-b\cdot e^{-\alpha x}$. Using an affine transformation, u can be normalized to

$$u(x)=1-e^{-\alpha x}.$$

(b) **Hyperbolic absolute risk aversion (HARA):** $\alpha(x) = (1-\gamma)/x$ on $S = (0, \infty)$ for some $\gamma < 1$. Up to affine transformations, we have

$$u(x) = \log x \quad \text{for } \gamma = 0,$$
$$u(x) = \frac{1}{\gamma} x^{\gamma} \quad \text{for } \gamma \neq 0.$$

Sometimes, these functions are also called CRRA utility functions, because their "relative risk aversion" $x\alpha(x)$ is constant. Of course, these utility functions can be shifted to any interval $S = (a, \infty)$. The "risk-neutral" case $\gamma = 1$ would correspond to an affine utility function u. ◇

Proposition 2.47. *Suppose that u and $\widetilde{u}$ are two utility functions on S which are twice continuously differentiable, and that α and $\widetilde{\alpha}$ are the corresponding Arrow–Pratt coefficients of absolute risk aversion. Then the following conditions are equivalent.*

(a) $\alpha(x) \geq \widetilde{\alpha}(x)$ *for all* $x \in S$.

(b) $u = F \circ \widetilde{u}$ *for a strictly increasing concave function* F.

(c) *The respective risk premiums ρ and $\widetilde{\rho}$ associated with u and $\widetilde{u}$ satisfy $\rho(\mu) \geq \widetilde{\rho}(\mu)$ for all $\mu \in \mathcal{M}$.*

Proof. (a)⇒(b): Since $\widetilde{u}$ is strictly increasing, we may define its inverse function, w. Then $F(t) := u\big(w(t)\big)$ is clearly increasing, twice differentiable, and satisfies $u = F \circ \widetilde{u}$. For showing that F is concave and strictly increasing we calculate the first two derivatives of w:

$$w' = \frac{1}{\widetilde{u}'(w)}, \qquad w'' = \widetilde{\alpha}(w) \cdot \frac{1}{\widetilde{u}'(w)^2}.$$

Now we can calculate the first two derivatives of F:

$$F' = u'(w) \cdot w' = \frac{u'(w)}{\widetilde{u}'(w)} > 0$$

and

$$\begin{aligned} F'' &= u''(w)(w')^2 + u'(w)w'' \\ &= \frac{u'(w)}{\widetilde{u}'(w)^2}\big[\,\widetilde{\alpha}(w) - \alpha(w)\,\big] \\ &\leq 0. \end{aligned} \tag{2.12}$$

This proves that F is concave and strictly increasing.

(b)⇒(c): Jensen's inequality implies that the respective certainty equivalents $c(\mu)$ and $\widetilde{c}(\mu)$ satisfy

$$\begin{aligned} u(c(\mu)) &= \int u \, d\mu = \int F \circ \widetilde{u} \, d\mu \\ &\leq F\Big(\int \widetilde{u} \, d\mu\Big) = F\big(\widetilde{u}\big(\widetilde{c}(\mu)\big)\big) = u(\widetilde{c}(\mu)). \end{aligned} \tag{2.13}$$

Hence, $\rho(\mu) = m(\mu) - c(\mu) \geq m(\mu) - \widetilde{c}(\mu) = \widetilde{\rho}(\mu)$.

(c) ⇒(a): If condition (a) is false, there exists an open interval $O \subset S$ such that $\widetilde{\alpha}(x) > \alpha(x)$ for all $x \in O$. Let $\widetilde{O} := \widetilde{u}(O)$, and denote again by w the inverse of $\widetilde{u}$. Then the function $F(t) = u\big(w(t)\big)$ will be strictly convex in the open interval $\widetilde{O}$ by (2.12). Thus, if μ is a measure with support in O, the inequality in (2.13) is reversed and is even strict – unless μ is concentrated at a single point. It follows that $\rho(\mu) < \widetilde{\rho}(\mu)$, which contradicts condition (c). □

In view of the underlying axioms, the paradigm of expected utility has a certain plausibility on a normative level, i.e., as a guideline of rational behavior in the face of risk. But this guideline should be applied with care: If pushed too far, it may lead to unplausible conclusions. In the remaining part of this section we discuss some of these issues. From now on, we assume that S is unbounded from above, so that $w + x \in S$ for any $x \in S$ and $w \geq 0$. So far, we have implicitly assumed that the preference relation $\succ$ on lotteries reflects the views of an economic agent in a given set of conditions, including a fixed level $w \geq 0$ of the agent's initial wealth. In particular, the utility function may vary as the level of wealth changes, and so it should really be indexed by w. Usually one assumes that u_w is obtained by simply shifting a fixed utility function u to the level w, i.e., $u_w(x) := u(w + x)$. Thus, a lottery μ is declined at a given level of wealth w if and only if

$$\int u(w + x)\,\mu(dx) < u(w).$$

Let us now return to the situation of Proposition 2.41 when μ is the distribution of an integrable random variable X on $(\Omega, \mathcal{F}, P)$, which is bounded from below by some number a in the interior of S. We view X as the net payoff of some financial bet, and we assume that the bet is *favorable* in the sense that

$$m(\mu) = E[\,X\,] > 0.$$

Remark 2.48. Even though the favorable bet X might be declined at a given level w due to risk aversion, it follows from Proposition 2.41 that it would be optimal to accept the bet at some smaller scale, i.e., there is some $\gamma^* > 0$ such that

$$E[\,u(w + \gamma^* X)\,] > u(w).$$

On the other hand, it follows from Lemma 2.50 below that the given bet X becomes acceptable at a sufficiently high level of wealth whenever the utility function is unbounded from above. ◇

Sometimes it is assumed that some favorable bet is declined at *every* level of wealth. The assumption that such a bet exists is not as innocent as it may look. In fact it has rather drastic consequences. In particular, we are going to see that it rules out all utility functions in Example 2.46 except for the class of exponential utilities.

Example 2.49. For any exponential utility function $u(x) = 1 - e^{-\alpha x}$ with constant risk aversion $\alpha > 0$, the induced preference order on lotteries does not at all depend on the initial wealth w. To see this, note that

$$\int u(w+x)\,\mu(dx) < \int u(w+x)\,\nu(dx)$$

is equivalent to

$$\int e^{-\alpha x}\,\mu(dx) > \int e^{-\alpha x}\,\nu(dx). \qquad \diamond$$

Let us now show that the rejection of some favorable bet μ at every wealth level w leads to a not quite plausible conclusion: At high levels of wealth, the agent would reject a bet ν with huge potential gain even though the potential loss is just a negligible fraction of the initial wealth.

Lemma 2.50. *If the favorable bet μ is rejected at any level of wealth, then the utility function u is bounded from above, and there exists $A > 0$ such that the bet*

$$\nu := \frac{1}{2}(\delta_{-A} + \delta_{\infty})$$

is rejected at any level of wealth.

Proof. We have assumed that X is bounded from below, i.e., μ is concentrated on $[a, \infty)$ for some $a < 0$, where a is in the interior of S. Moreover, we can choose $b > 0$ such that

$$\widetilde{\mu}(B) := \mu\big(B \cap [a,b]\big) + \delta_b(B)\cdot\mu\big((b,\infty)\big)$$

is still favorable. Since u is increasing, we have

$$\int u(w+x)\,\widetilde{\mu}(dx) \le \int u(w+x)\,\mu(dx) < u(w)$$

for any $w \ge 0$, i.e., also the lottery $\widetilde{\mu}$ is rejected at any level of wealth. It follows that

$$\int_{[0,b]} \big[\,u(w+x) - u(w)\,\big]\,\widetilde{\mu}(dx) < \int_{[a,0)} \big[\,u(w) - u(w+x)\,\big]\,\widetilde{\mu}(dx).$$

Let us assume for simplicity that u is differentiable; the general case requires only minor modifications. Then the previous inequality implies

$$u'(w+b)\,m^{+}(\widetilde{\mu}) < u'(w+a)\,m^{-}(\widetilde{\mu}),$$

where

$$m^{+}(\widetilde{\mu}) := \int_{[0,b]} x\,\widetilde{\mu}(dx) > \int_{[a,0]} (-x)\,\widetilde{\mu}(dx) =: m^{-}(\widetilde{\mu}),$$

due to the fact that $\widetilde{\mu}$ is favorable. Thus,

$$\frac{u'(w+b)}{u'(w-|a|)} < \frac{m^-(\widetilde{\mu})}{m^+(\widetilde{\mu})} =: \gamma < 1$$

for any w, hence

$$u'\big(x + n(|a| + b)\big) < \gamma^n u'(x)$$

for any x in the interior of S. This exponential decay of the derivative implies $u(\infty) := \lim_{x\uparrow\infty} u(x) < \infty$. More precisely, if $A := n(|a| + b)$ for some n, then

$$\begin{aligned} u(\infty) - u(x) &= \sum_{k=0}^{\infty} \int_{x+kA}^{x+(k+1)A} u'(y)\, dy \\ &= \sum_{k=0}^{\infty} \int_{x-A}^{x} u'(z + (k+1)A)\, dz \\ &< \sum_{k=0}^{\infty} \gamma^{(k+1)n} \int_{x-A}^{x} u'(z)\, dz \\ &= \frac{\gamma^n}{1-\gamma^n}\big(u(x) - u(x-A)\big). \end{aligned}$$

Take n such that $\gamma^n \le 1/2$. Then we obtain

$$u(\infty) - u(x) < u(x) - u(x-A),$$

i.e.,

$$\frac{1}{2}\big(u(\infty) + u(x-A)\big) < u(x)$$

for all x such that $x - A \in S$. □

Example 2.51. For an exponential utility function $u(x) = 1 - e^{-\alpha x}$, the bet ν defined in the preceding lemma is rejected at any level of wealth as soon as $A > \frac{1}{\alpha}\log 2$. ◇

Suppose now that the lottery $\mu \in \mathcal{M}$ is played not only once but n times in a row. For instance, one can think of an insurance company selling identical policies to a large number of individual customers. More precisely, let $(\Omega, \mathcal{F}, P)$ be a probability space supporting a sequence $X_1, X_2, \ldots$ of independent random variables with common distribution μ. The value of X_i will be interpreted as the outcome of the i^{th} drawing of the lottery μ. The accumulated payoff of n successive independent repetitions of the financial bet X_1 is given by

$$Z_n := \sum_{i=1}^{n} X_i,$$

and we assume that this accumulated payoff takes values in S; this is the case if, e.g., $S = [0, \infty)$.

Remark 2.52. It may happen that an agent refuses the single favorable bet X at any level of wealth but feels tempted by a sufficiently large series $X_1, \dots, X_n$ of independent repetitions of the same bet. It is true that, by the weak law of large numbers, the probability

$$P[\, Z_n < 0\,] = P\Big[\, \frac{1}{n}\sum_{i=1}^{n} X_i < m(\mu) - \varepsilon \,\Big]$$

(for $\varepsilon := m(\mu)$) of incurring a cumulative loss at the end of the series converges to 0 as $n \uparrow \infty$. Nevertheless, the decision of accepting n repetitions is not consistent with the decision to reject the single bet at any wealth level w. In fact, for $W_k := w + Z_k$ we obtain

$$\begin{aligned} E[\, u(W_n)\,] &= E\big[\, E[\, u(W_{n-1} + X_n) \mid X_1, \dots, X_{n-1}\,]\big] \\ &= E\Big[\, \int u(W_{n-1} + x)\, \mu(dx) \Big] \\ &< E[\, u(W_{n-1})\,] < \cdots < u(w), \end{aligned}$$

i.e., the bet described by Z_n should be rejected as well. ◇

Let us denote by μ_n the distribution of the accumulated payoff Z_n. The lottery μ_n has the mean $m(\mu_n) = n \cdot m(\mu)$, the certainty equivalent $c(\mu_n)$, and the associated risk premium $\rho(\mu_n) = n \cdot m(\mu) - c(\mu_n)$. We are interested in the asymptotic behavior of these quantities for large n. Kolmogorov's law of large numbers states that the average outcome $\frac{1}{n} Z_n$ converges P-a.s. to the constant $m(\mu)$. Therefore, one might guess that a similar averaging effect occurs on the level of the relative certainty equivalents

$$c_n := \frac{c(\mu_n)}{n}$$

and of the relative risk premiums

$$\rho_n := \frac{\rho(\mu_n)}{n} = m(\mu) - c_n.$$

Does c_n converge to $m(\mu)$, and is there a successive reduction of the relative risk premiums ρ_n as n grows to infinity? Applying our heuristic (2.11) to the present situation yields

$$\rho_n \approx \frac{1}{2n}\, \alpha\big(m(\mu_n)\big)\, \mathrm{var}(\mu_n) = \frac{1}{2}\, \alpha\big(n \cdot m(\mu)\big)\, \mathrm{var}(\mu).$$

Thus, one should expect that ρ_n tends to zero only if the Arrow–Pratt coefficient $\alpha(x)$ becomes arbitrarily small as x becomes large, i.e., if the utility function is *decreasingly risk averse*. This guess is confirmed by the following two examples.

Example 2.53. Suppose that $u(x) = 1 - e^{-\alpha x}$ is a CARA utility function with constant risk aversion $\alpha > 0$ and assume that μ is such that $\int e^{-\alpha x}\,\mu(dx) < \infty$. Then, with the notation introduced above,

$$\int e^{-\alpha x}\,\mu_n(dx) = E\Big[\prod_{i=1}^{n} e^{-\alpha X_i}\Big] = \Big(\int e^{-\alpha x}\,\mu(dx)\Big)^n.$$

Hence, the certainty equivalent of μ_n is given by

$$c(\mu_n) = -\frac{n}{\alpha}\log\int e^{-\alpha x}\,\mu(dx) = n\cdot c(\mu).$$

It follows that c_n and ρ_n are independent of n. In particular, the relative risk premiums are not reduced if the lottery is drawn more than once. ◇

The second example displays a different behavior. It shows that for HARA utility functions the relative risk premiums will indeed decrease to 0. In particular, the lottery μ_n will become attractive for large enough n as soon as the price of the single lottery μ is less than $m(\mu)$.

Example 2.54. Suppose that μ is a non-degenerate lottery concentrated on $(0, \infty)$, and that u is a HARA utility function of index $\gamma \in [0, 1)$. If $\gamma > 0$ then $u(x) = \frac{1}{\gamma}x^\gamma$ and $c(\mu_n) = E[(Z_n)^\gamma]^{1/\gamma}$, hence

$$c_n = \frac{c(\mu_n)}{n} = E\bigg[\Big(\frac{1}{n}Z_n\Big)^\gamma\bigg]^{1/\gamma} < m(\mu).$$

If $\gamma = 0$ then $u(x) = \log x$, and the relative certainty equivalent satisfies

$$\log c_n = \log c(\mu_n) - \log n = E\bigg[\log\Big(\frac{1}{n}Z_n\Big)\bigg].$$

Thus, we have

$$u(c_n) = E\bigg[u\Big(\frac{1}{n}Z_n\Big)\bigg]$$

for any $\gamma \in [0, 1)$. By symmetry,

$$\frac{1}{n+1}Z_{n+1} = E[\,X_k \mid Z_{n+1}\,] \quad \text{for } k = 1, \dots, n+1;$$

see part II of §20 in [19]. It follows that

$$\frac{1}{n+1}Z_{n+1} = E\bigg[\frac{1}{n}Z_n \;\bigg|\; Z_{n+1}\bigg]. \tag{2.14}$$

Since u is strictly concave and since μ is non-degenerate, we get

$$\begin{aligned} u(c_{n+1}) &= E\Big[\, u\Big(E\Big[\frac{1}{n} Z_n \;\Big|\; Z_{n+1} \Big]\Big) \Big] \\ &> E\Big[\, E\Big[u\Big(\frac{1}{n} Z_n \Big) \;\Big|\; Z_{n+1} \Big] \Big] \\ &= u(c_n), \end{aligned}$$

i.e., the relative certainty equivalents are *strictly increasing* and the relative risk premiums ρ_n are *strictly decreasing*. By Kolmogorov's law of large numbers,

$$\frac{1}{n} Z_n \longrightarrow m(\mu) \quad P\text{-a.s.} \tag{2.15}$$

Thus, by Fatou's lemma (we assume for simplicity that μ is concentrated on $[\varepsilon, \infty)$ for some $\varepsilon > 0$ if $\gamma = 0$),

$$\liminf_{n\uparrow\infty} u(c_n) \ge E\Big[\liminf_{n\uparrow\infty} u\Big(\frac{1}{n} Z_n \Big) \Big] = u\big(m(\mu)\big),$$

hence

$$\lim_{n\uparrow\infty} c_n = m(\mu) \quad \text{and} \quad \lim_{n\uparrow\infty} \rho_n = 0.$$

Suppose that the price of μ is given by $\pi \in \big(c(\mu), m(\mu)\big)$. At initial wealth $w = 0$, the agent would decline a single bet. But, in contrast to the situation in Remark 2.52, a series of n repetitions of the same bet would now become attractive for large enough n, since $c(\mu_n) = nc_n > n\pi$ for

$$n \ge n_0 := \min\{ k \in \mathbb{N} \mid c_k > \pi \} < \infty. \qquad \diamond$$

Remark 2.55. The identity (2.14) can also be written as

$$\frac{1}{n+1} Z_{n+1} = E\Big[\frac{1}{n} Z_n \;\Big|\; \mathcal{A}_{n+1} \Big] = E[\, X_1 \mid \mathcal{A}_{n+1} \,]$$

where $\mathcal{A}_{n+1} = \sigma(Z_{n+1}, Z_{n+2}, \dots)$. This means that the stochastic process $\frac{1}{n} Z_n$, $n = 1, 2 \dots$, is a *reverse martingale*. In particular, Kolmogorov's law of large numbers (2.15) can be regarded as a special case of the convergence theorem for reverse martingales; see part II of §20 in [19]. $\diamond$

2.4 Uniform preferences

So far, we have considered preference relations on distributions defined in terms of a fixed utility function u. In this section, we focus on the question whether one distribution is preferred over another, regardless of the choice of a particular utility function.

For simplicity, we take $S = \mathbb{R}$ as the set of possible payoffs. Let $\mathcal{M}$ be the set of all $\mu \in \mathcal{M}_1(\mathbb{R})$ with well-defined and finite expectation

$$m(\mu) = \int x\, \mu(dx).$$

Recall from Definition 2.37 that a *utility function* on $\mathbb{R}$ is a strictly concave and strictly increasing function $u : \mathbb{R} \to \mathbb{R}$. Since each concave function u is dominated by an affine function, the existence of $m(\mu)$ implies the existence of the integral $\int u\, d\mu$ as an extended real number in $[-\infty, \infty)$.

Definition 2.56. Let ν and μ be lotteries in $\mathcal{M}$. We say that the lottery μ is *uniformly preferred* over ν and we write

$$\mu \succcurlyeq_{\text{uni}} \nu$$

if

$$\int u\, d\mu \geq \int u\, d\nu \quad \text{for all utility functions } u.$$

Thus, $\mu \succcurlyeq_{\text{uni}} \nu$ holds if and only if every risk-averse agent will prefer μ over ν, regardless of which utility function the agent is actually using. In this sense, $\mu \succcurlyeq_{\text{uni}} \nu$ expresses a uniform preference for μ over ν. Sometimes, $\succcurlyeq_{\text{uni}}$ is also called *second order stochastic dominance*; the notion of *first order stochastic dominance* will be introduced in Definition 2.69.

Remark 2.57. The binary relation $\succcurlyeq_{\text{uni}}$ is a *partial order* on $\mathcal{M}$, i.e., $\succcurlyeq_{\text{uni}}$ satisfies the following three properties:

- *Reflexivity:* $\mu \succcurlyeq_{\text{uni}} \mu$ for all $\mu \in \mathcal{M}$.
- *Transitivity:* $\mu \succcurlyeq_{\text{uni}} \nu$ and $\nu \succcurlyeq_{\text{uni}} \lambda$ imply $\mu \succcurlyeq_{\text{uni}} \lambda$.
- *Antisymmetry:* $\mu \succcurlyeq_{\text{uni}} \nu$ and $\nu \succcurlyeq_{\text{uni}} \mu$ imply $\mu = \nu$.

The first two properties are obvious, the third is derived in Remark 2.59. Moreover, $\succcurlyeq_{\text{uni}}$ is *monotone* and *risk-averse* in the sense that

$$\delta_y \succcurlyeq_{\text{uni}} \delta_x \text{ for } y \geq x, \quad \text{and} \quad \delta_{m(\mu)} \succcurlyeq_{\text{uni}} \mu \text{ for all } \mu \in \mathcal{M}.$$

Note, however, that $\succcurlyeq_{\text{uni}}$ is *not* a weak preference relation in the sense of Definition 2.2, since it is not complete, see Remark 2.3. ◇

In the following theorem, we will give a number of equivalent formulations of the statement $\mu \succcurlyeq_{\text{uni}} \nu$. One of them needs the notion of a *stochastic kernel* on $\mathbb{R}$. This is a mapping

$$Q : \mathbb{R} \longrightarrow \mathcal{M}_1(\mathbb{R})$$

such that $x \mapsto Q(x, A)$ is measurable for each fixed Borel set $A \subset \mathbb{R}$. See Appendix A.3 for the notion of a quantile function, which will be used in condition (e).

Theorem 2.58. *For any pair $\mu, \nu \in \mathcal{M}$ the following conditions are equivalent.*

(a) $\mu \succcurlyeq_{\text{uni}} \nu$.

(b) $\int f\, d\mu \geq \int f\, d\nu$ *for all increasing concave functions* f.

(c) *For all* $c \in \mathbb{R}$

$$\int (c-x)^+ \,\mu(dx) \leq \int (c-x)^+ \,\nu(dx).$$

(d) *If* F_μ *and* F_ν *denote the distribution functions of* μ *and* ν*, then*

$$\int_{-\infty}^{c} F_\mu(x)\, dx \leq \int_{-\infty}^{c} F_\nu(x)\, dx \quad \textit{for all } c \in \mathbb{R}.$$

(e) *If* q_μ *and* q_ν *are quantile functions for* μ *and* ν*, then*

$$\int_0^t q_\mu(s)\, ds \geq \int_0^t q_\nu(s)\, ds \quad \textit{for } 0 < t \leq 1.$$

(f) *There exists a probability space* $(\Omega, \mathcal{F}, P)$ *with random variables* X_μ *and* X_ν *having respective distributions* μ *and* ν *such that*

$$E[\, X_\nu \mid X_\mu \,] \leq X_\mu \quad P\text{-a.s.}$$

(g) *There exists a stochastic kernel* $Q(x, dy)$ *on* $\mathbb{R}$ *such that* $Q(x,\cdot) \in \mathcal{M}$ *and* $m(Q(x,\cdot)) \leq x$ *for all* x *and such that* $\nu = \mu Q$*, where* μQ *denotes the measure*

$$\mu Q(A) := \int Q(x, A)\, \mu(dx) \quad \textit{for Borel sets } A \subset \mathbb{R}.$$

Below we will show the following implications between the conditions of the theorem:

$$\text{(e)} \iff \text{(d)} \iff \text{(c)} \iff \text{(b)} \iff \text{(a)} \Longleftarrow \text{(g)} \Longleftarrow \text{(f)}. \tag{2.16}$$

The difficult part is the proof that (b) implies (f). It will be deferred to Section 2.6, where we will prove a multidimensional variant of this result; cf. Theorem 2.93.

Proof of (2.16). (e)⇔(d): This follows from Lemma A.22.

(d)⇔(c): By Fubini's theorem,

$$\begin{aligned}
\int_{-\infty}^{c} F_\mu(y)\, dy &= \int_{-\infty}^{c} \int_{(-\infty,y]} \mu(dz)\, dy \\
&= \int \int \mathrm{I}_{\{z \leq y \leq c\}}\, dy\, \mu(dz) \\
&= \int (c-z)^+ \,\mu(dz).
\end{aligned}$$

(c)⇔(b): Condition (b) implies (c) because $f(x) := -(c-x)^+$ is concave and increasing. In order to prove the converse assertion, we take an increasing concave function f and let $h := -f$. Then h is convex and decreasing, and its increasing right-hand derivative $h' := h'_+$ can be regarded as a "distribution function" of a non-negative Radon measure γ on $\mathbb{R}$,

$$h'(b) = h'(a) + \gamma\big((a,b]\big) \quad \text{for } a < b;$$

see Appendix A.1. As in (1.11):

$$h(x) = h(b) - h'(b)\,(b-x) + \int_{(-\infty,b]} (z-x)^+\,\gamma(dz) \quad \text{for } x < b.$$

Using $h'(b) \le 0$, Fubini's theorem, and condition (c), we obtain that

$$\begin{aligned}\int\limits_{(-\infty,b]} h\,d\mu &= h(b) - h'(b)\int (b-x)^+\,\mu(dx) + \int\limits_{(-\infty,b]}\int (z-x)^+\,\mu(dx)\,\gamma(dz)\\ &\le h(b) - h'(b)\int (b-x)^+\,\nu(dx) + \int\limits_{(-\infty,b]}\int (z-x)^+\,\nu(dx)\,\gamma(dz)\\ &= \int_{(-\infty,b]} h\,d\nu.\end{aligned}$$

Taking $b \uparrow \infty$ yields $\int f\,d\mu \ge \int f\,d\nu$.

(a)⇔(b): That (b) implies (a) is obvious. For the proof of the converse implication, choose any utility function u_0 for which both $\int u_0\,d\mu$ and $\int u_0\,d\nu$ are finite. For instance, one can take

$$u_0(x) := \begin{cases} x - e^{x/2} + 1 & \text{if } x \le 0,\\ \sqrt{x+1} - 1 & \text{if } x \ge 0.\end{cases}$$

Then, for f concave and increasing and for $\alpha \in [0,1)$,

$$u_\alpha(x) := \alpha f(x) + (1-\alpha)u_0(x)$$

is a utility function. Hence,

$$\int f\,d\mu = \lim_{\alpha\uparrow 1}\int u_\alpha\,d\mu \ge \lim_{\alpha\uparrow 1}\int u_\alpha\,d\nu = \int f\,d\nu.$$

(f)⇒(g): By considering the joint distribution of X_μ and X_ν, we may reduce our setting to the situation in which $\Omega = \mathbb{R}^2$ and where X_μ and X_ν are the respective projections on the first and second coordinates, i.e., for $\omega = (x,y) \in \Omega = \mathbb{R}^2$ we have $X_\mu(\omega) = x$ and $X_\nu(\omega) = y$. Let $Q(x,dy)$ be a *regular conditional distribution* of X_ν given X_μ, i.e., a stochastic kernel on $\mathbb{R}$ such that

$$P[\,X_\nu \in A \mid X_\mu\,](\omega) = Q(X_\mu(\omega), A)$$

for all Borel sets $A \subseteq \mathbb{R}$ and for P-a.e. $\omega \in \Omega$ (see, e.g., Theorem 44.3 of [19] for an existence proof). Clearly, $\nu = \mu Q$. Condition (f) implies that

$$X_\mu(\omega) \geq E[\, X_\nu \mid X_\mu \,](\omega) = \int y\, Q(X_\mu(\omega), dy) \quad \text{for } P\text{-a.e. } \omega \in \Omega.$$

Hence, Q satisfies

$$\int y\, Q(x, dy) \leq x \quad \text{for } \mu\text{-a.e. } x.$$

By modifying Q on a μ-null set (e.g., by putting $Q(x,\cdot) := \delta_x$ there), this inequality can be achieved for all $x \in \mathbb{R}$.

(g) ⇒(a): Let u be a utility function. Jensen's inequality applied to the measure $Q(x, dy)$ implies

$$\int u(y)\, Q(x, dy) \leq u\big(m\big(Q(x,\cdot)\big)\big) \leq u(x).$$

Hence,

$$\int u\, d\nu = \int \int u(y)\, Q(x, dy)\, \mu(dx) \leq \int u\, d\mu,$$

completing the proof of the set of implications (2.16). □

Remark 2.59. Let us note some consequences of the preceding theorem. First, taking in condition (b) the increasing concave function $f(x) = x$ yields

$$m(\mu) \geq m(\nu) \ \text{ if } \ \mu \succcurlyeq_{\text{uni}} \nu,$$

i.e., the expectation $m(\cdot)$ is increasing with respect to $\succcurlyeq_{\text{uni}}$.

Next, suppose that μ and ν are such that

$$\int (c-x)^+\, \mu(dx) = \int (c-x)^+\, \nu(dx) \quad \text{for all } c.$$

Then we have both $\mu \succcurlyeq_{\text{uni}} \nu$ and $\nu \succcurlyeq_{\text{uni}} \mu$, and condition (d) of the theorem implies that the respective distribution functions satisfy

$$\int_{-\infty}^{c} F_\mu(x)\, dx = \int_{-\infty}^{c} F_\nu(x)\, dx \quad \text{for all } c.$$

Differentiating with respect to c gives the identity $\mu = \nu$, i.e., a measure $\mu \in \mathcal{M}$ is uniquely determined by the integrals $\int (c-x)^+\, \mu(dx)$ for all $c \in \mathbb{R}$. In particular, $\succcurlyeq_{\text{uni}}$ is antisymmetric. ◇

The following proposition characterizes the partial order $\succcurlyeq_{\text{uni}}$ considered on the set of all normal distributions $N(m, \sigma^2)$. Recall that the standard normal distribution $N(0, 1)$ is defined by its density function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}, \quad x \in \mathbb{R}.$$

The corresponding distribution function is usually denoted

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\,dy, \quad x \in \mathbb{R}.$$

More generally, the normal distribution $N(m, \sigma^2)$ with mean $m \in \mathbb{R}$ and variance $\sigma^2 > 0$ is given by the density function

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R}.$$

Proposition 2.60. *For two normal distributions, we have $N(m, \sigma^2) \succcurlyeq_{\text{uni}} N(\widetilde{m}, \widetilde{\sigma}^2)$ if and only if both $m \geq \widetilde{m}$ and $\sigma^2 \leq \widetilde{\sigma}^2$ hold.*

Proof. In order to prove necessity, note that $N(m, \sigma^2) \succcurlyeq_{\text{uni}} N(\widetilde{m}, \widetilde{\sigma}^2)$ implies that

$$e^{-\alpha m + \alpha^2\sigma^2/2} = \int e^{-\alpha x}\, N(m, \sigma^2)(dx) \leq \int e^{-\alpha x}\, N(\widetilde{m}, \widetilde{\sigma}^2)(dx) = e^{-\alpha\widetilde{m} + \alpha^2\widetilde{\sigma}^2/2}.$$

Hence, for $\alpha > 0$,

$$m - \frac{1}{2}\alpha\sigma^2 \geq \widetilde{m} - \frac{1}{2}\alpha\widetilde{\sigma}^2,$$

which gives $m \geq \widetilde{m}$ by letting $\alpha \downarrow 0$ and $\sigma^2 \leq \widetilde{\sigma}^2$ for $\alpha \uparrow \infty$.

We show sufficiency first in the case $m = \widetilde{m} = 0$. Note that the distribution function of $N(0, \sigma^2)$ is given by $\Phi(x/\sigma)$. Since $\varphi'(x) = -x\varphi(x)$,

$$\frac{d}{d\sigma} \int_{-\infty}^{c} \Phi\Big(\frac{x}{\sigma}\Big)\,dx = \int_{-\infty}^{c} \varphi\Big(\frac{x}{\sigma}\Big) \cdot \frac{-x}{\sigma^2}\,dx = \varphi\Big(\frac{c}{\sigma}\Big) > 0.$$

Note that interchanging differentiation and integration is justified by dominated convergence. Thus, we have shown that $\sigma \mapsto \int_{-\infty}^{c} \Phi(x/\sigma)\,dx$ is strictly increasing for all c, and $N(0, \sigma^2) \succcurlyeq_{\text{uni}} N(0, \widetilde{\sigma}^2)$ follows from part (d) of Theorem 2.58.

Now we turn to the case of arbitrary expectations m and $\widetilde{m}$. Let u be a utility function. Then

$$\int u\,dN(m, \sigma^2) = \int u(m + x)\, N(0, \sigma^2)(dx) \geq \int u(\widetilde{m} + x)\, N(0, \sigma^2)(dx),$$

because $m \geq \widetilde{m}$. Since $x \mapsto u(\widetilde{m} + x)$ is again a utility function, we obtain from the preceding step of the proof that

$$\int u(\widetilde{m} + x)\, N(0, \sigma^2)(dx) \geq \int u(\widetilde{m} + x)\, N(0, \widetilde{\sigma}^2)(dx) = \int u\,dN(\widetilde{m}, \widetilde{\sigma}^2),$$

and $N(m, \sigma^2) \succcurlyeq_{\text{uni}} N(\widetilde{m}, \widetilde{\sigma}^2)$ follows. □

Remark 2.61. Let us indicate an alternative proof for the sufficiency part of Proposition 2.60 that uses condition (g) instead of (d) in Theorem 2.58. To this end, we define a stochastic kernel by $Q(x,\cdot) := N(x+\widetilde{m}-m, \hat{\sigma}^2)$, where $\hat{\sigma}^2 := \widetilde{\sigma}^2 - \sigma^2 > 0$. Then $m(Q(x,\cdot)) = x + \widetilde{m} - m \le x$ and

$$N(m,\sigma^2)\,Q = N(m,\sigma^2) * N(\widetilde{m}-m,\hat{\sigma}^2) = N(m+\widetilde{m}-m,\sigma^2+\hat{\sigma}^2) = N(m,\widetilde{\sigma}^2),$$

where $*$ denotes convolution. Hence, $N(m,\sigma^2) \succcurlyeq_{\text{uni}} N(\widetilde{m},\widetilde{\sigma}^2)$ follows. ◇

The following corollary investigates the relation $\mu \succcurlyeq_{\text{uni}} \nu$ for lotteries with the same expectation. A multidimensional version of this result will be given in Corollary 2.94 below.

Corollary 2.62. *For all $\mu, \nu \in \mathcal{M}$ the following conditions are equivalent.*

(a) $\mu \succcurlyeq_{\text{uni}} \nu$ *and* $m(\mu) = m(\nu)$.

(b) $\int f\,d\mu \ge \int f\,d\nu$ *for all (not necessarily increasing) concave functions* f.

(c) $m(\mu) \ge m(\nu)$ *and* $\int (x-c)^+\,\mu(dx) \le \int (x-c)^+\,\nu(dx)$ *for all* $c \in \mathbb{R}$.

(d) *There exists a probability space* $(\Omega, \mathcal{F}, P)$ *with random variables* X_μ *and* X_ν *having respective distributions* μ *and* ν *such that*

$$E[\,X_\nu \mid X_\mu\,] = X_\mu \quad P\text{-a.s.}$$

(e) *There exists a "mean-preserving spread"* Q, *i.e., a stochastic kernel on* $\mathbb{R}$ *such that* $m(Q(x,\cdot)) = x$ *for all* $x \in S$, *such that* $\nu = \mu Q$.

Proof. (a)⇒(e): Condition (g) of Theorem 2.58 yields a stochastic kernel Q such that $\nu = \mu Q$ and $m\big(Q(x,\cdot)\big) \le x$. Due to the assumption $m(\mu) = m(\nu)$, Q must satisfy $m\big(Q(x,\cdot)\big) = x$ at least for μ-a.e. x. By modifying Q on the μ-null set where $m\big(Q(x,\cdot)\big) < x$ (e.g. by putting $Q(x,\cdot) := \delta_x$ there), we obtain a kernel as needed for condition (e).

(e)⇒(b): Since

$$\int f(y)\,Q(x,dy) \le f\big(m\big(Q(x,\cdot)\big)\big) = f(x)$$

by Jensen's inequality, we obtain

$$\int f\,d\nu = \int\int f(y)\,Q(x,dy)\,\mu(dx) \le \int f\,d\mu.$$

(b)⇒(c): Just take the concave functions $f(x) = -(x-c)^+$, and $f(x) = x$.

(c) ⇒ (a): Note that

$$\int (x-c)^+ \,\mu(dx) = \int_{(c,\infty)} x\,\mu(dx) - c + c\,\mu\big((-\infty, c]\big).$$

The existence of $m(\mu)$ implies that $c\,\mu\big((-\infty, c]\big) \to 0$ as $c \downarrow -\infty$. Hence, we deduce from the second condition in (c) that $m(\mu) \le m(\nu)$, i.e., the two expectations are in fact identical. Now we can apply the following "put-call parity" (compare also (1.10))

$$\int (c-x)^+ \,\mu(dx) = c - m(\mu) + \int (x-c)^+ \,\mu(dx)$$

to see that our condition (c) implies the third condition of Theorem 2.58 and, thus, $\mu \succcurlyeq_{\text{uni}} \nu$.

(d) ⇔ (a): Condition (d) implies both $m(\mu) = m(\nu)$ and condition (f) of Theorem 2.58, and this implies our condition (a). Conversely, assume that (a) holds. Then Theorem 2.58 provides random variables X_μ and X_ν having the respective distributions μ and ν such that $E[\, X_\nu \mid X_\mu \,] \le X_\mu$. Since X_μ and X_ν have the same mean, this inequality must in fact be an almost-sure equality, and we obtain condition (d). □

Let us denote by

$$\operatorname{var}(\mu) := \int \big(x - m(\mu)\big)^2 \,\mu(dx) = \int x^2 \,\mu(dx) - m(\mu)^2 \in [0, \infty]$$

the *variance* of a lottery $\mu \in \mathcal{M}$.

Remark 2.63. If μ and ν are two lotteries in $\mathcal{M}$ such that $m(\mu) = m(\nu)$ and $\mu \succcurlyeq_{\text{uni}} \nu$, then $\operatorname{var}(\mu) \le \operatorname{var}(\nu)$. This follows immediately by taking the concave function $f(x) := -x^2$ in condition (b) of Corollary 2.62. ◇

In the financial context, comparisons of portfolios with known payoff distributions often use a *mean-variance* approach based on the relation

$$\mu \succcurlyeq \nu \quad :\Longleftrightarrow \quad m(\mu) \ge m(\nu) \text{ and } \operatorname{var}(\mu) \le \operatorname{var}(\nu).$$

For normal distributions μ and ν, we have seen that the relation $\mu \succcurlyeq \nu$ is equivalent to $\mu \succcurlyeq_{\text{uni}} \nu$. Beyond this special case, the equivalence typically fails as illustrated by the following example and by Proposition 2.67 below.

Example 2.64. Let μ be the uniform distribution on the interval $[-1, 1]$, so that $m(\mu) = 0$ and $\operatorname{var}(\mu) = 1/3$. For ν we take $\nu = p\delta_{-1/2} + (1-p)\delta_2$. With the choice of $p = 4/5$ we obtain $m(\nu) = 0$ and $1 = \operatorname{var}(\nu) > \operatorname{var}(\mu)$. However,

$$\frac{1}{16} = \int \Big(-\frac{1}{2} - x\Big)^+ \mu(dx) > \int \Big(-\frac{1}{2} - x\Big)^+ \nu(dx) = 0,$$

so $\mu \succcurlyeq_{\text{uni}} \nu$ does not hold. ◇

Remark 2.65. Let μ and ν be two lotteries in $\mathcal{M}$. We will write $\mu \succcurlyeq_{\text{con}} \nu$ if

$$\int f\, d\mu \geq \int f\, d\nu \quad \text{for all concave functions } f \text{ on } \mathbb{R}. \tag{2.17}$$

Note that $\mu \succcurlyeq_{\text{con}} \nu$ implies that $m(\mu) = m(\nu)$, because both $f(x) = x$ and $\widetilde{f}(x) = -x$ are concave. Corollary 2.62 shows that $\succcurlyeq_{\text{con}}$ coincides with our uniform partial order $\succcurlyeq_{\text{uni}}$ if we compare two measures which have the same mean. The partial order $\succcurlyeq_{\text{con}}$ is sometimes called *concave stochastic order*. It was proposed in [170] and [171] to express the view that μ is *less risky* than ν. The inverse relation $\mu \succcurlyeq_{\text{bal}} \nu$ defined by

$$\int f\, d\mu \geq \int f\, d\nu \quad \text{for all convex functions } f \text{ on } \mathbb{R} \tag{2.18}$$

is sometimes called *balayage order* or *convex stochastic order*. ◇

The following class of asset distributions is widely used in Finance.

Definition 2.66. A real-valued random variable Y on some probability space $(\Omega, \mathcal{F}, P)$ is called *log-normally distributed* with parameters $\alpha \in \mathbb{R}$ and $\sigma \geq 0$ if it can be written as

$$Y = \exp(\alpha + \sigma X), \tag{2.19}$$

where X has a standard normal law $N(0, 1)$.

Clearly, any log-normally distributed random variable Y on $(\Omega, \mathcal{F}, P)$ takes P-a.s. strictly positive values. Recall from above the standard notations φ and Φ for the density and the distribution function of the standard normal law $N(0, 1)$. We obtain from (2.19) the distribution function

$$P[\, Y \leq y\,] = \Phi\Big(\frac{\log y - \alpha}{\sigma}\Big), \quad 0 < y < \infty,$$

and the density

$$\psi(y) = \frac{1}{\sigma \cdot y}\, \varphi\Big(\frac{\log y - \alpha}{\sigma}\Big) \cdot I_{(0,\infty)}(y) \tag{2.20}$$

of the log-normally distributed random variable Y. Its p^{th} moment is given by the formula

$$E[\, Y^p\,] = \exp\Big(p\alpha + \frac{1}{2}\, p^2\sigma^2\Big).$$

In particular, the law μ of Y has the expectation

$$m(\mu) = E[\, Y\,] = \exp\Big(\alpha + \frac{1}{2}\, \sigma^2\Big)$$

and the variance

$$\operatorname{var}(\mu) = \exp\big(2\alpha + \sigma^2\big)\big(\exp(\sigma^2) - 1\big).$$

Proposition 2.67. *Let μ and $\widetilde{\mu}$ be two log-normal distributions with parameters (α, σ) and $(\widetilde{\alpha}, \widetilde{\sigma})$, respectively. Then $\mu \succcurlyeq_{\text{uni}} \widetilde{\mu}$ holds if and only if $\sigma^2 \le \widetilde{\sigma}^2$ and $\alpha + \frac{1}{2}\sigma^2 \ge \widetilde{\alpha} + \frac{1}{2}\widetilde{\sigma}^2$.*

Proof. First suppose that $\sigma^2 \le \widetilde{\sigma}^2$ and $m(\mu) \ge m(\widetilde{\mu})$. We define a kernel $Q(x, \cdot)$ as the law of $x \cdot \exp(\lambda + \beta Z)$ where Z is a standard normal random variable. Now suppose that μ is represented by (2.19) with X independent of Z, and let f denote a bounded measurable function. It follows that

$$\int f \, d(\mu Q) = E\big[f\big(e^{\alpha+\sigma X} \cdot e^{\lambda+\beta Z}\big)\big] = E\big[f\big(e^{\alpha+\lambda+(\sigma^2+\beta^2)^{1/2} U}\big)\big],$$

where

$$U = \frac{\sigma X + \beta Z}{\sqrt{\sigma^2 + \beta^2}}$$

is also $N(0, 1)$-distributed. Thus, μQ is a log-normal distribution with parameters $(\alpha + \lambda, \sqrt{\sigma^2 + \beta^2}\,)$. By taking $\beta := \sqrt{\widetilde{\sigma}^2 - \sigma^2}$ and $\lambda := \widetilde{\alpha} - \alpha$, we can represent $\widetilde{\mu}$ as $\widetilde{\mu} = \mu Q$. With this parameter choice,

$$\lambda = \widetilde{\alpha} - \alpha = \log m(\widetilde{\mu}) - \log m(\mu) - \frac{1}{2}(\widetilde{\sigma}^2 - \sigma^2) \le -\frac{\beta^2}{2}.$$

We have thus $m(Q(x, \cdot)) \le x$ for all x, and so $\mu \succcurlyeq_{\text{uni}} \widetilde{\mu}$ follows from condition (g) of Theorem 2.58.

As to the converse implication, the inequality $m(\mu) \ge m(\widetilde{\mu})$ is already clear. To prove $\sigma^2 \le \widetilde{\sigma}^2$, let $\nu := \mu \circ \log^{-1}$ and $\widetilde{\nu} := \widetilde{\mu} \circ \log^{-1}$ so that $\nu = N(\alpha, \sigma^2)$ and $\widetilde{\nu} = N(\widetilde{\alpha}, \widetilde{\sigma}^2)$. For $\varepsilon > 0$ we define the concave increasing function $f_\varepsilon(x) := \log(\varepsilon + x)$. If u is a concave increasing function on $\mathbb{R}$, the function $u \circ f_\varepsilon$ is a concave and increasing function on $[0, \infty)$, which can be extended to a concave increasing function v_ε on the full real line. Therefore,

$$\int u \, d\nu = \lim_{\varepsilon \downarrow 0} \int v_\varepsilon \, d\mu \ge \lim_{\varepsilon \downarrow 0} \int v_\varepsilon \, d\widetilde{\mu} = \int u \, d\widetilde{\nu}. \tag{2.21}$$

Consequently, $\nu \succcurlyeq_{\text{uni}} \widetilde{\nu}$ and Proposition 2.60 yields $\sigma^2 \le \widetilde{\sigma}^2$. □

Remark 2.68. The inequality (2.21) shows that if $\nu = N(\alpha, \sigma^2)$, $\widetilde{\nu} = N(\widetilde{\alpha}, \widetilde{\sigma}^2)$ and μ and $\widetilde{\mu}$ denote the images of ν and $\widetilde{\nu}$ under the map $x \mapsto e^x$, then $\mu \succcurlyeq_{\text{uni}} \widetilde{\mu}$ implies $\nu \succcurlyeq_{\text{uni}} \widetilde{\nu}$. However, the converse implication "$\nu \succcurlyeq_{\text{uni}} \widetilde{\nu} \Rightarrow \mu \succcurlyeq_{\text{uni}} \widetilde{\mu}$" fails, as can be seen by increasing $\widetilde{\sigma}$ until $m(\widetilde{\mu}) > m(\mu)$. ◇

Because of its relation to the analysis of the Black–Scholes formula for option prices, we will now sketch a second proof of Proposition 2.67.

Second proof of Proposition 2.67. Let

$$Y_{m,\sigma} := m \cdot \exp\left(\sigma X - \frac{\sigma^2}{2}\right)$$

for a standard normally distributed random variable X. Then

$$E\big[\,(Y_{m,\sigma} - c)^+\,\big] = m\,\Phi(d_+) - c\,\Phi(d_-) \quad \text{with } d_\pm = \frac{\log\frac{x}{c} \pm \frac{1}{2}\sigma^2}{\sigma}\,;$$

see Example 5.57 in Chapter 5. Calculating the derivative of this expectation with respect to $\sigma > 0$, one finds that

$$\frac{d}{d\sigma} E\big[\,(Y_{m,\sigma} - c)^+\,\big] = \frac{d}{d\sigma}\big(m\,\Phi(d_+) - c\,\Phi(d_-)\big) = x\,\varphi(d_+) > 0,$$

see (5.35) in Chapter 5. The law $\mu_{m,\sigma}$ of $Y_{m,\sigma}$ satisfies $m(\mu_{m,\sigma}) = m$ for all $\sigma > 0$. Condition (c) of Corollary 2.62 implies that $\mu_{m,\sigma}$ is decreasing in $\sigma > 0$ with respect to $\succcurlyeq_{\text{uni}}$ and hence also with respect to $\succcurlyeq_{\text{con}}$, i.e., $\mu_{m,\sigma} \succcurlyeq_{\text{con}} \mu_{m,\widetilde{\sigma}}$ if and only if $\sigma \le \widetilde{\sigma}$. For two different expectations m and $\widetilde{m}$, simply use the monotonicity of the function $u(y) := (y - c)^+$ to conclude

$$\begin{aligned}\int u\, d\mu_{m,\sigma} &= E\big[u\big(m \cdot \exp(\sigma X - \sigma^2/2)\big)\,\big]\\ &\ge E\big[u\big(\widetilde{m} \cdot \exp(\sigma X - \sigma^2/2)\big)\,\big]\\ &\ge \int u\, d\mu_{\widetilde{m},\widetilde{\sigma}},\end{aligned}$$

provided that $m \ge \widetilde{m}$ and $0 < \sigma \le \widetilde{\sigma}$. □

The partial order $\succcurlyeq_{\text{uni}}$ was defined in terms of integrals against increasing concave functions. By taking the larger class of all concave functions as integrands, we arrived at the partial order $\succcurlyeq_{\text{con}}$ defined by (2.17) and characterized in Corollary 2.62. In the remainder of this section, we will briefly discuss the partial order of *stochastic dominance*, which is induced by increasing instead of concave functions:

Definition 2.69. Let μ and ν be two arbitrary probability measures on $\mathbb{R}$. We say that μ *stochastically dominates* ν and we write $\mu \succcurlyeq_{\text{mon}} \nu$ if

$$\int f\, d\mu \ge \int f\, d\nu \quad \text{for all bounded increasing functions } f \in C(\mathbb{R}).$$

Stochastic dominance is sometimes also called *first order stochastic dominance*. It is indeed a partial order on $\mathcal{M}_1(\mathbb{R})$: Reflexivity and transitivity are obvious, and antisymmetry follows, e.g., from the equivalence (a)⇔(b) below. As will be shown by the following theorem, the relation $\mu \succcurlyeq_{\text{mon}} \nu$ means that the distribution μ is "higher" than the distribution ν. In our one-dimensional situation, we can provide a complete proof of this fact by using elementary properties of distribution functions. The general version of this result, given in Theorem 2.95, will require different techniques.

Theorem 2.70. *For $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ the following conditions are equivalent.*

(a) $\mu \succcurlyeq_{\text{mon}} \nu$.

(b) *The distribution functions of μ and ν satisfy $F_\mu(x) \le F_\nu(x)$ for all x.*

(c) *Any pair of quantile functions for μ and ν satisfies $q_\mu(t) \ge q_\nu(t)$ for a.e. $t \in (0, 1)$.*

(d) *There exists a probability space $(\Omega, \mathcal{F}, P)$ with random variables X_μ and X_ν with distributions μ and ν such that $X_\mu \ge X_\nu$ P-a.s.*

(e) *There exists a stochastic kernel $Q(x, dy)$ on $\mathbb{R}$ such that $Q(x, (-\infty, x]) = 1$ and such that $\nu = \mu Q$.*

In particular, $\mu \succcurlyeq_{\text{mon}} \nu$ implies $\mu \succcurlyeq_{\text{uni}} \nu$.

Proof. (a) ⇒(b): Note that $F_\mu(x) = \mu\big((-\infty, x]\big)$ can be written as

$$F_\mu(x) = 1 - \int I_{(x,\infty)}(y)\,\mu(dy).$$

It is easy to construct a sequence of increasing continuous functions with values in $[0, 1]$ which increase to $I_{(x,\infty)}$ for each x. Hence,

$$\int I_{(x,\infty)}(y)\,\mu(dy) \ge \int I_{(x,\infty)}(y)\,\nu(dy) = 1 - F_\nu(x).$$

(b)⇔(c): This follows from the definition of a quantile function and from Lemma A.17.

(c) ⇒(d): Let $(\Omega, \mathcal{F}, P)$ be a probability space supporting a random variable U with a uniform distribution on $(0, 1)$. Then $X_\mu := q_\mu(U)$ and $X_\nu := q_\nu(U)$ satisfy $X_\mu \ge X_\nu$ P-almost surely. Moreover, it follows from Lemma A.19 that they have the distributions μ and ν.

(d) ⇒(e): This is proved as in Theorem 2.58 by using regular conditional distributions.

(e) ⇒(a): Condition (e) implies that $x \ge y$ for $Q(x, \cdot)$-a.e. y. Hence, if f is bounded and increasing, then

$$\int f(y)\,Q(x, dy) \le \int f(x)\,Q(x, dy) = f(x).$$

Therefore,

$$\int f\,d\nu = \int\int f(y)\,Q(x, dy)\,\mu(dx) \le \int f\,d\mu.$$

Finally, due to the equivalence (a) ⇔ (b) above and the equivalence (a) ⇔ (d) in Theorem 2.58, $\mu \succcurlyeq_{\text{mon}} \nu$ implies $\mu \succcurlyeq_{\text{uni}} \nu$. □

Remark 2.71. It is clear from conditions (d) or (e) of Theorem 2.70 that the set of bounded, increasing, and continuous functions in Definition 2.69 can be replaced by the set of all increasing functions for which the two integrals make sense. Thus, $\mu \succcurlyeq_{\text{mon}} \nu$ for $\mu, \nu \in \mathcal{M}$ implies $\mu \succcurlyeq_{\text{uni}} \nu$, and in particular $m(\mu) \geq m(\nu)$. Moreover, condition (d) shows that $\mu \succcurlyeq_{\text{mon}} \nu$ together with $m(\mu) = m(\nu)$ implies $\mu = \nu$. ◇

2.5 Robust preferences on asset profiles

In this section, we discuss the structure of preferences for assets on a more fundamental level. Instead of assuming that the distributions of assets are known and that preferences are defined on a set of probability measures, we will take as our basic objects the assets themselves. An asset will be viewed as a function which associates real-valued payoffs to possible scenarios. More precisely, $\mathcal{X}$ will denote a set of bounded measurable functions X on some measurable set $(\Omega, \mathcal{F})$. We emphasize that no a priori probability measure is given on $(\Omega, \mathcal{F})$. In other words, we are facing uncertainty instead of risk.

We assume that $\mathcal{X}$ is endowed with a preference relation $\succ$. In view of the financial interpretation, it is natural to assume that $\succ$ is *monotone* in the sense that

$$Y \succeq X \quad \text{if } Y(\omega) \geq X(\omega) \text{ for all } \omega \in \Omega.$$

Under a suitable condition of *continuity*, we could apply the results of Section 2.1 to obtain a numerical representation of $\succ$. L. J. Savage introduced a set of additional axioms which guarantee there is a numerical representation of the special form

$$U(X) = E_Q[\, u(X)\,] = \int u\big(X(\omega)\big)\, Q(d\omega) \quad \text{for all } X \in \mathcal{X} \tag{2.22}$$

where Q is a probability measure on $(\Omega, \mathcal{F})$ and u is a function on $\mathbb{R}$. The measure Q specifies the subjective view of the probabilities of events which is implicit in the preference relation $\succ$. Note that the function $u : \mathbb{R} \to \mathbb{R}$ is determined by restricting U to the class of constant functions on $(\Omega, \mathcal{F})$. Clearly, the monotonicity of $\succ$ is equivalent to the condition that u is an increasing function.

Definition 2.72. A numerical representation of the form (2.22) will be called a *Savage representation* of the preference relation $\succ$.

Remark 2.73. Let $\mu_{Q,X}$ denote the distribution of X under the subjective measure Q. Clearly, the preference order $\succ$ on $\mathcal{X}$ given by (2.22) induces a preference order on

$$\mathcal{M}_Q := \{\, \mu_{Q,X} \mid X \in \mathcal{X}\,\}$$

with von Neumann–Morgenstern representation

$$U_Q(\mu_{Q,X}) := U(X) = E_Q[\, u(X)\,] = \int u \, d\mu_{Q,X},$$

i.e.,

$$U_Q(\mu) = \int u(x)\,\mu(dx) \quad \text{for } \mu \in \mathcal{M}_Q.$$

On this level, Section 2.3 specifies the conditions on U_Q which guarantee that u is a (strictly concave and strictly increasing) utility function. $\diamond$

Remark 2.74. Even if an economic agent with preferences $\succ$ would accept the view that scenarios $\omega \in \Omega$ are generated in accordance to a given objective probability measure P on $(\Omega, \mathcal{F})$, the preference order $\succ$ on $\mathcal{X}$ may be such that the subjective measure Q appearing in the Savage representation (2.22) is different from the objective measure P. Suppose, for example, that P is Lebesgue measure restricted to $\Omega = [0, 1]$, and that $\mathcal{X}$ is the space of bounded right-continuous increasing functions on $[0, 1]$. Let $\mu_{P,X}$ denote the distribution of X under P. By Lemma A.19, every probability measure on $\mathbb{R}$ with bounded support is of the form $\mu_{P,X}$ for some $X \in \mathcal{X}$, i.e.,

$$\mathcal{M}_b(\mathbb{R}) = \{\, \mu_{P,X} \mid X \in \mathcal{X} \,\}.$$

Suppose the agent agrees that, objectively, $X \in \mathcal{X}$ can be identified with the lottery $\mu_{P,X}$, so that the preference relation on $\mathcal{X}$ could be viewed as a preference relation on $\mathcal{M}_b(\mathbb{R})$ with numerical representation

$$U^*(\mu_{P,X}) := U(X).$$

This does not imply that U^* satisfies the assumptions of Section 2.2; in particular, the preference relation on $\mathcal{M}_b(\mathbb{R})$ may violate the independence axiom. In fact, the agent might take a pessimistic view and distort P by putting more emphasis on unfavorable scenarios. For example, the agent could replace P by the subjective measure

$$Q := \alpha\delta_0 + (1-\alpha)P$$

for some $\alpha \in (0, 1)$ and specify preferences by a Savage representation in terms of u and Q. In this case,

$$\begin{aligned} U^*(\mu_{P,X}) &= E_Q[\, u(X) \,] = \int u \, d\mu_{Q,X} \\ &= \alpha u\big(X(0)\big) + (1-\alpha) E_P[\, u(X) \,] \\ &= \alpha u\big(X(0)\big) + (1-\alpha) \int u \, d\mu_{P,X}. \end{aligned}$$

Note that $X(0) = \ell(\mu_{P,X})$ for

$$\ell(\mu) := \inf(\operatorname{supp}\mu) = \sup\big\{\, a \in \mathbb{R} \mid \mu\big((-\infty, a)\big) = 0 \,\big\},$$

where $\operatorname{supp}\mu$ is the support of μ. Hence, replacing P by Q corresponds to a non-linear distortion on the level of lotteries: $\mu = \mu_{P,X}$ is distorted to the lottery $\mu^* = \mu_{Q,X}$ given by

$$\mu^* = \alpha\, \delta_{\ell(\mu)} + (1-\alpha)\mu,$$

and the preference relation on lotteries has the numerical representation

$$U^*(\mu) = \int u(x)\,\mu^*(dx) \quad \text{for } \mu \in \mathcal{M}_b(\mathbb{R}).$$

Let us now show that such a subjective distortion of objective lotteries provides a possible explanation of the Allais paradox. Consider the lotteries μ_i and ν_i, $i = 1, 2$, described in Example 2.32. Clearly,

$$\mu_1^* = \mu_1 \quad \text{and} \quad \nu_1^* = \alpha\,\delta_0 + (1-\alpha)\nu_1,$$

while

$$\mu_2^* = \alpha\,\delta_0 + (1-\alpha)\mu_2 \quad \text{and} \quad \nu_1^* = \alpha\,\delta_0 + (1-\alpha)\nu_1.$$

For the particular choice $u(x) = x$ we have $U^*(\nu_2) > U^*(\mu_2)$, and for $\alpha > 9/2409$ we obtain $U^*(\mu_1) > U^*(\nu_1)$, in accordance with the observed preferences $\nu_2 \succ \mu_2$ and $\mu_1 \succ \nu_1$ described in Example 2.32.

For a systematic discussion of preferences described in terms of a subjective distortion of lotteries we refer to [131]. In Section 4.6, we will discuss the role of distortions in the context of risk measures, and in particular the connection to Yaari's "dual theory of choice under risk" [198]. ◇

Even in its general form (2.22), however, the paradigm of expected utility has a limited scope as illustrated by the following example.

Example 2.75 (Ellsberg paradox). You are faced with a choice between two urns, each containing 100 balls which are either red or black. In the first urn, the proportion p of red balls is know; assume, e.g., $p = 0.49$. In the second urn, the proportion $\widetilde{p}$ is unknown. Suppose that you get 1000 € if you draw a red ball and 0 € otherwise. In this case, most people would choose the first urn. Naturally, they make the same choice if you get 1000 € for drawing a black ball and 0 € for a red one. But this behavior is not compatible with the paradigm of expected utility: For any subjective probability $\widetilde{p}$ of drawing a red ball in the second urn, the first choice would imply $p > \widetilde{p}$, the second would yield $1 - p > 1 - \widetilde{p}$, and this is a contradiction. ◇

For this reason, we are going to make one further conceptual step beyond the Savage representation before we start to prove a representation theorem for preferences on $\mathcal{X}$. Instead of a single measure Q, let us consider a whole class $\mathcal{Q}$ of measures on $(\Omega, \mathcal{F})$. Our aim is to characterize those preference relations on $\mathcal{X}$ which admit a representation of the form

$$U(X) = \inf_{Q \in \mathcal{Q}} E_Q[\,u(X)\,]. \tag{2.23}$$

This may be viewed as a *robust* version of the paradigm of expected utility: The agent has in mind a whole collection of possible probabilistic views of the given set of scenarios and takes a *worst-case approach* in evaluating the expected utility of a given payoff.

It will be convenient to extend the discussion to the following framework where payoffs can be lotteries. Let $\mathcal{X}$ denote the space of all bounded measurable functions on $(\Omega, \mathcal{F})$. We are going to embed $\mathcal{X}$ into a certain space $\widetilde{\mathcal{X}}$ of functions $\widetilde{X}$ on $(\Omega, \mathcal{F})$ with values in the convex set

$$\mathcal{M}_b(\mathbb{R}) = \left\{ \mu \in \mathcal{M}_1(\mathbb{R}) \mid \mu([-c, c]) = 1 \text{ for some } c \geq 0 \right\}$$

of boundedly supported Borel probability measures on $\mathbb{R}$. More precisely, $\widetilde{\mathcal{X}}$ is defined as the convex set of all those stochastic kernels $\widetilde{X}(\omega, dy)$ from $(\Omega, \mathcal{F})$ to $\mathbb{R}$ for which there exists a constant $c \geq 0$ such that

$$\widetilde{X}\big(\omega, [-c, c]\big) = 1 \quad \text{for all } \omega \in \Omega.$$

The space $\mathcal{X}$ can be embedded into $\widetilde{\mathcal{X}}$ by virtue of the mapping

$$\mathcal{X} \ni X \longmapsto \delta_X \in \widetilde{\mathcal{X}}. \tag{2.24}$$

In this way, $\mathcal{X}$ can be identified with the set of all $\widetilde{X} \in \widetilde{\mathcal{X}}$ for which the measure $\widetilde{X}(\omega, \cdot)$ is a Dirac measure. A preference order on $\mathcal{X}$ defined by (2.23) clearly extends to $\widetilde{\mathcal{X}}$ by

$$\widetilde{U}(\widetilde{X}) = \inf_{Q \in \mathcal{Q}} \int \int u(y)\, \widetilde{X}(\omega, dy)\, Q(d\omega) = \inf_{Q \in \mathcal{Q}} E_Q[\, \widetilde{u}(\widetilde{X})\,] \tag{2.25}$$

where $\widetilde{u}$ is the affine function on $\mathcal{M}_b(\mathbb{R})$ defined by

$$\widetilde{u}(\mu) = \int u \, d\mu, \quad \mu \in \mathcal{M}_b(\mathbb{R}).$$

Remark 2.76. Restricting the preference order $\succ$ on $\widetilde{\mathcal{X}}$ obtained from (2.25) to the constant maps $\widetilde{X}(\omega) = \mu$ for $\mu \in \mathcal{M}_b(\mathbb{R})$, we obtain a preference order on $\mathcal{M}_b(\mathbb{R})$, and on this level we know how to characterize risk aversion by the property that u is strictly concave. ◇

Example 2.77. Let us show how the Ellsberg paradox fits into our extended setting, and how it can be resolved by a suitable choice of the set $\mathcal{Q}$. For $\Omega = \{0, 1\}$ define

$$\widetilde{X}_0(\omega) := p\, \delta_{1000} + (1-p)\delta_0, \quad \widetilde{X}_1(\omega) := (1-p)\delta_{1000} + p\, \delta_0,$$

and

$$\widetilde{Z}_i(\omega) := \delta_{1000} \cdot \mathrm{I}_{\{i\}}(\omega) + \delta_0 \cdot \mathrm{I}_{\{1-i\}}(\omega), \quad i = 0, 1.$$

Take

$$\mathcal{Q} := \left\{ q\, \delta_1 + (1-q)\delta_0 \mid a \leq q \leq b \right\}$$

with $[a, b] \subset [0, 1]$. For any increasing function u, the functional

$$\widetilde{U}(\widetilde{X}) := \inf_{Q \in \mathcal{Q}} E_Q[\, \widetilde{u}(\widetilde{X})\,]$$

satisfies

$$\widetilde{U}(\widetilde{X}_i) > \widetilde{U}(\widetilde{Z}_i), \quad i = 0, 1,$$

as soon as $a < p < b$, in accordance with the preferences described in Example 2.75. ◇

Let us now formulate those properties of a preference order $\succ$ on the convex set $\widetilde{\mathcal{X}}$ which are crucial for a representation of the form (2.25). For $\widetilde{X}, \widetilde{Y} \in \widetilde{\mathcal{X}}$ and $\alpha \in (0, 1)$, (2.25) implies

$$\begin{aligned}\widetilde{U}\big(\alpha\widetilde{X} + (1-\alpha)\widetilde{Y}\big) &= \inf_{Q\in\mathcal{Q}} \big(\alpha\, E_Q[\,\widetilde{u}(\widetilde{X})\,] + (1-\alpha)\, E_Q[\,\widetilde{u}(\widetilde{Y})\,]\big)\\ &\geq \alpha\,\widetilde{U}(\widetilde{X}) + (1-\alpha)\widetilde{U}(\widetilde{Y}).\end{aligned}$$

In contrast to the Savage case $\mathcal{Q} = \{Q\}$, we can no longer expect equality, except for the case of *certainty* $\widetilde{Y}(\omega) \equiv \mu$. If $\widetilde{X} \sim \widetilde{Y}$, then $\widetilde{U}(\widetilde{X}) = \widetilde{U}(\widetilde{Y})$, and the lower bound reduces to $\widetilde{U}(\widetilde{X}) = \widetilde{U}(\widetilde{Y})$. Thus, $\succ$ satisfies the following two properties:

Uncertainty aversion: If $\widetilde{X}, \widetilde{Y} \in \widetilde{\mathcal{X}}$ are such that $\widetilde{X} \sim \widetilde{Y}$, then

$$\alpha\widetilde{X} + (1-\alpha)\widetilde{Y} \succeq \widetilde{X} \quad \text{for all } \alpha \in [0, 1].$$

Certainty independence: For $\widetilde{X}, \widetilde{Y} \in \widetilde{\mathcal{X}}$, $\widetilde{Z} \equiv \mu \in \mathcal{M}_b(\mathbb{R})$, and $\alpha \in (0, 1]$ we have

$$\widetilde{X} \succ \widetilde{Y} \quad \Longleftrightarrow \quad \alpha\widetilde{X} + (1-\alpha)\widetilde{Z} \succ \alpha\widetilde{Y} + (1-\alpha)\widetilde{Z}.$$

Remark 2.78. In order to motivate the term "uncertainty aversion", consider the situation of the preceding example. Suppose that an agent is indifferent between the choices $\widetilde{Z}_0$ and $\widetilde{Z}_1$, which both involve the same kind of uncertainty. For $\alpha \in (0, 1)$, the convex combination $\widetilde{Y} := \alpha\widetilde{Z}_0 + (1-\alpha)\widetilde{Z}_1$, which is weakly preferred to both $\widetilde{Z}_0$ and $\widetilde{Z}_1$ in the case of uncertainty aversion, takes the form

$$\widetilde{Y}(\omega) = \begin{cases} \alpha\,\delta_{1000} + (1-\alpha)\delta_0 & \text{for } \omega = 1,\\ \alpha\,\delta_0 + (1-\alpha)\delta_{1000} & \text{for } \omega = 0,\end{cases}$$

i.e., uncertainty is reduced in favor of risk. For $\alpha = 1/2$, the resulting lottery $\widetilde{Y}(\omega) \equiv \frac{1}{2}(\delta_{1000} + \delta_0)$ is independent of the scenario ω, i.e., uncertainty is completely replaced by risk. ◇

Remark 2.79. The axiom of "certainty independence" extends the independence axiom for preferences on lotteries to our present setting, but only under the restriction that one of the two contingent lotteries $\widetilde{X}$ and $\widetilde{Y}$ is *certain*, i.e., does not depend on the scenario $\omega \in \Omega$. Without this restriction, the extended independence axiom would lead to the Savage representation in its original form (2.22). There are good reasons for *not* requiring full independence for all $\widetilde{Z} \in \widetilde{\mathcal{X}}$. As an example, take $\Omega = \{0, 1\}$ and define $\widetilde{X}(\omega) = \delta_\omega$, $\widetilde{Y}(\omega) = \delta_{1-\omega}$, and $\widetilde{Z} = \widetilde{X}$. An agent may prefer $\widetilde{X}$ over $\widetilde{Y}$, thus expressing the implicit view that scenario 1 is somewhat more likely than scenario 0. At the same time, the agent may like the idea of *hedging* against the occurrence of scenario 0, and this could mean that the certain lottery

$$\frac{1}{2}\big(\widetilde{Y}+\widetilde{Z}\big)(\cdot) \equiv \frac{1}{2}(\delta_0+\delta_1)$$

is preferred over the contingent lottery

$$\frac{1}{2}\big(\widetilde{X}+\widetilde{Z}\big)(\cdot) \equiv \widetilde{X}(\cdot),$$

thus violating the independence assumption in its unrestricted form. In general, the role of $\widetilde{Z}$ as a hedge against scenarios unfavorable for $\widetilde{Y}$ requires that $\widetilde{Y}$ and $\widetilde{Z}$ are *not comonotone*, i.e.,

$$\exists\, \omega,\ \eta \in \Omega: \quad \widetilde{Y}(\omega) \succ \widetilde{Y}(\eta),\ \widetilde{Z}(\omega) \prec \widetilde{Z}(\eta). \tag{2.26}$$

Thus, the wish to hedge would still be compatible with the following enforcement of certainty independence, called

- *comonotonic independence:* For $\widetilde{X},\ \widetilde{Y},\ \widetilde{Z} \in \widetilde{\mathcal{X}}$ and $\alpha \in (0,1]$

  $$\widetilde{X} \succ \widetilde{Y} \quad \Longleftrightarrow \quad \alpha\widetilde{X} + (1-\alpha)\widetilde{Z} \succ \alpha\widetilde{Y} + (1-\alpha)\widetilde{Z}.$$

  whenever $\widetilde{Y}$ and $\widetilde{Z}$ are *comonotone* in the sense that (2.26) does *not* occur. ◇

From now on, we assume that $\succ$ is a given preference order on $\widetilde{\mathcal{X}}$. The set $\mathcal{M}_b(\mathbb{R})$ will be regarded as a subset of $\widetilde{\mathcal{X}}$ by identifying a constant function $\widetilde{Z} \equiv \mu$ with its value $\mu \in \mathcal{M}_b(\mathbb{R})$. We assume that $\succ$ possesses the following properties:

- *Uncertainty aversion.*
- *Certainty independence.*
- *Monotonicity:* If $\widetilde{Y}(\omega) \succeq \widetilde{X}(\omega)$ for all $\omega \in \Omega$, then $\widetilde{Y} \succeq \widetilde{X}$. Moreover, $\succ$ is compatible with the usual order on $\mathbb{R}$, i.e., $\delta_y \succ \delta_x$ if and only if $y > x$.
- *Continuity:* The following analogue of the *Archimedean axiom* holds on $\widetilde{\mathcal{X}}$: If $\widetilde{X}, \widetilde{Y}, \widetilde{Z} \in \widetilde{\mathcal{X}}$ are such that $\widetilde{Z} \succ \widetilde{Y} \succ \widetilde{X}$, then there are $\alpha, \beta \in (0,1)$ with

  $$\alpha\widetilde{Z} + (1-\alpha)\widetilde{X} \succ \widetilde{Y} \succ \beta\widetilde{Z} + (1-\beta)\widetilde{X}.$$

  Moreover, for all $c > 0$ the restriction of $\succ$ to $\mathcal{M}_1([-c,c])$ is continuous with respect to the weak topology.

Let us denote by

$$\mathcal{M}_{1,f} := \mathcal{M}_{1,f}(\Omega, \mathcal{F})$$

the class of all set functions $Q : \mathcal{F} \to [0,1]$ which are normalized to $Q[\,\Omega\,] = 1$ and which are *finitely additive*, i.e., $Q[\,A \cup B\,] = Q[\,A\,] + Q[\,B\,]$ for all disjoint $A, B \in \mathcal{F}$. By $E_Q[\,X\,]$ we denote the integral of X with respect to $Q \in \mathcal{M}_{1,f}$; see Appendix A.6. With $\mathcal{M}_1(\Omega, \mathcal{F})$ we denote the σ-additive members of $\mathcal{M}_{1,f}$, that is, the class of all probability measures on $(\Omega, \mathcal{F})$.

Theorem 2.80. *Consider a preference order $\succ$ on $\widetilde{\mathcal{X}}$ satisfying the four properties listed above.*

(a) *There exists a strictly increasing function $u \in C(\mathbb{R})$ and a convex set $\mathcal{Q} \subset \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ such that*

$$\widetilde{U}(\widetilde{X}) = \min_{Q \in \mathcal{Q}} E_Q\Big[\int u(x)\,\widetilde{X}(\cdot, dx)\Big]$$

is a numerical representation of $\succ$. Moreover, u is unique up to positive affine transformations.

(b) *If the induced preference order $\succ$ on $\mathcal{X}$, viewed as a subset of $\widetilde{\mathcal{X}}$ as in* (2.24)*, satisfies the following additional continuity property*

$$X \succ Y \text{ and } X_n \nearrow X \quad \Longrightarrow \quad X_n \succ Y \quad \text{for all large } n, \tag{2.27}$$

then the set functions in $\mathcal{Q}$ are in fact probability measures, i.e., each $Q \in \mathcal{Q}$ is σ-additive. In this case, the induced preference order on $\mathcal{X}$ has the robust Savage representation

$$U(X) = \min_{Q \in \mathcal{Q}} E_Q[\,u(X)\,] \quad \text{for } X \in \mathcal{X}$$

with $\mathcal{Q} \subset \mathcal{M}_1(\Omega, \mathcal{F})$.

Remark 2.81. Even without its axiomatic foundation, the robust Savage representation is highly plausible as it stands, since it may be viewed as a worst-case approach to the problem of model uncertainty. This aspect will be of particular relevance in our discussion of risk measures in Chapter 4. ◇

The proof of Theorem 2.80 needs some preparation.

When restricted to $\mathcal{M}_b(\mathbb{R})$, viewed as a subset of $\widetilde{\mathcal{X}}$, the axiom of certainty independence is just the *independence axiom* of the von Neumann–Morgenstern theory. Thus, the preference relation $\succ$ on $\mathcal{M}_b(\mathbb{R})$ satisfies the assumptions of Corollary 2.29, and we obtain the existence of a continuous function $u : \mathbb{R} \to \mathbb{R}$ such that

$$\widetilde{u}(\mu) := \int u(x)\,\mu(dx) \tag{2.28}$$

is a numerical representation of $\succ$ on the set $\mathcal{M}_b(\mathbb{R})$. Moreover, u is unique up to positive affine transformations. The second part of our monotonicity assumption implies that u is strictly increasing. Without loss of generality, we assume $u(0) = 0$ and $u(1) = 1$.

Remark 2.82. In view of the representation (2.28), it follows as in (2.9) that any $\mu \in \mathcal{M}_b(\mathbb{R})$ admits a unique *certainty equivalent* $c(\mu) \in \mathbb{R}$ for which

$$\mu \sim \delta_{c(\mu)}.$$

Thus, if $X \in \mathcal{X}$ is defined for $\widetilde{X} \in \widetilde{\mathcal{X}}$ as $X(\omega) := c\big(\widetilde{X}(\omega)\big)$, then the first part of our monotonicity assumption yields

$$\widetilde{X} \sim \delta_X, \tag{2.29}$$

and so the preference relation $\succ$ on $\widetilde{\mathcal{X}}$ is uniquely determined by its restriction to $\mathcal{X}$. ◇

Lemma 2.83. *There exists a unique extension $\widetilde{U}$ of the functional $\widetilde{u}$ in* (2.28) *as a numerical representation of $\succ$ on $\widetilde{\mathcal{X}}$.*

Proof. For $\widetilde{X} \in \widetilde{\mathcal{X}}$ let $c > 0$ be such that $\widetilde{X}(\omega, [-c, c]) = 1$ for all $\omega \in \Omega$. Then

$$\widetilde{u}(\delta_{-c}) \le \widetilde{u}\big(\widetilde{X}(\omega)\big) \le \widetilde{u}(\delta_c) \quad \text{for all } \omega \in \Omega,$$

and our monotonicity assumption implies that

$$\delta_c \succeq \widetilde{X} \succeq \delta_{-c}.$$

We will show below that there exists a unique $\alpha \in [0, 1]$ such that

$$\widetilde{X} \sim (1 - \alpha)\delta_{-c} + \alpha\delta_c. \tag{2.30}$$

Once this has been achieved, the only possible choice for $\widetilde{U}(\widetilde{X})$ is

$$\widetilde{U}(\widetilde{X}) := \widetilde{u}\big((1 - \alpha)\delta_{-c} + \alpha\delta_c\big) = (1 - \alpha)\widetilde{u}(\delta_{-c}) + \alpha\widetilde{u}(\delta_c).$$

This definition of $\widetilde{U}$ provides a numerical representation of $\succ$ on $\widetilde{\mathcal{X}}$.

The proof of the existence of a unique $\alpha \in [0, 1]$ with (2.30) is similar to the proof of Lemma 2.24. Uniqueness follows from the monotonicity

$$\beta > \alpha \quad \Longrightarrow \quad (1 - \beta)\delta_{-c} + \beta\delta_c \succ (1 - \alpha)\delta_{-c} + \alpha\delta_c, \tag{2.31}$$

which is an immediate consequence of the von Neumann–Morgenstern representation. Now we let

$$\alpha := \sup\big\{\, \gamma \in [0, 1] \mid \widetilde{X} \succeq (1 - \gamma)\delta_{-c} + \gamma\delta_c \,\big\}.$$

We have to exclude the two following cases:

$$\widetilde{X} \succ (1 - \alpha)\delta_{-c} + \alpha\delta_c \tag{2.32}$$

$$(1 - \alpha)\delta_{-c} + \alpha\delta_c \succ \widetilde{X}. \tag{2.33}$$

In the case (2.32), our continuity axiom yields some $\beta \in (0, 1)$ for which

$$\widetilde{X} \succ \beta[\,(1 - \alpha)\delta_{-c} + \alpha\delta_c\,] + (1 - \beta)\delta_c = (1 - \gamma)\delta_{-c} + \gamma\delta_c$$

where $\gamma = \beta\alpha + (1-\beta) > \alpha$, in contradiction to the definition of α.

If (2.33) holds, then the same argument as above yields $\beta \in (0,1)$ with

$$\beta\alpha\delta_c + (1-\beta\alpha)\delta_{-c} \succ \widetilde{X}.$$

By our definition of α there must be some $\gamma \in (\beta\alpha, \alpha)$ with

$$\widetilde{X} \succeq (1-\gamma)\delta_{-c} + \gamma\delta_c \succ \beta\alpha\delta_c + (1-\beta\alpha)\delta_{-c},$$

where the second relation follows from (2.31). This, however, is a contradiction. □

Via the embedding (2.24), Lemma 2.83 induces a numerical representation U of $\succ$ on $\mathcal{X}$ given by

$$U(X) := \widetilde{U}(\delta_X). \tag{2.34}$$

The following proposition clarifies the properties of the functional U and provides the key to a robust Savage representation of the preference order $\succ$ on $\mathcal{X}$.

Proposition 2.84. *Given u of* (2.28) *and the numerical representation U on $\mathcal{X}$ constructed via Lemma* 2.83 *and* (2.34)*, there exists a unique functional $J : \mathcal{X} \to \mathbb{R}$ such that*

$$U(X) = J\big(u(X)\big) \quad \textit{for all } X \in \mathcal{X}, \tag{2.35}$$

and such that the following four properties are satisfied:

- Monotonicity: *If $Y(\omega) \geq X(\omega)$ for all ω, then $J(Y) \geq J(X)$.*
- Concavity: *If $\lambda \in [0,1]$ then $J\big(\lambda X + (1-\lambda)Y\big) \geq \lambda J(X) + (1-\lambda)J(Y)$.*
- Positive homogeneity: *$J(\lambda X) = \lambda J(X)$ for $\lambda \geq 0$.*
- Cash invariance: *$J(X+z) = J(X) + z$ for all $z \in \mathbb{R}$.*

Moreover, any functional J with these four properties is Lipschitz continuous on $\mathcal{X}$ with respect to the supremum norm $\|\cdot\|$, i.e.,

$$|J(X) - J(Y)| \leq \|X - Y\| \quad \textit{for all } X, Y \in \mathcal{X}.$$

Proof. Denote by $\mathcal{X}_u$ the space of all $X \in \mathcal{X}$ which take values in the range $u(\mathbb{R})$ of u. Clearly, $\mathcal{X}_u$ coincides with the range of the non-linear transformation $\mathcal{X} \ni X \mapsto u(X)$. Note that this transformation is bijective since u is strictly increasing due to our assumption of monotonicity. Thus, J is well-defined on $\mathcal{X}_u$ via (2.35). We show next that this J has the four properties of the assertion.

Monotonicity is obvious. For positive homogeneity on $\mathcal{X}_u$, it suffices to show that $J(\lambda X) = \lambda J(X)$ for $X \in \mathcal{X}_u$ and $\lambda \in (0,1]$. Let $X_0 \in \mathcal{X}$ be such that $u(X_0) = X$. We define $\widetilde{Z} \in \widetilde{\mathcal{X}}$ by

$$\widetilde{Z} := \lambda\delta_{X_0} + (1-\lambda)\delta_0.$$

By (2.29), $\widetilde{Z} \sim \delta_Z$ where Z is given by

$$\begin{aligned} Z(\omega) &= c\big(\lambda\delta_{X_0(\omega)} + (1-\lambda)\delta_0\big) \\ &= u^{-1}\big(\lambda u(X_0(\omega)) + (1-\lambda)u(0)\big) \\ &= u^{-1}\big(\lambda u(X_0(\omega))\big), \end{aligned}$$

where we have used our convention $u(0) = 0$. It follows that $u(Z) = \lambda u(X_0) = \lambda X$, and so

$$J(\lambda X) = U(Z) = \widetilde{U}(\widetilde{Z}). \tag{2.36}$$

As in (2.30), one can find $\nu \in \mathcal{M}_b(\mathbb{R})$ such that $\nu \sim \delta_{X_0}$. Certainty independence implies that

$$\widetilde{Z} = \lambda\delta_{X_0} + (1-\lambda)\delta_0 \sim \lambda\nu + (1-\lambda)\delta_0.$$

Hence,

$$\widetilde{U}(\widetilde{Z}) = \widetilde{u}(\lambda\nu + (1-\lambda)\delta_0) = \lambda\widetilde{u}(\nu) = \lambda U(X_0) = \lambda J(X).$$

This shows that J is positively homogeneous on $\mathcal{X}_u$.

Since the range of u is an interval, we can extend J from $\mathcal{X}_u$ to all of $\mathcal{X}$ by positive homogeneity, and this extension, again denoted J, is also monotone and positively homogeneous.

Let us now show that J is cash invariant. First note that

$$J(1) = \frac{J\big(u(x)\big)}{u(x)} = \frac{\widetilde{u}(\delta_x)}{u(x)} = 1$$

for any x such that $u(x) \neq 0$. Now take $X \in \mathcal{X}$ and $z \in \mathbb{R}$. By positive homogeneity, we may assume without loss of generality that $2X \in \mathcal{X}_u$ and $2z \in u(\mathbb{R})$. Then there are $X_0 \in \mathcal{X}$ such that $2X = u(X_0)$ as well as $z_0, x_0 \in \mathbb{R}$ with $2z = u(z_0)$ and $2J(X) = u(x_0)$. Note that $\delta_{X_0} \sim \delta_{x_0}$. Thus, certainty independence yields

$$\widetilde{Z} := \frac{1}{2}(\delta_{X_0} + \delta_{z_0}) \sim \frac{1}{2}(\delta_{x_0} + \delta_{z_0}) =: \mu.$$

On the one hand, it follows that

$$\widetilde{U}(\widetilde{Z}) = U(\mu) = \frac{1}{2}u(x_0) + \frac{1}{2}u(z_0) = J(X) + z.$$

On the other hand, the same reasoning which lead to (2.36) shows that

$$\widetilde{U}(\widetilde{Z}) = J(X + z).$$

As to concavity, we need only show that $J(\frac{1}{2}X + \frac{1}{2}Y) \geq \frac{1}{2}J(X) + \frac{1}{2}J(Y)$ for $X, Y \in \mathcal{X}_u$, by positive homogeneity. Let $X_0, Y_0 \in \mathcal{X}$ be such that $X = u(X_0)$ and $Y = u(Y_0)$. If $J(X) = J(Y)$, then $\delta_{X_0} \sim \delta_{Y_0}$, and uncertainty aversion gives

$$\widetilde{Z} := \frac{1}{2}(\delta_{X_0} + \delta_{Y_0}) \succeq \delta_{X_0},$$

which by the same arguments as above yields

$$\widetilde{U}(\widetilde{Z}) = J\left(\frac{1}{2}X + \frac{1}{2}Y\right) \geq J(X) = \frac{1}{2}\big(J(X) + J(Y)\big).$$

The case in which $J(X) > J(Y)$ can be reduced to the previous one by letting $z := J(X) - J(Y)$, and by replacing Y by $Y_z := Y + z$. Cash invariance then implies that

$$\begin{aligned} J\left(\frac{1}{2}X + \frac{1}{2}Y\right) + \frac{1}{2}z &= J\left(\frac{1}{2}X + \frac{1}{2}Y_z\right) \\ &\geq \frac{1}{2}\big(J(X) + J(Y_z)\big) \\ &= \frac{1}{2}\big(J(X) + J(Y)\big) + \frac{1}{2}z. \end{aligned}$$

Now we show the Lipschitz continuity of J. If $X, Y \in \mathcal{X}$ then $X \leq Y + \|X - Y\|$, and so $J(X) \leq J(Y) + \|X - Y\|$, by monotonicity and cash invariance. Reversing the roles of X and Y yields

$$|J(X) - J(Y)| \leq \|X - Y\|. \qquad \square$$

Let us now show that a function with the four properties established in the preceding proposition can be represented in terms of a family of set functions in the class $\mathcal{M}_{1,f}$.

Proposition 2.85. *A functional $J : \mathcal{X} \to \mathbb{R}$ is monotone, concave, positively homogeneous, and cash invariant if and only if there exists a set $\mathcal{Q} \subset \mathcal{M}_{1,f}$ such that*

$$J(X) = \inf_{Q \in \mathcal{Q}} E_Q[\,X\,], \quad X \in \mathcal{X}.$$

Moreover, the set $\mathcal{Q}$ can always be chosen to be convex and such that the infimum above is attained, i.e.,

$$J(X) = \min_{Q \in \mathcal{Q}} E_Q[\,X\,], \quad X \in \mathcal{X}.$$

Proof. The necessity of the four properties is obvious. Conversely, we will construct for any $X \in \mathcal{X}$ a finitely additive set function Q_X such that $J(X) = E_{Q_X}[\,X\,]$ and $J(Y) \leq E_{Q_X}[\,Y\,]$ for all $Y \in \mathcal{X}$. Then

$$J(Y) = \min_{Q \in \mathcal{Q}_0} E_Q[\,Y\,] \quad \text{for all } Y \in \mathcal{X} \tag{2.37}$$

where $\mathcal{Q}_0 := \{\, Q_X \mid X \in \mathcal{X} \,\}$. Clearly, (2.37) remains true if we replace $\mathcal{Q}_0$ by its convex hull $\mathcal{Q} := \operatorname{conv} \mathcal{Q}_0$.

To construct Q_X for a given $X \in \mathcal{X}$, we define three convex sets in $\mathcal{X}$ by

$$\mathcal{B} := \{\, Y \in \mathcal{X} \mid J(Y) > 1 \,\},$$

$$\mathcal{C}_1 := \{ Y \in \mathcal{X} \mid Y \le 1 \}, \quad \text{and} \quad \mathcal{C}_2 := \left\{ Y \in \mathcal{X} \;\middle|\; Y \le \frac{X}{J(X)} \right\}.$$

The convexity of $\mathcal{C}_1$ and $\mathcal{C}_2$ implies that the convex hull of their union is given by

$$\mathcal{C} := \operatorname{conv}(\mathcal{C}_1 \cup \mathcal{C}_2) = \big\{ \alpha Y_1 + (1-\alpha)Y_2 \mid Y_i \in \mathcal{C}_i \text{ and } \alpha \in [0,1] \big\}.$$

Since $Y \in \mathcal{C}$ is of the form $Y = \alpha Y_1 + (1-\alpha)Y_2$ for some $Y_i \in \mathcal{C}_i$ and $\alpha \in [0,1]$,

$$J(Y) \le J(\alpha + (1-\alpha)Y_2) = \alpha + (1-\alpha)J(Y_2) \le 1,$$

and so $\mathcal{B}$ and $\mathcal{C}$ are disjoint. Let $\mathcal{X}$ be endowed with the supremum norm $\|Y\| := \sup_{\omega\in\Omega} |Y(\omega)|$. Then $\mathcal{C}_1$, and hence $\mathcal{C}$, contains the unit ball in $\mathcal{X}$. In particular, $\mathcal{C}$ has non-empty interior. Thus, we may apply the separation argument in the form of Theorem A.54, which yields a non-zero continuous linear functional ℓ on $\mathcal{X}$ such that

$$c := \sup_{Y\in\mathcal{C}} \ell(Y) \le \inf_{Z\in\mathcal{B}} \ell(Z).$$

Since $\mathcal{C}$ contains the unit ball, c must be strictly positive, and there is no loss of generality in assuming $c = 1$. In particular, $\ell(1) \le 1$ as $1 \in \mathcal{C}$. On the other hand, any constant $b > 1$ is contained in $\mathcal{B}$, and so

$$\ell(1) = \lim_{b\downarrow 1} \ell(b) \ge c = 1.$$

Hence, $\ell(1) = 1$.

If $A \in \mathcal{F}$ then $I_{A^c} \in \mathcal{C}_1 \subset \mathcal{C}$, which implies that

$$\ell(I_A) = \ell(1) - \ell(I_{A^c}) \ge 1 - 1 = 0.$$

By Theorem A.50 there exists a finitely additive set function $Q_X \in \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ such that $\ell(Y) = E_{Q_X}[\,Y\,]$ for any $Y \in \mathcal{X}$.

It remains to show that $E_{Q_X}[\,Y\,] \ge J(Y)$ for all $Y \in \mathcal{X}$, with equality for $Y = X$. By the cash invariance of J, we need only consider the case in which $J(Y) > 0$. Then

$$Y_n := \frac{Y}{J(Y)} + \frac{1}{n} \in \mathcal{B},$$

and $Y_n \to Y/J(Y)$ uniformly, whence

$$\frac{E_{Q_X}[\,Y\,]}{J(Y)} = \lim_{n\uparrow\infty} E_{Q_X}[\,Y_n\,] \ge 1.$$

On the other hand, $X/J(X) \in \mathcal{C}_2 \subset \mathcal{C}$ yields the inequality

$$\frac{E_{Q_X}[\,X\,]}{J(X)} \le c = 1. \qquad \square$$

We are now ready to complete the proof of the main result in this section.

Proof of Theorem 2.80. (a): By Remark 2.82, it suffices to consider the induced preference relation $\succ$ on $\mathcal{X}$ once the function u has been determined. According to Lemma 2.83 and the two Propositions 2.84 and 2.85, there exists a convex set $\mathcal{Q} \subset \mathcal{M}_{1,f}$ such that

$$U(X) = \min_{Q \in \mathcal{Q}} E_Q[\, u(X)\,]$$

is a numerical representation of $\succ$ on $\mathcal{X}$. This proves the first part of the assertion.

(b): The assumption (2.27) applied to $X \equiv 1$ and $Y \equiv b < 1$ gives that any sequence with $X_n \nearrow 1$ is such that $X_n \succ b$ for large enough n. We claim that this implies that $U(X_n) \nearrow u(1) = 1$. Otherwise, $U(X_n)$ would increase to some number $a < 1$. Since u is continuous and strictly increasing, we may take b such that $a < u(b) < 1$. But then $U(X_n) > U(b) = u(b) > a$ for large enough n, which is a contradiction.

In particular, we obtain that for any increasing sequence $A_n \in \mathcal{F}$ such that $\bigcup_n A_n = \Omega$

$$\lim_{n\uparrow\infty} \min_{Q \in \mathcal{Q}} Q[\, A_n\,] = \lim_{n\uparrow\infty} U(\mathrm{I}_{A_n}) = 1.$$

But this means that each $Q \in \mathcal{Q}$ satisfies $\lim_n Q[\, A_n\,] = 1$, which is equivalent to the σ-additivity of Q. □

The continuity assumption (2.27), required for *all* $X_n \in \mathcal{X}$, is actually quite strong. In a topological setting, our discussion of risk measures in Chapter 4 will imply the following version of the representation theorem.

Proposition 2.86. *Consider a preference order $\succ$ as in Theorem* 2.80. *Suppose that Ω is a Polish space with Borel field $\mathcal{F}$ and that* (2.27) *holds if X_n and X are continuous. Then there exists a class of probability measures $\mathcal{Q} \subset \mathcal{M}_1(\Omega, \mathcal{F})$ such that the induced preference order on $\mathcal{X}$ has the* robust Savage representation

$$U(X) = \min_{Q \in \mathcal{Q}} E_Q[\, u(X)\,] \quad \textit{for continuous } X \in \mathcal{X}.$$

Proof. As in the proof of Theorem 2.80, the continuity property of $\succ$ implies the corresponding continuity property of U, and hence of the functional J in (2.35). The result follows by combining Proposition 2.84, which reduces the representation of U to a representation of J, with Proposition 4.25 applied to the coherent risk measure $\rho := -J$. □

Finally, we consider an alternative setting where we fix in advance a reference measure P on $(\Omega, \mathcal{F})$. In this context, $\mathcal{X}$ will be identified with the space $L^\infty(\Omega, \mathcal{F}, P)$, and the representation of preferences will involve measures which are absolutely continuous with respect to P. Note, however, that this passage from measurable functions to equivalence classes of random variables in $L^\infty(\Omega, \mathcal{F}, P)$, and from arbitrary probability measures to absolutely continuous measures, involves a certain loss of robustness in the face of model uncertainty.

Theorem 2.87. *Let $\succ$ be a preference relation as in Theorem* 2.80*, and assume that*

$$X \sim Y \quad \textit{whenever } X = Y \ P\text{-a.s.}$$

(a) *There exists a robust Savage representation of the form*

$$U(X) = \inf_{Q \in \mathcal{Q}} E_Q\big[\, u(X)\,\big], \quad X \in \mathcal{X},$$

where $\mathcal{Q}$ consists of probability measures on $(\Omega, \mathcal{F})$ which are absolutely continuous with respect to P, if and only if $\succ$ satisfies the following condition of continuity from above:

$$Y \succ X \text{ and } X_n \searrow X \ P\text{-a.s.} \quad \Longrightarrow \quad Y \succ X_n \quad P\text{-a.s. for all large } n.$$

(b) *There exists a representation of the form*

$$U(X) = \min_{Q \in \mathcal{Q}} E_Q\big[\, u(X)\,\big], \quad X \in \mathcal{X},$$

where $\mathcal{Q}$ consists of probability measures on $(\Omega, \mathcal{F})$ which are absolutely continuous with respect to P, if and only if $\succ$ satisfies the following condition of continuity from below:

$$X \succ Y \text{ and } X_n \nearrow X \ P\text{-a.s.} \quad \Longrightarrow \quad X_n \succ Y \quad P\text{-a.s. for all large } n.$$

Proof. As in the proof of Theorem 2.80, the continuity property of $\succ$ implies the corresponding continuity property of U, and hence of the functional J in (2.35). The results follow by combining Proposition 2.84, which reduces the representation of U to a representation of J, with Corollary 4.34 and Corollary 4.35 applied to the coherent risk measure $\rho := -J$. □

2.6 Probability measures with given marginals

In this section, we study the construction of probability measures with given marginals. In particular, this will yield the missing implication in the characterization of uniform preference in Theorem 2.58, but the results in this section are of independent interest. We focus on the following basic question: Suppose μ_1 and μ_2 are two probability measures on S, and Λ is a convex set of probability measures on $S \times S$; when does Λ contain some $\overline{\mu}$ which has μ_1 and μ_2 as marginals?

The answer to this question will be given in a general topological setting. Let S be a Polish space, and let us fix a continuous function ψ on S with values in $[1, \infty)$. As in Section 2.2 and in Appendix A.6, we use ψ as a *gauge function* in order to define the space of measures

$$\mathcal{M}_1^{\psi}(S) := \Big\{ \mu \in \mathcal{M}_1(S) \;\Big|\; \int \psi(x)\, \mu(dx) < \infty \Big\}$$

and the space of continuous test functions

$$C_\psi(S) := \left\{ f \in C(S) \mid \exists c : |f(x)| \le c \cdot \psi(x) \text{ for all } x \in S \right\}.$$

The ψ-weak topology on $\mathcal{M}_1^\psi(S)$ is the coarsest topology such that

$$\mathcal{M}_1^\psi(S) \ni \mu \mapsto \int f \, d\mu$$

is a continuous mapping for all $f \in C_\psi(S)$; see Appendix A.6 for details. On the product space $S \times S$, we take the gauge function

$$\overline{\psi}(x, y) := \psi(x) + \psi(y),$$

and define the corresponding set $\mathcal{M}_1^{\overline{\psi}}(S \times S)$, which will be endowed with the $\overline{\psi}$-weak topology.

Theorem 2.88. *Suppose that $\Lambda \subset \mathcal{M}_1^{\overline{\psi}}(S \times S)$ is convex and closed in the $\overline{\psi}$-weak topology, and that μ_1, μ_2 are probability measures in $\mathcal{M}_1^\psi(S)$. Then there exists some $\overline{\mu} \in \Lambda$ with marginal distributions μ_1 and μ_2 if and only if*

$$\int f_1 \, d\mu_1 + \int f_2 \, d\mu_2 \le \sup_{\lambda \in \Lambda} \int \big(f_1(x) + f_2(y)\big) \, \lambda(dx, dy) \quad \textit{for all } f_1, f_2 \in C_\psi(S).$$

Theorem 2.88 is due to V. Strassen [191]. Its proof boils down to an application of the Hahn–Banach theorem; the difficult part consists in specifying the right topological setting. First, let us investigate the relations between $\mathcal{M}_1^{\overline{\psi}}(S \times S)$ and $\mathcal{M}_1^\psi(S)$. To this end, we define mappings

$$\pi_i : \mathcal{M}_1^{\overline{\psi}}(S \times S) \to \mathcal{M}_1^\psi(S), \qquad i = 1, 2,$$

that yield the i^{th} marginal distribution of a measure $\lambda \in \mathcal{M}_1^{\overline{\psi}}(S \times S)$:

$$\int f \, d(\pi_1 \lambda) = \int f(x) \, \lambda(dx, dy) \quad \text{and} \quad \int f \, d(\pi_2 \lambda) = \int f(y) \, \lambda(dx, dy),$$

for all $f \in C_\psi(S)$.

Lemma 2.89. *π_1 and π_2 are continuous and affine mappings from $\mathcal{M}_1^{\overline{\psi}}(S \times S)$ to $\mathcal{M}_1^\psi(S)$.*

Proof. Suppose that λ_n converges to λ in $\mathcal{M}_1^{\overline{\psi}}(S \times S)$. For $f \in C_\psi(S)$ let $\overline{f}(x, y) := f(x)$. Clearly, $\overline{f} \in C_{\overline{\psi}}(S \times S)$, and thus

$$\int f \, d(\pi_1 \lambda_n) = \int \overline{f} \, d\lambda_n \longrightarrow \int \overline{f} \, d\lambda = \int f \, d(\pi_1 \lambda).$$

Therefore, π_1 is continuous, and the same is true of π_2. Affinity is obvious. □

Now, let us consider the linear space

$$E := \{ \alpha\mu - \beta\nu \mid \mu, \nu \in \mathcal{M}_1^{\psi}(S), \ \alpha, \beta \in \mathbb{R} \}$$

spanned by $\mathcal{M}_1^{\psi}(S)$. For $\rho = \alpha\mu - \beta\nu \in E$ the integral $\int f \, d\rho$ against a function $f \in C_{\psi}(S)$ is well-defined and given by

$$\int f \, d\rho = \alpha \int f \, d\mu - \beta \int f \, d\nu.$$

In particular, $\rho \mapsto \int f \, d\rho$ is linear functional on E, so we can regard $C_{\psi}(S)$ as a subset of the algebraic dual E^* of E. Note that $\int f \, d\rho = \int f \, d\widetilde{\rho}$ for all $f \in C_{\psi}(S)$ implies $\rho = \widetilde{\rho}$, i.e., $C_{\psi}(S)$ separates the points of E. We endow E with the coarsest topology $\sigma(E, C_{\psi}(S))$ for which all maps

$$E \ni \rho \mapsto \int f \, d\rho, \quad f \in C_{\psi}(S),$$

are continuous; see Definition A.57. With this topology, E becomes a locally convex topological vector space.

Lemma 2.90. *Under the above assumptions, $\mathcal{M}_1^{\psi}(S)$ is a closed convex subset of E, and the relative topology of the embedding coincides with the ψ-weak topology.*

Proof. The sets of the form

$$U_{\varepsilon}(\rho; f_1, \dots, f_n) := \bigcap_{i=1}^{n} \Big\{ \widetilde{\rho} \in E \;\Big|\; \Big| \int f_i \, d\rho - \int f_i \, d\widetilde{\rho} \Big| < \varepsilon \Big\}$$

with $\rho \in E$, $n \in \mathbb{N}$, $f_i \in C_{\psi}(S)$, and $\varepsilon > 0$ form a base of the topology $\sigma(E, C_{\psi}(S))$. Thus, if $U \subset E$ is open, then every point $\mu \in U \cap \mathcal{M}_1^{\psi}(S)$ possesses some neighborhood $U_{\varepsilon}(\mu; f_1, \dots, f_n) \subset U$. But $U_{\varepsilon}(\mu; f_1, \dots, f_n) \cap \mathcal{M}_1^{\psi}(S)$ is an open neighborhood of μ in the ψ-weak topology. Hence, $U \cap \mathcal{M}_1^{\psi}(S)$ is open in the ψ-weak topology. Similarly, one shows that every open set $V \subset \mathcal{M}_1^{\psi}(S)$ is of the form $V = U \cap \mathcal{M}_1^{\psi}(S)$ for some open subset U of E. This shows that the relative topology $\mathcal{M}_1^{\psi}(S) \cap \sigma\big(E, C_{\psi}(S)\big)$ coincides with the ψ-weak topology.

Moreover, $\mathcal{M}_1^{\psi}(S)$ is an intersection of closed subsets of E:

$$\mathcal{M}_1^{\psi}(S) = \Big\{ \rho \in E \;\Big|\; \int 1 \, d\rho = 1 \Big\} \cap \bigcap_{\substack{f \in C_{\psi}(S) \\ f \geq 0}} \Big\{ \rho \in E \;\Big|\; \int f \, d\rho \geq 0 \Big\}.$$

Therefore, $\mathcal{M}_1^{\psi}(S)$ is closed in E. □

Next, let E^2 denote the product space $E \times E$. We endow E^2 with the product topology for which the sets $U \times V$ with $U, V \in \sigma(E, C_\psi(S))$ form a neighborhood base. Clearly, E^2 is a locally convex topological vector space.

Lemma 2.91. *Every continuous linear functional ℓ on E^2 is of the form*

$$\ell(\rho_1, \rho_2) = \int f_1 \, d\rho_1 + \int f_2 \, d\rho_2$$

for some $f_1, f_2 \in C_\psi(S)$.

Proof. By linearity, ℓ is of the form $\ell(\rho_1, \rho_2) = \ell_1(\rho_1) + \ell_2(\rho_2)$, where $\ell_1(\rho_1) := \ell(\rho_1, 0)$ and $\ell_2(\rho_2) := \ell(0, \rho_2)$. By continuity of ℓ, the set

$$V := \ell^{-1}\big((-1, 1)\big)$$

is open in E^2 and contains the point $(0, 0)$. Hence, there are two open neighborhoods $U_1, U_2 \subset E$ such that $(0, 0) \in U_1 \times U_2 \subset V$. Therefore,

$$0 \in U_i \subset \ell_i^{-1}\big((-1, 1)\big) \quad \text{for } i = 1, 2,$$

i.e., 0 is an interior point of $\ell_i^{-1}\big((-1, 1)\big)$. It follows that the ℓ_i are continuous at 0, which in view of their linearity implies continuity everywhere on E. Finally, we may conclude from Proposition A.58 that each ℓ_i is of the form $\ell_i(\rho) = \int f_i \, d\rho_i$ for some $f_i \in C_\psi(S)$. □

The proof of the following lemma uses the characterization of compact sets for the ψ-weak topology that is stated in Corollary A.46. It is here that we need our assumption that S is Polish.

Lemma 2.92. *If Λ is a closed convex subset of $\mathcal{M}_1^{\bar\psi}(S \times S)$, then*

$$H_\Lambda := \{ (\pi_1\lambda, \pi_2\lambda) \mid \lambda \in \Lambda \}$$

is a closed convex subset of E^2.

Proof. It is enough to show that H_Λ is closed in $\mathcal{M}_1^\psi(S)^2 := \mathcal{M}_1^\psi(S) \times \mathcal{M}_1^\psi(S)$, because Lemma 2.90 implies that the relative topology induced by E^2 on $\mathcal{M}_1^\psi(S)^2$ coincides with the product topology for the ψ-weak topology. This is a metric topology by Corollary A.44. So let $(\mu_n, \nu_n) \in H_\Lambda$, $n \in \mathbb{N}$, be a sequence converging to some $(\mu, \nu) \in \mathcal{M}_1^\psi(S)^2$ in the product topology. Since both sequences $(\mu_n)_{n\in\mathbb{N}}$ and $(\nu_n)_{n\in\mathbb{N}}$ are relatively compact for the ψ-weak topology, Corollary A.46 yields functions $\phi_i : S \to [1, \infty]$, $i = 1, 2$, such that sets of the form $K_k^i := \{\phi_i \le k\psi\}$, $k \in \mathbb{N}$, are relatively compact in S and such that

$$\sup_{n\in\mathbb{N}} \int \phi_1 \, d\mu_n + \sup_{n\in\mathbb{N}} \int \phi_2 \, d\nu_n < \infty \, .$$

For each n, there exists $\lambda_n \in \Lambda$ such that $\pi_1\lambda_n = \mu_n$ and $\pi_2\lambda_n = \nu_n$. Hence, if we let $\overline{\phi}(x, y) := \phi_1(x) + \phi_2(y)$, then

$$\sup_{n\in\mathbb{N}} \int \overline{\phi}\, d\lambda_n = \sup_{n\in\mathbb{N}} \Big(\int \phi_1\, d\mu_n + \int \phi_2\, d\nu_n \Big) < \infty .$$

Moreover, we claim that each set $\{\overline{\phi} \le k\overline{\psi}\}$ is relatively compact in $S \times S$. To prove this claim, let $l_i \in \mathbb{N}$ be such that

$$l_i \ge \sup_{x\in K_k^i} \psi(x) .$$

Then, since $\psi \ge 1$,

$$\{\overline{\phi} \le k\overline{\psi}\} \subset K_k^1 \times K^2_{k(1+l_1)} \cup K^1_{k(1+l_2)} \times K_k^2 ,$$

and the right-hand side is a relatively compact set in $S \times S$. It follows from Corollary A.46 that the sequence $(\lambda_n)_{n\in\mathbb{N}}$ is relatively compact for the $\overline{\psi}$-weak topology. Any accumulation point λ of this sequence belongs to the closed set Λ. Moreover, λ has marginal distributions μ and ν, since the projections π_i are continuous according to Lemma 2.89. Hence $(\mu, \nu) \in H_\Lambda$. □

Proof of Theorem 2.88. Let $\mu_1, \mu_2 \in \mathcal{M}_1^\psi(S)$ be given. Since H_Λ is closed and convex in E^2 by Lemma 2.92, we may apply Theorem A.56 with $\mathcal{B} := \{(\mu_1, \mu_2)\}$ and $\mathcal{C} := H_\Lambda$: We conclude that $(\mu_1, \mu_2) \notin H_\Lambda$ if and only if there exists a linear functional ℓ on E^2 such that

$$\ell(\mu_1, \mu_2) > \sup_{(\nu_1,\nu_2)\in H_\Lambda} \ell(\nu_1, \nu_2) = \sup_{\lambda\in\Lambda} \ell(\pi_1\lambda, \pi_2\lambda).$$

Applying Lemma 2.91 to ℓ completes the assertion. □

We will now use Theorem 2.88 to deduce the remaining implication of Theorem 2.58. We consider here a more general, d-dimensional setting. Let $x = (x^1, \dots, x^d)$ and $y = (y^1, \dots, y^d)$ be two d-dimensional vectors. We will say that $x \le y$ if $x^i \le y^i$ for all i. A function on $\mathbb{R}^d$ is called increasing, if it is increasing with respect to the partial order $\le$.

Theorem 2.93. *Suppose μ_1 and μ_2 are Borel probability measures on $\mathbb{R}^d$ with $\int |x|\mu_i(dx) < \infty$ for $i = 1, 2$. Then the following assertions are equivalent.*

(a) $\displaystyle\int f\, d\mu_1 \ge \int f\, d\mu_2$ *for all increasing concave functions f on $\mathbb{R}^d$.*

(b) *There exists a probability space $(\Omega, \mathcal{F}, P)$ with random variables X_1 and X_2 having distributions μ_1 and μ_2, respectively, such that*

$$E[\, X_2 \mid X_1\,] \le X_1 \quad P\text{-a.s.}$$

(c) *There exists a kernel $Q(x, dy)$ on $\mathbb{R}^d$ such that*

$$\int_{\mathbb{R}^d} y\, Q(x, dy) \le x \quad \text{for all } x \in \mathbb{R}^d$$

and such that $\mu_2 = \mu_1 Q$.

Proof. (a)$\Rightarrow$(b): We will apply Theorem 2.88 with $S := \mathbb{R}^d$ and with the gauge functions $\psi(x) := 1 + |x|$ and $\overline{\psi}(x, y) := \psi(x) + \psi(y)$. We denote by $C_b(\mathbb{R}^d)$ the set of bounded and continuous functions on $\mathbb{R}^d$. Let

$$\Lambda := \bigcap_{f \in C_b(\mathbb{R}^d)} \Big\{ \lambda \in \mathcal{M}_1^{\overline{\psi}}(\mathbb{R}^d \times \mathbb{R}^d) \;\Big|\; \int y f(x)\, \lambda(dx, dy) \le \int x f(x)\, \lambda(dx, dy) \Big\}.$$

Each single set of the intersection is convex and closed in $\mathcal{M}_1^{\overline{\psi}}(\mathbb{R}^d \times \mathbb{R}^d)$, because the functions $g(x, y) := y f(x)$ and $\widetilde{g}(x, y) := x f(x)$ belong to $C_{\overline{\psi}}(\mathbb{R}^d \times \mathbb{R}^d)$ for $f \in C_b(S)$. Therefore, Λ itself is convex and closed.

Suppose we can show that Λ contains an element P that has μ_1 and μ_2 as marginal distributions. Then we can take $\Omega := \mathbb{R}^d \times \mathbb{R}^d$ with its Borel σ-algebra $\mathcal{F}$, and let X_1 and X_2 denote the canonical projections on the first and the second components, respectively. By definition, X_i will have the distribution μ_i, and

$$E\big[\, E[\, X_2 \mid X_1\,] f(X_1) \,\big] = E[\, X_2 f(X_1) \,] \le E[\, X_1 f(X_1) \,] \quad \text{for all } f \in C_b(\mathbb{R}^d).$$

By monotone class arguments, we may thus conclude that

$$E[\, X_2 \mid X_1 \,] \le X_1 \qquad P\text{-a.s.}$$

so that the assertion will follow.

It remains to prove the existence of P. To this end, we will apply Theorem 2.88 with the set Λ defined above. Take a pair $f_1, f_2 \in C_\psi(\mathbb{R}^d)$, and let

$$\widetilde{f}_2(x) := \inf \big\{ g(x) \mid g \text{ is concave, increasing, and dominates } f_2 \big\}.$$

Then $\widetilde{f}_2$ is concave, increasing, and dominates f_2. In fact, $\widetilde{f}_2$ is the smallest function with these properties. We have

$$\begin{aligned}
\int f_1\, d\mu_1 + \int f_2\, d\mu_2 &\le \int f_1\, d\mu_1 + \int \widetilde{f}_2\, d\mu_2 \\
&\le \int (f_1 + \widetilde{f}_2)\, d\mu_1 \\
&\le \sup_{x \in \mathbb{R}^d} \big(f_1(x) + \widetilde{f}_2(x) \big) =: r_0.
\end{aligned}$$

We will establish the condition in Theorem 2.88 for our set Λ by showing that for $r < r_0$ we have

$$r < \sup_{\lambda \in \Lambda} \int \big(f_1(x) + f_2(y) \big)\, \lambda(dx, dy).$$

To this end, let for $z \in \mathbb{R}^d$

$$\Lambda_z := \Big\{ \nu \in \mathcal{M}_1^{\psi}(\mathbb{R}^d) \;\Big|\; \int x\, \nu(dx) \le z \Big\}$$

and

$$g_2(z) := \sup \Big\{ \int f_2 \, d\nu \;\Big|\; \nu \in \Lambda_z \Big\}.$$

Then g_2 is increasing and $g_2(z) \ge f_2(z)$, because $\delta_z \in \Lambda_z$. Moreover, if $\nu_1 \in \Lambda_{z_1}$ and $\nu_2 \in \Lambda_{z_2}$, then

$$\alpha\nu_1 + (1-\alpha)\nu_2 \in \Lambda_{\alpha z_1 + (1-\alpha) z_2}$$

for $\alpha \in [0,1]$. Therefore, g_2 is concave, and we conclude that $g_2 \ge \widetilde{f}_2$ (recall that $\widetilde{f}_2$ is the smallest increasing and concave function dominating f_2). Hence, $r < f_1(z) + g_2(z)$ for some $z \in \mathbb{R}^d$, i.e., there exists some $\nu \in \Lambda_z$ such that the product measure $\lambda := \delta_z \otimes \nu$ satisfies

$$r < f_1(z) + \int f_2 \, d\nu = \int \big(f_1(x) + f_2(y) \big)\, \lambda(dx, dy).$$

But $\lambda = \delta_z \otimes \nu \in \Lambda$.

(b) ⇒ (c): This follows as in the proof of the implication (f) ⇒ (g) of Theorem 2.58 by using regular conditional distributions.

(c) ⇒ (a): As in the proof of (g) ⇒ (a) of Theorem 2.58, this follows by an application of Jensen's inequality. □

By the same arguments as for Corollary 2.62, we obtain the following result from Theorem 2.93.

Corollary 2.94. *Suppose μ_1 and μ_2 are Borel probability measures on $\mathbb{R}^d$ such that $\int |x| \mu_i(dx) < \infty$, for $i = 1, 2$. Then the following conditions are equivalent.*

(a) $\int f \, d\mu_1 \ge \int f \, d\mu_2$ *for all concave functions f on $\mathbb{R}^d$.*

(b) *There exists a probability space $(\Omega, \mathcal{F}, P)$ with random variables X_1 and X_2 having distributions μ_1 and μ_2, respectively, such that*

$$E[\, X_2 \mid X_1 \,] = X_1 \quad P\text{-a.s.}$$

(c) *There exists a kernel $Q(x, dy)$ on $\mathbb{R}^d$ such that*

$$\int y \, Q(x, dy) = x \quad \text{for all } x \in \mathbb{R}^d$$

(i.e., Q is a mean-preserving spread*) and such that $\mu_2 = \mu_1 Q$.*

We conclude this section with a generalization of Theorem 2.70. Let S be a Polish space which is endowed with a preference order $\succ$. We will assume that $\succ$ is continuous in the sense of Definition 2.8. A function on S will be called *increasing* if it is increasing with respect to $\succ$.

Theorem 2.95. *For two Borel probability measures μ_1 and μ_2 on S, the following conditions are equivalent.*

(a) $\int f\,d\mu_1 \geq \int f\,d\mu_2$ *for all bounded, increasing, and measurable functions f on S.*

(b) *There exists a probability space $(\Omega, \mathscr{F}, P)$ with random variables X_1 and X_2 having distributions μ_1 and μ_2, respectively, such that $X_1 \succeq X_2$ P-a.s.*

(c) *There exists a kernel Q on S such that $\mu_2 = \mu_1 Q$ and*

$$Q\big(x, \{\, y \mid x \succeq y \,\}\big) = 1 \quad \textit{for all } x \in S.$$

Proof. (a)$\Rightarrow$(b): We will apply Theorem 2.88 with the gauge function $\psi \equiv 1$, so that $\mathcal{M}_1^\psi(S)$ is just the space $\mathcal{M}_1(S)$ of all Borel probability measures on S with the usual weak topology. Then $\overline{\psi} \equiv 2$ which is equivalent to taking $\overline{\psi} := 1$. Let

$$M := \{\, (x, y) \in S \times S \mid x \succeq y \,\}.$$

This set M is closed in $S \times S$ by Proposition 2.11. Hence, the portmanteau theorem in the form of Theorem A.38 implies that the convex set

$$\Lambda := \{\, \lambda \in \mathcal{M}_1(S \times S) \mid \lambda(M) = 1 \,\}$$

is closed in $\mathcal{M}_1(S \times S)$. For $f_2 \in C_b(S)$, let

$$\widetilde{f}_2(x) := \sup\{\, f_2(y) \mid x \succeq y \,\}.$$

Then $\widetilde{f}_2$ is bounded, increasing, and dominates f_2. Therefore, if $f_1 \in C_b(S)$,

$$\begin{aligned}
\int f_1\,d\mu_1 + \int f_2\,d\mu_2 &\leq \int f_1\,d\mu_1 + \int \widetilde{f}_2\,d\mu_2 \\
&\leq \int (f_1 + \widetilde{f}_2)\,d\mu_1 \\
&\leq \sup_{x \in S} \big(f_1(x) + \widetilde{f}_2(x)\big) \\
&= \sup_{x \succeq y} \big(f_1(x) + f_2(y)\big).
\end{aligned}$$

If $x \succeq y$, then the product measure $\lambda := \delta_x \otimes \delta_y$ is contained in Λ, and so

$$\sup_{x \succeq y} \big(f_1(x) + f_2(y)\big) = \sup_{\lambda \in \Lambda} \int \big(f_1(x) + f_2(y)\big)\,\lambda(dx, dy).$$

Hence, all assumptions of Theorem 2.88 are satisfied, and we conclude that there exists a probability measure $P \in \Lambda$ with marginals μ_1 and μ_2. Taking $\Omega := S \times S$ and X_i as the projection on the i^{th} coordinate finishes the proof of (a) $\Rightarrow$(b).

(b) $\Rightarrow$(c) follows as in the proof of Theorem 2.58 by using regular conditional distributions.

(c) $\Rightarrow$(a) is proved as the corresponding implication of Theorem 2.70. □

Chapter 3

Optimality and equilibrium

Consider an investor whose preferences can be expressed in terms of expected utility. In Section 3.1, we discuss the problem of constructing a portfolio which maximizes the expected utility of the resulting payoff. The existence of an optimal solution is equivalent to the absence of arbitrage opportunities. This leads to an alternative proof of the "fundamental theorem of asset pricing", and to a specific choice of an equivalent martingale measure defined in terms of marginal utility. Section 3.2 contains a detailed case study describing the interplay between exponential utility and relative entropy. In Section 3.3, the optimization problem is formulated for general contingent claims. Typically, optimal profiles will be non-linear functions of a given market portfolio, and this is one source of the demand for financial derivatives. Section 3.4 introduces the idea of market equilibrium. Prices of risky assets will no longer be given in advance; they will be derived as equilibrium prices in a microeconomic setting, where different agents demand contingent claims in accordance with their preferences and with their budget constraints.

3.1 Portfolio optimization and the absence of arbitrage

Let us consider the one-period market model of Section 1.1 in which $d+1$ assets are priced at time 0 and at time 1. Prices at time 0 are given by the price system

$$\overline{\pi} = (\pi^0, \pi) = (\pi^0, \pi^1, \dots, \pi^d) \in \mathbb{R}_+^{d+1},$$

prices at time 1 are modeled by the price vector

$$\overline{S} = (S^0, S) = (S^0, S^1, \dots, S^d)$$

consisting of non-negative random variables S^i defined on some probability space $(\Omega, \mathcal{F}, P)$. The 0^{th} asset models a riskless bond, and so we assume that

$$\pi^0 = 1 \quad \text{and} \quad S^0 \equiv 1 + r$$

for some constant $r > -1$. At time $t = 0$, an investor chooses a portfolio

$$\overline{\xi} = (\xi^0, \xi) = (\xi^0, \xi^1, \dots, \xi^d) \in \mathbb{R}^{d+1}$$

where ξ^i represents the amount of shares of the i^{th} asset. Such a portfolio $\overline{\xi}$ requires an initial investment $\overline{\pi} \cdot \overline{\xi}$ and yields at time 1 the random payoff $\overline{\xi} \cdot \overline{S}$.

Consider a risk-averse economic agent whose preferences are described in terms of a utility function $\widetilde{u}$, and who wishes to invest a given amount w into the financial market. Recall from Definition 2.37 that a real-valued function $\widetilde{u}$ is called a utility function if it is continuous, strictly increasing, and strictly concave. A rational choice of the investor's portfolio $\overline{\xi} = (\xi^0, \xi)$ will be based on the expected utility

$$E\big[\,\widetilde{u}(\overline{\xi} \cdot \overline{S})\,\big] \tag{3.1}$$

of the payoff $\overline{\xi} \cdot \overline{S}$ at time 1, where the portfolio $\overline{\xi}$ satisfies the *budget constraint*

$$\overline{\pi} \cdot \overline{\xi} \le w. \tag{3.2}$$

Thus, the problem is to maximize the expected utility (3.1) among all portfolios $\overline{\xi} \in \mathbb{R}^{d+1}$ which satisfy the budget constraint (3.2). Here we make the implicit assumption that the payoff $\overline{\xi} \cdot \overline{S}$ is P-a.s. contained in the domain of definition of the utility function $\widetilde{u}$.

In a first step, we remove the constraint (3.2) by considering instead of (3.1) the expected utility of the *discounted net gain*

$$\frac{\overline{\xi} \cdot \overline{S}}{1+r} - \overline{\pi} \cdot \overline{\xi} = \xi \cdot Y$$

earned by a portfolio $\overline{\xi} = (\xi^0, \xi)$. Here Y is the d-dimensional random vector with components

$$Y^i = \frac{S^i}{1+r} - \pi^i, \quad i = 1, \dots, d.$$

For any portfolio $\overline{\xi}$ with $\overline{\pi} \cdot \overline{\xi} < w$, adding the risk-free investment $w - \overline{\pi} \cdot \overline{\xi}$ would lead to the strictly better portfolio $(\xi^0 + w - \overline{\pi} \cdot \overline{\xi}, \xi)$. Thus, we can focus on portfolios $\overline{\xi}$ which satisfy $\overline{\pi} \cdot \overline{\xi} = w$, and then the payoff is an affine function of the discounted net gain:

$$\overline{\xi} \cdot \overline{S} = (1+r)(\xi \cdot Y + w).$$

Moreover, for any $\xi \in \mathbb{R}^d$ there exists a unique numéraire component $\xi^0 \in \mathbb{R}$ such that the portfolio $\overline{\xi} := (\xi^0, \xi)$ satisfies $\overline{\pi} \cdot \overline{\xi} = w$.

Let u denote the following transformation of our original utility function $\widetilde{u}$:

$$u(y) := \widetilde{u}\big((1+r)(y+w)\big).$$

Note that u is again a utility function, and that CARA and (shifted) HARA utility functions are transformed into utility functions in the same class.

Clearly, the original constrained utility maximization problem is equivalent to the unconstrained problem of maximizing the expected utility $E[\,u(\xi \cdot Y)\,]$ among all $\xi \in \mathbb{R}^d$ such that $\xi \cdot Y$ is contained in the domain D of u.

Assumption 3.1. *We assume one of the following two cases:*

(a) $D = \mathbb{R}$. *In this case, we will admit all portfolios* $\xi \in \mathbb{R}^d$, *but we assume that* u *is bounded from above.*

(b) $D = [a, \infty)$ *for some* $a < 0$. *In this case, we only consider portfolios which satisfy the constraint*

$$\xi \cdot Y \geq a \quad P\text{-a.s.},$$

and we assume that the expected utility generated by such portfolios is finite, i.e.,

$$E[\, u(\xi \cdot Y)\,] < \infty \quad \text{for all } \xi \in \mathbb{R}^d \text{ with } \xi \cdot Y \geq a \ P\text{-a.s.}$$

Remark 3.2. Part (a) of this assumption is clearly satisfied in the case of an exponential utility function $u(x) = 1 - e^{-\alpha x}$. Domains of the form $D = [a, \infty)$ appear, for example, in the case of (shifted) HARA utility functions $u(x) = \log(x - b)$ for $b < a$ and $u(x) = \frac{1}{\gamma}(x - c)^\gamma$ for $c \leq a$ and $0 < \gamma < 1$. The integrability assumption in (b) holds if $E[\,|Y|\,] < \infty$, because any concave function is bounded above by an affine function. ◇

In order to simplify notations, let us denote by

$$\mathcal{S}(D) := \{\, \xi \in \mathbb{R}^d \mid \xi \cdot Y \in D \ P\text{-a.s.}\,\}$$

the set of admissible portfolios for D. Clearly, $\mathcal{S}(D) = \mathbb{R}^d$ if $D = \mathbb{R}$. Our aim is to find some $\xi^* \in \mathcal{S}(D)$ which is optimal in the sense that it maximizes the expected utility $E[\, u(\xi \cdot Y)\,]$ among all $\xi \in \mathcal{S}(D)$. In this case, ξ^* will be an optimal investment strategy into the risky assets. Complementing ξ^* with a suitable numéraire component ξ^0 yields a portfolio $\overline{\xi}^* = (\xi^0, \xi^*)$ which maximizes the expected utility $E[\, \widetilde{u}(\overline{\xi} \cdot \overline{S})\,]$ under the budget constraint $\overline{\pi} \cdot \overline{\xi} = w$. Our first result in this section will relate the existence of such an optimal portfolio to the absence of arbitrage opportunities.

Theorem 3.3. *Suppose that the utility function* $u : D \to \mathbb{R}$ *satisfies Assumption* 3.1. *Then there exists a maximizer of the expected utility*

$$E[\, u(\xi \cdot Y)\,], \quad \xi \in \mathcal{S}(D),$$

if and only if the market model is arbitrage-free. Moreover, there exists at most one maximizer if the market model is non-redundant in the sense of Definition 1.13.

Proof. The uniqueness part of the assertion follows immediately from the strict concavity of the function $\xi \mapsto E[\, u(\xi \cdot Y)\,]$ for non-redundant market models. As to existence, we may assume without loss of generality that our model is non-redundant. If the non-redundance condition (1.8) does not hold, then we define a linear space $N \subset \mathbb{R}^d$ by

$$N := \{\, \eta \in \mathbb{R}^d \mid \eta \cdot Y = 0 \ P\text{-a.s.}\,\}.$$

Clearly, Y takes P-a.s. values in the orthogonal complement $N^{\perp}$ of N. Moreover, the no-arbitrage condition (1.3) holds for all $\xi \in \mathbb{R}^d$ if and only if it is satisfied for all $\xi \in N^{\perp}$. By identifying $N^{\perp}$ with some $\mathbb{R}^n$, we arrive at a situation in which the non-redundance condition (1.8) is satisfied and where we may apply our result for non-redundant market models.

If the model admits arbitrage opportunities, then a maximizer ξ^* of the expected utility $E[\,u(\xi \cdot Y)\,]$ cannot exist: Adding to ξ^* some non-zero $\eta \in \mathbb{R}^d$ for which $\eta \cdot Y \geq 0$ P-a.s., which exists by Lemma 1.3, would yield a contradiction to the optimality of ξ^*, because then

$$E[\,u(\xi^* \cdot Y)\,] < E[\,u((\xi^* + \eta) \cdot Y)\,].$$

From now on, we assume that the market model is arbitrage-free. Let us first consider the case in which $D = [a, \infty)$ for some $a \in (-\infty, 0)$. Then $\mathcal{S}(D)$ is compact. In order to prove this claim, suppose by way of contradiction that (ξ_n) is a diverging sequence in $\mathcal{S}(D)$. By choosing a subsequence if necessary, we may assume that $\eta_n := \xi_n/|\xi_n|$ converges to some unit vector $\eta \in \mathbb{R}^d$. Clearly,

$$\eta \cdot Y = \lim_{n\uparrow\infty} \frac{\xi_n \cdot Y}{|\xi_n|} \geq \lim_{n\uparrow\infty} \frac{a}{|\xi_n|} = 0 \quad P\text{-a.s.},$$

and so non-redundance implies that $\overline{\eta} := (-\pi \cdot \eta, \eta)$ is an arbitrage-opportunity.

In the next step, we show that our assumptions guarantee the continuity of the function

$$\mathcal{S}(D) \ni \xi \mapsto E[\,u(\xi \cdot Y)\,],$$

which, in view of the compactness of $\mathcal{S}(D)$, will imply the existence of a maximizer of the expected utility. To this end, it suffices to construct an integrable random variable which dominates $u(\xi \cdot Y)$ for all $\xi \in \mathcal{S}(D)$. Define $\eta \in \mathbb{R}^d$ by

$$\eta^i := 0 \vee \max_{\xi \in \mathcal{S}(D)} \xi^i < \infty.$$

Then, $\eta \cdot S \geq \xi \cdot S$ for $\xi \in \mathcal{S}(D)$, and hence

$$\xi \cdot Y = \frac{\xi \cdot S}{1+r} - \pi \cdot \xi \leq \frac{\eta \cdot S}{1+r} - 0 \wedge \min_{\xi' \in \mathcal{S}(D)} \pi \cdot \xi'.$$

Note that $\eta \cdot Y$ is bounded below by $-\pi \cdot \eta$ and that there exists some $\alpha \in (0, 1]$ such that $\alpha\pi \cdot \eta < |a|$. Hence $\alpha\eta \in \mathcal{S}(D)$, and our assumptions imply $E[\,u(\alpha\eta \cdot Y)\,] < \infty$. Applying Lemma 3.4 below first with $b := \alpha\pi \cdot \eta$ and then with $b := -0 \wedge \min_{\xi' \in \mathcal{S}(D)} \pi \cdot \xi'$ shows that

$$E\left[\,u\left(\frac{\eta \cdot S}{1+r} - 0 \wedge \min_{\xi' \in \mathcal{S}(D)} \pi \cdot \xi'\right)\right] < \infty.$$

This concludes the proof of the theorem in case $D = [a, \infty)$.

Let us now turn to the case of a utility function on $D = \mathbb{R}$ which is bounded from above. We will reduce the assertion to a general existence criterion for minimizers of lower semicontinuous convex functions on $\mathbb{R}^d$, given in Lemma 3.5 below. It will be applied to the convex function $h(\xi) := -E[\,u(\xi \cdot Y)\,]$. We must show that h is lower semicontinuous. Take a sequence $(\xi_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ converging to some ξ. By part (a) of Assumption 3.1, the random variables $-u(\xi_n \cdot Y)$ are uniformly bounded below, and so we may apply Fatou's lemma:

$$\liminf_{n\uparrow\infty} h(\xi_n) = \liminf_{n\uparrow\infty} E[\,-u(\xi_n \cdot Y)\,] \;\geq E[\,-u(\xi \cdot Y)\,] = h(\xi).$$

Thus, h is lower semicontinuous.

By our non-redundance assumption, h is strictly convex and admits at most one minimizer. We claim that the absence of arbitrage opportunities is equivalent to the following condition:

$$\lim_{\alpha\uparrow\infty} h(\alpha\,\xi) = +\infty \quad \text{for all non-zero } \xi \in \mathbb{R}^d. \tag{3.3}$$

This is just the condition (3.4) required in Lemma 3.5. It follows from (1.3) and (1.8) that a non-redundant market model is arbitrage-free if and only if each non-zero $\xi \in \mathbb{R}^d$ satisfies $P[\xi \cdot Y < 0] > 0$. Since the utility function u is strictly increasing and concave, the set $\{\xi \cdot Y < 0\}$ can be described as

$$\{\,\xi \cdot Y < 0\,\} = \{\,\lim_{\alpha\uparrow\infty} u(\alpha\,\xi \cdot Y) = -\infty\,\} \quad \text{for } \xi \in \mathbb{R}^d.$$

The probability of the right-hand set is strictly positive if and only if

$$\lim_{\alpha\uparrow\infty} E[\,u(\alpha\xi \cdot Y)\,] = -\infty,$$

because u is bounded from above. This observation proves that the absence of arbitrage opportunities is equivalent to the condition (3.3) and completes the proof. □

Lemma 3.4. *If $D = [a, \infty)$, $b < |a|$, $0 < \alpha \leq 1$, and X is a non-negative random variable, then*

$$E[\,u(\alpha X - b)\,] < \infty \quad \Longrightarrow \quad E[\,u(X)\,] < \infty.$$

Proof. As in (A.1) in the proof of Proposition A.4, we obtain that

$$\frac{u(X) - u(0)}{X - 0} \leq \frac{u(\alpha X) - u(0)}{\alpha X - 0} \leq \frac{u(\alpha X - b) - u(-b)}{\alpha X - b - (-b)}.$$

Multiplying by αX shows that $u(X)$ can be dominated by a multiple of $u(\alpha X - b)$ plus some constant. □

Lemma 3.5. *Suppose $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex and lower semicontinuous function with $h(0) < \infty$. Then h attains its infimum provided that*

$$\lim_{\alpha \uparrow \infty} h(\alpha\,\xi) = +\infty \quad \text{for all non-zero } \xi \in \mathbb{R}^d. \tag{3.4}$$

Moreover, if h is strictly convex on $\{h < \infty\}$, then also the converse implication holds: the existence of a minimizer implies (3.4).

Proof. First suppose that (3.4) holds. We will show below that the "level sets" $\{ x \mid h(x) \le c \}$ of h are bounded – hence compact – for $c > \inf h$. Once the compactness of the level sets is established, it follows that the set

$$\{ x \in \mathbb{R}^d \mid h(x) = \inf h \} = \bigcap_{c > \inf h} \{ x \in \mathbb{R}^d \mid h(x) \le c \}$$

of minimizers of h is non-empty as an intersection of decreasing and non-empty compact sets.

Suppose $c > \inf h$ is such that the level set $\{h \le c\}$ is not compact, and take a sequence (x_n) in $\{h \le c\}$ such that $|x_n| \to \infty$. By passing to a subsequence if necessary, we may assume that $x_n/|x_n|$ converges to some non-zero ξ. For any $\alpha > 0$,

$$\begin{aligned} h(\alpha\xi) &\le \liminf_{n\uparrow\infty} h\left(\alpha \frac{x_n}{|x_n|}\right) = \liminf_{n\uparrow\infty} h\left(\frac{\alpha}{|x_n|}\, x_n + \left(1 - \frac{\alpha}{|x_n|}\right) 0\right) \\ &\le \liminf_{n\uparrow\infty} \left(\frac{\alpha}{|x_n|}\, c + \left(1 - \frac{\alpha}{|x_n|}\right) h(0)\right) \\ &= h(0). \end{aligned}$$

Thus, we arrive at a contradiction to condition (3.4). This completes the proof of the existence of a minimizer under assumption (3.4).

In order to prove the converse implication, suppose that the strictly convex function h has a minimizer x^* but that there exists a non-zero $\xi \in \mathbb{R}^d$ violating (3.4), i.e., there exists a sequence $(\alpha_n)_{n\in\mathbb{N}}$ and some $c < \infty$ such that $\alpha_n \uparrow \infty$ but $h(\alpha_n\xi) \le c$ for all n. Let

$$x_n := \lambda_n x^* + (1 - \lambda_n)\alpha_n\xi$$

where λ_n is such that $|x^* - x_n| = 1$, which is possible for all large enough n. By the compactness of the Euclidean unit sphere centered in x^*, we may assume that x_n converges to some x. Then necessarily $|x - x^*| = 1$. As $\alpha_n\xi$ diverges, we must have that $\lambda_n \to 1$. By using our assumption that $h(\alpha_n\xi)$ is bounded, we obtain

$$h(x) \le \liminf_{n\uparrow\infty} h(x_n) \le \lim_{n\uparrow\infty} \big(\lambda_n h(x^*) + (1 - \lambda_n) h(\alpha_n\xi)\big) = h(x^*).$$

Hence, x is another minimizer of h besides x^*, contradicting the strict convexity of h. Thus, (3.4) must hold if the strictly convex function h takes on its infimum. □

Remark 3.6. Note that the proof of Theorem 3.3 under Assumption 3.1 (a) did not use the fact that the components of Y are bounded from below. The result remains true for arbitrary Y. ◇

We turn now to a characterization of the solution ξ^* of our utility maximization problem for continuously differentiable utility functions.

Proposition 3.7. *Let u be a continuously differentiable utility function on D such that $E[\,u(\xi \cdot Y)\,]$ is finite for all $\xi \in \mathcal{S}(D)$. Suppose that ξ^* is a solution of the utility maximization problem, and that one of the following two sets of conditions is satisfied:*

- *u is defined on $D = \mathbb{R}$ and is bounded from above.*
- *u is defined on $D = [a, \infty)$, and ξ^* is an interior point of $\mathcal{S}(D)$.*

Then

$$u'(\xi^* \cdot Y)\,|Y| \in \mathcal{L}^1(P),$$

and the following "first-order condition" holds:

$$E[\,u'(\xi^* \cdot Y)\,Y\,] = 0. \tag{3.5}$$

Proof. For $\xi \in \mathcal{S}(D)$ and $\varepsilon \in (0, 1]$ let $\xi_\varepsilon := \varepsilon\xi + (1-\varepsilon)\xi^*$, and define

$$\Delta_\varepsilon := \frac{u(\xi_\varepsilon \cdot Y) - u(\xi^* \cdot Y)}{\varepsilon}.$$

The concavity of u implies that $\Delta_\varepsilon \geq \Delta_\delta$ for $\varepsilon \leq \delta$, and so

$$\Delta_\varepsilon \nearrow u'(\xi^* \cdot Y)\,(\xi - \xi^*) \cdot Y \quad \text{as } \varepsilon \downarrow 0.$$

Note that our assumptions imply that $u(\xi \cdot Y) \in \mathcal{L}^1(P)$ for all $\xi \in \mathcal{S}(D)$. In particular, we have $\Delta_1 \in \mathcal{L}^1(P)$, so that monotone convergence and the optimality of ξ^* yield that

$$0 \geq E[\,\Delta_\varepsilon\,] \nearrow E[\,u'(\xi^* \cdot Y)\,(\xi - \xi^*) \cdot Y\,] \quad \text{as } \varepsilon \downarrow 0. \tag{3.6}$$

In particular, the expectation on the right-hand side of (3.6) is finite.

Both sets of assumptions imply that ξ^* is an interior point of $\mathcal{S}(D)$. Hence, we deduce from (3.6) by letting $\eta := \xi - \xi^*$ that

$$E[\,u'(\xi^* \cdot Y)\,\eta \cdot Y\,] \leq 0$$

for all η in a small ball centered in the origin of $\mathbb{R}^d$. Replacing η by $-\eta$ shows that the expectation must vanish. □

Remark 3.8. Let us comment on the assumption that the optimal ξ^* is an interior point of $\mathcal{S}(D)$:

(a) If the non-redundance condition (1.8) is not satisfied, then either each or none of the solutions to the utility maximization problem is contained in the interior of $\mathcal{S}(D)$. This can be seen by using the reduction argument given at the beginning of the proof of Theorem 3.3.

(b) Note that $\xi \cdot Y$ is bounded below by $-\pi \cdot \xi$ in case ξ has only non-negative components. Thus, the interior of $\mathcal{S}(D)$ is always non-empty.

(c) As shown by the following example, the optimal ξ^* need not be contained in the interior of $\mathcal{S}(D)$ and, in this case, the first-order condition (3.5) will generally fail. ◇

Example 3.9. Take $r = 0$, and let S^1 be integrable but unbounded. We choose $D = [a, \infty)$ with $a := -\pi^1$, and we assume that $P[\, S^1 \leq \varepsilon \,] > 0$ for all $\varepsilon > 0$. Then $\mathcal{S}(D) = [0, 1]$. If $0 < E[\, S^1 \,] < \pi^1$ then Example 2.43 shows that the optimal investment is given by $\xi^* = 0$, and so ξ^* lies in the boundary of $\mathcal{S}(D)$. Thus, if u is sufficiently smooth,

$$E[\, u'(\xi^* \cdot Y)\, Y \,] = u'(0)\big(E[\, S^1 \,] - \pi^1\big) < 0.$$

The intuitive reason for this failure of the first-order condition is that taking a short position in the asset would be optimal as soon as $E[\, S^1 \,] < \pi^1$. This choice, however, is ruled out by the constraint $\xi \in \mathcal{S}(D)$. ◇

Proposition 3.7 yields a formula for the density of a particular equivalent risk-neutral measure. Recall that P^* is risk-neutral if and only if $E^*[\, Y \,] = 0$.

Corollary 3.10. *Suppose that the market model is arbitrage-free and that the assumptions of Proposition* 3.7 *are satisfied for a utility function $u : D \to \mathbb{R}$ and an associated maximizer ξ^* of the expected utility $E[u(\xi \cdot Y)]$. Then*

$$\frac{dP^*}{dP} = \frac{u'(\xi^* \cdot Y)}{E[u'(\xi^* \cdot Y)]} \tag{3.7}$$

defines an equivalent risk neutral measure.

Proof. Proposition 3.7 states that $u'(\xi^* \cdot Y)Y$ is integrable with respect to P and that its expectation vanishes. Hence, we may conclude that P^* is an equivalent risk-neutral measure if we can show that P^* is well-defined by (3.7), i.e., if $u'(\xi^* \cdot Y) \in \mathcal{L}^1(P)$. Let

$$c := \sup\{\, u'(x) \mid x \in D \text{ and } |x| \leq |\xi^*|\,\} \leq \begin{cases} u'(a) & \text{for } D = [a, \infty), \\ u'(-|\xi^*|) & \text{for } D = \mathbb{R}, \end{cases}$$

which is finite by our assumption that u is continuously differentiable on all of D. Thus,

$$0 \leq u'(\xi^* \cdot Y) \leq c + u'(\xi^* \cdot Y)|Y| \cdot \mathrm{I}_{\{|Y| \geq 1\}},$$

and the right-hand side has a finite expectation. □

Remark 3.11. Corollary 3.10 yields an independent and constructive proof of the "fundamental theorem of asset pricing" in the form of Theorem 1.6: Suppose that the model is arbitrage-free. If Y is P-a.s. bounded, then so is $u(\xi^* \cdot Y)$, and the measure P^* of (3.7) is an equivalent risk-neutral measure with a bounded density dP^*/dP. If Y is unbounded, then we may consider the bounded random vector

$$\widetilde{Y} := \frac{Y}{1+|Y|},$$

which also satisfies the no-arbitrage condition (1.3). Let $\widetilde{\xi}^*$ be a maximizer of the expected utility $E[\, u(\xi \cdot \widetilde{Y})\,]$. Then an equivalent risk-neutral measure P^* is defined through the bounded density

$$\frac{dP^*}{dP} := c \cdot \frac{u'(\widetilde{\xi}^* \cdot \widetilde{Y})}{1+|Y|},$$

where c is an appropriate normalizing constant. ◇

Example 3.12. Consider the exponential utility function

$$u(x) = 1 - e^{-\alpha x}$$

with constant absolute risk aversion $\alpha > 0$. The requirement that $E[\, u(\xi \cdot Y)\,]$ is finite is equivalent to the condition

$$E[\, e^{\xi \cdot Y}\,] < \infty \quad \text{for all } \xi \in \mathbb{R}^d.$$

If ξ^* is a maximizer of the expected utility, then the density of the equivalent risk neutral measure P^* in (3.7) takes the particular form

$$\frac{dP^*}{dP} = \frac{e^{-\alpha \xi^* \cdot Y}}{E[\, e^{-\alpha \xi^* \cdot Y}\,]}.$$

In fact, P^* is independent of α since ξ^* maximizes the expected utility $1 - E[\, e^{-\alpha \xi \cdot Y}\,]$ if and only if $\lambda^* := -\alpha \xi^*$ is a minimizer of the *moment generating function*

$$Z(\lambda) := E[\, e^{\lambda \cdot Y}\,], \quad \lambda \in \mathbb{R}^d,$$

of Y. In Corollary 3.25 below, the measure P^* will be characterized by the fact that it minimizes the *relative entropy* with respect to P among the risk-neutral measures in $\mathcal{P}$; see Definition 3.20 below. ◇

3.2 Exponential utility and relative entropy

In this section we give a more detailed study of the problem of portfolio optimization with respect to a CARA utility function

$$u(x) = 1 - e^{-\alpha x}$$

for $\alpha > 0$. As in the previous Section 3.1, the problem is to maximize the expected utility

$$E[\,u(\xi \cdot Y)\,]$$

of the discounted net gain $\xi \cdot Y$ earned by an investment into risky assets. The key assumption for this problem is that

$$E[\,u(\xi \cdot Y)\,] > -\infty \quad \text{for all } \xi \in \mathbb{R}^d. \tag{3.8}$$

Recall from Example 3.12 that the maximization of $E[\,u(\xi \cdot Y)\,]$ is reduced to the minimization of the moment generating function

$$Z(\lambda) := E[\,e^{\lambda \cdot Y}\,], \quad \lambda \in \mathbb{R}^d,$$

which does not depend on the risk aversion α. The key assumption (3.8) is equivalent to the condition that

$$Z(\lambda) < \infty \quad \text{for all } \lambda \in \mathbb{R}^d. \tag{3.9}$$

Throughout this section, we will always assume that (3.9) holds. But we will not need the assumption that Y is bounded from below (which in our financial market model follows from assuming that asset prices are non-negative); all results remain true for general random vectors Y; see also Remarks 1.7 and 3.6.

Lemma 3.13. *The condition* (3.9) *is equivalent to*

$$E[\,e^{\alpha|Y|}\,] < \infty \quad \textit{for all } \alpha > 0.$$

Proof. Clearly, the condition in the statement of the lemma implies (3.9). To prove the converse assertion, take a constant $c > 0$ such that $|x| \le c \sum_{i=1}^d |x^i|$ for $x \in \mathbb{R}^d$. By Hölder's inequality,

$$E[\,e^{\alpha|Y|}\,] \le E\Big[\,\exp\Big(\alpha c \sum_{i=1}^d |Y^i|\Big)\Big] \le \prod_{i=1}^d E[\,e^{\alpha c d|Y^i|}\,]^{1/d}.$$

In order to show that the i^{th} factor on the right is finite, take $\lambda \in \mathbb{R}^d$ such that $\lambda^i = \alpha c d$ and $\lambda^j = 0$ for $j \neq i$. With this choice,

$$E[\,e^{\alpha c d|Y^i|}\,] \le E[\,e^{\lambda \cdot Y}\,] + E[\,e^{-\lambda \cdot Y}\,],$$

which is finite by (3.9). □

Definition 3.14. The *exponential family* of P with respect to Y is the set of measures

$$\{\, P_\lambda \mid \lambda \in \mathbb{R}^d \,\}$$

defined via

$$\frac{dP_\lambda}{dP} = \frac{e^{\lambda \cdot Y}}{Z(\lambda)}.$$

Example 3.15. Suppose that the risky asset S^1 has under P a Poisson distribution with parameter $\alpha > 0$, i.e., S^1 takes values in $\{0, 1, \dots\}$ and satisfies

$$P[\,S^1 = k\,] = e^{-\alpha}\frac{\alpha^k}{k!}, \quad k = 0, 1, \dots .$$

Then (3.9) is satisfied for $Y := S^1 - \pi^1$, and S^1 has under P_λ a Poisson distribution with parameter $e^\lambda \alpha$. Hence, the exponential family of P generates the family of all Poisson distributions. ◇

Example 3.16. Let Y have a standard normal distribution $N(0, 1)$. Then (3.9) is satisfied, and the distribution of Y under P_λ is equal to the normal distribution $N(\lambda, 1)$ with mean λ and variance 1. ◇

Remark 3.17. Two parameters λ and λ' in $\mathbb{R}^d$ determine the same element in the exponential family of P if and only if $(\lambda - \lambda') \cdot Y = 0$ P-almost surely. It follows that the mapping

$$\lambda \longmapsto P_\lambda$$

is injective provided that the non-redundance condition holds in the form

$$\xi \cdot Y = 0 \ P\text{-a.s.} \quad \Longrightarrow \quad \xi = 0. \tag{3.10}$$

◇

In the sequel, we will be interested in the barycenters of the members of the exponential family of P with respect to Y. We denote

$$m(\lambda) := E_\lambda[\,Y\,] = \frac{1}{Z(\lambda)} E[\,Y e^{\lambda \cdot Y}\,], \quad \lambda \in \mathbb{R}^d.$$

The next lemma shows that $m(\lambda)$ can be obtained as the gradient of the logarithmic moment generating function.

Lemma 3.18. *Z is a smooth function on $\mathbb{R}^d$, and the gradient of* $\log Z$ *at λ is the expectation of Y under P_λ:*

$$(\nabla \log Z)(\lambda) = E_\lambda[\,Y\,] = m(\lambda).$$

Moreover, the Hessian of $\log Z$ *at λ equals the covariance matrix* $(\mathrm{cov}_{P_\lambda}(Y^i, Y^j))_{i,j}$ *of Y under the measure P_λ:*

$$\frac{\partial^2}{\partial \lambda^i \, \partial \lambda^j} \log Z(\lambda) = \mathrm{cov}_{P_\lambda}(Y^i, Y^j) = E_\lambda[\,Y^i Y^j\,] - E_\lambda[\,Y^i\,] E_\lambda[\,Y^j\,].$$

In particular, $\log Z$ *is convex.*

Proof. Observe that

$$\left|\frac{\partial e^{\lambda\cdot x}}{\partial \lambda^i}\right| = |x^i|\, e^{\lambda\cdot x} \le \exp\big[(1+|\lambda|)\cdot|x|\big].$$

Hence, Lemma 3.13 and Lebesgue's dominated convergence theorem justify the interchanging of differentiation and integration (see the "differentiation lemma" in [20], §16, for details). □

The following corollary summarizes the results we have obtained so far. Recall from Section 1.5 the notion of the convex hull $\Gamma(\nu)$ of the support of a measure ν on $\mathbb{R}^d$ and the definition of the relative interior $\operatorname{ri} C$ of a convex set C.

Corollary 3.19. *Denote by $\mu := P \circ Y^{-1}$ the distribution of Y under P. Then the function*

$$\lambda \longmapsto \lambda \cdot m_0 - \log Z(\lambda)$$

takes on its maximum if and only if m_0 is contained in the relative interior of the convex hull of the support of μ, i.e., if and only if

$$m_0 \in \operatorname{ri}\Gamma(\mu).$$

In this case, any maximizer λ^* satisfies

$$m_0 = m(\lambda^*) = E_{\lambda^*}[\,Y\,].$$

In particular, the set $\{m(\lambda) \mid \lambda \in \mathbb{R}^d\}$ coincides with $\operatorname{ri}\Gamma(\mu)$. Moreover, if the non-redundance condition (3.10) *holds, then there exists at most one maximizer λ^*.*

Proof. Taking $\widetilde{Y} := Y - m_0$ reduces the problem to the situation where $m_0 = 0$. Applying Theorem 3.3 with the utility function $u(z) = 1 - e^{-z}$ shows that the existence of a maximizer λ^* of $-\log Z$ is equivalent to the absence of arbitrage opportunities. Corollary 3.10 states that $m(\lambda^*) = 0$ and that 0 belongs to $M_b(\mu)$, where $M_b(\mu)$ was defined in Lemma 1.42. An application of Theorem 1.48 completes the proof. □

It will turn out that the maximization problem of the previous corollary is closely related to the following concept.

Definition 3.20. The *relative entropy* of a probability measure Q with respect to P is defined as

$$H(Q|P) := \begin{cases} E\left[\dfrac{dQ}{dP}\log\dfrac{dQ}{dP}\right] & \text{if } Q \ll P, \\ +\infty & \text{otherwise.} \end{cases}$$

Remark 3.21. Jensen's inequality applied to the strictly convex function $h(x) = x\log x$ yields

$$H(Q|P) = E\left[h\left(\frac{dQ}{dP}\right)\right] \ge h(1) = 0, \tag{3.11}$$

with equality if and only if $Q = P$. ◇

Example 3.22. Let Ω be a finite set and $\mathcal{F}$ be its power set. Every probability Q on $(\Omega, \mathcal{F})$ is absolutely continuous with respect to the uniform distribution P. Let us denote $Q(\omega) := Q[\{\omega\}]$. Clearly,

$$H(Q|P) = \sum_{\omega\in\Omega} Q(\omega) \log \frac{Q(\omega)}{P(\omega)} = \sum_{\omega\in\Omega} Q(\omega) \log Q(\omega) + \log |\Omega|.$$

The quantity

$$H(Q) := -\sum_{\omega\in\Omega} Q(\omega) \log Q(\omega)$$

is usually called the *entropy* of Q. Observe that $H(P) = \log |\Omega|$, so that

$$H(Q|P) = H(P) - H(Q).$$

Since the left-hand side is non-negative by (3.11), the uniform distribution P has maximal entropy among all probability distributions on $(\Omega, \mathcal{F})$. ◇

Example 3.23. Let $\mu = N(m, \sigma^2)$ denote the normal distribution with mean m and variance σ^2 on $\mathbb{R}$. Then, for $\widetilde{\mu} = N(\widetilde{m}, \widetilde{\sigma}^2)$

$$\frac{d\widetilde{\mu}}{d\mu}(x) = \frac{\sigma}{\widetilde{\sigma}} \exp\left[-\frac{(x-\widetilde{m})^2}{2\widetilde{\sigma}^2} + \frac{(x-m)^2}{2\sigma^2} \right],$$

and hence

$$H(\widetilde{\mu}|\mu) = \frac{1}{2}\left(\log \frac{\sigma^2}{\widetilde{\sigma}^2} - 1 + \frac{\sigma^2}{\widetilde{\sigma}^2} \right) + \frac{1}{2}\left(\frac{m-\widetilde{m}}{\sigma} \right)^2. \qquad ◇$$

The following result shows that P_λ is the unique minimizer of the relative entropy $H(Q|P)$ among all probability measures Q with $E_Q[Y] = E_\lambda[Y]$.

Theorem 3.24. *Let $m_0 := m(P_{\lambda_0})$ for some given $\lambda_0 \in \mathbb{R}^d$. Then, for any probability measure Q on $(\Omega, \mathcal{F})$ such that $E_Q[Y] = m_0$,*

$$H(Q|P) \geq H(P_{\lambda_0}|P) = \lambda_0 \cdot m_0 - \log Z(\lambda_0),$$

and equality holds if and only if $Q = P_{\lambda_0}$. Moreover, λ_0 maximizes the function

$$\lambda \cdot m_0 - \log Z(\lambda)$$

over all $\lambda \in \mathbb{R}^d$.

Proof. Let Q be a probability measure on $(\Omega, \mathcal{F})$ such that $E_Q[Y] = m_0$. We show first that for all $\lambda \in \mathbb{R}^d$

$$H(Q|P) = H(Q|P_\lambda) + \lambda \cdot m_0 - \log Z(\lambda). \tag{3.12}$$

To this end, note that both sides of (3.12) are infinite if $Q \not\ll P$. Otherwise

$$\frac{dQ}{dP} = \frac{dQ}{dP_\lambda} \cdot \frac{dP_\lambda}{dP} = \frac{dQ}{dP_\lambda} \cdot \frac{e^{\lambda \cdot Y}}{Z(\lambda)},$$

and taking logarithms and integrating with respect to Q yields (3.12).

Since $H(Q|P_\lambda) \geq 0$ according to (3.11), we get from (3.12) that

$$H(Q|P) \geq \lambda \cdot m_0 - \log Z(\lambda) \tag{3.13}$$

for all $\lambda \in \mathbb{R}^d$ and all measures Q such that $E_Q[\,Y\,] = m_0$. Moreover, equality holds in (3.13) if and only if $H(Q|P_\lambda) = 0$, which is equivalent to $Q = P_\lambda$. In this case, λ must be such that $m(\lambda) = m_0$. In particular, for any such λ

$$H(P_\lambda|P) = \lambda \cdot m_0 - \log Z(\lambda).$$

Thus, λ_0 maximizes the right-hand side of (3.13), and P_{λ_0} minimizes the relative entropy on the set

$$M_0 := \{\, Q \mid E_Q[\,Y\,] = m_0 \,\}.$$

But the relative entropy $H(Q|P)$ is a strictly convex functional of Q, and so it can have at most one minimizer in the convex set M_0. Thus, any λ with $m(\lambda) = m_0$ induces the same measure P_{λ_0}. □

Taking $m_0 = 0$ in the preceding theorem yields a special equivalent risk-neutral measure in our financial market model, namely the *entropy-minimizing risk neutral measure*. Sometimes it is also called the *Esscher transform* of P. Recall our assumption (3.9).

Corollary 3.25. *Suppose the market model is arbitrage-free. Then there exists a unique equivalent risk-neutral measure $P^* \in \mathcal{P}$ which minimizes the relative entropy $H(\hat{P}|P)$ over all $\hat{P} \in \mathcal{P}$. The density of P^* is of the form*

$$\frac{dP^*}{dP} = \frac{e^{\lambda^* \cdot Y}}{E[\,e^{\lambda^* \cdot Y}\,]},$$

where λ^* denotes a minimizer of the moment generating function $E[\,e^{\lambda \cdot Y}\,]$ of Y.

Proof. This follows immediately from Corollary 3.19 and Theorem 3.24. □

By combining Theorem 3.24 with Remark 3.17, we obtain the following corollary. It clarifies the question of uniqueness in the representation of points in the relative interior of $\Gamma(P \circ Y^{-1})$ as barycenters of the exponential family.

Corollary 3.26. *If the non-redundance condition* (3.10) *holds, then*

$$\lambda \longmapsto m(\lambda)$$

is a bijective mapping from $\mathbb{R}^d$ to $\operatorname{ri}\Gamma(P \circ Y^{-1})$.

Remark 3.27. It follows from Corollary 3.19 and Theorem 3.24 that for all $m \in \operatorname{ri}\Gamma(P \circ Y^{-1})$

$$\min_{E_Q[\,Y\,]=m} H(Q|P) = \max_{\lambda \in \mathbb{R}^d}[\lambda \cdot m - \log Z(\lambda)]. \tag{3.14}$$

Here, the right-hand side is the *Fenchel–Legendre transform* of the convex function $\log Z$ evaluated at $m \in \mathbb{R}^d$. ◇

The following theorem shows that the variational principle (3.14) remains true for all $m \in \mathbb{R}^d$, if we replace "min" and "max" by "inf" and "sup".

Theorem 3.28. *For $m \in \mathbb{R}^d$*

$$\inf_{E_Q[\,Y\,]=m} H(Q|P) = \sup_{\lambda \in \mathbb{R}^d}[\lambda \cdot m - \log Z(\lambda)].$$

The proof of this theorem relies on the following two general lemmas.

Lemma 3.29. *For any probability measure Q,*

$$\begin{aligned} H(Q|P) &= \sup_{Z \in \mathcal{L}^\infty(\Omega,\mathcal{F},P)} \big(E_Q[\,Z\,] - \log E[\,e^Z\,]\big) \\ &= \sup\big\{E_Q[\,Z\,] - \log E[\,e^Z\,] \mid e^Z \in \mathcal{L}^1(P)\big\}. \end{aligned} \tag{3.15}$$

The second supremum is attained by $Z := \log \frac{dQ}{dP}$ if $Q \ll P$.

Proof. We first show $\geq$ in (3.15). To this end, we may assume that $H(Q|P) < \infty$. For Z with $e^Z \in \mathcal{L}^1(P)$ let P^Z be defined by

$$\frac{dP^Z}{dP} = \frac{e^Z}{E[\,e^Z\,]}.$$

Then P^Z is equivalent to P and

$$\log \frac{dQ}{dP} = \log \frac{dQ}{dP^Z} + \log \frac{dP^Z}{dP}.$$

Integrating with respect to Q gives

$$H(Q|P) = H(Q|P^Z) + E_Q[\,Z\,] - \log E[\,e^Z\,].$$

Since $H(Q|P^Z) \geq 0$ by (3.11), we have proved that $H(Q|P)$ is larger than or equal to both suprema on the right of (3.15).

To prove the reverse inequality, consider first the case $Q \not\ll P$. Take $Z_n := n\mathrm{I}_A$ where A is such that $Q[\,A\,] > 0$ and $P[\,A\,] = 0$. Then, as $n \uparrow \infty$,

$$E_Q[\,Z_n\,] - \log E[\,e^{Z_n}\,] = n \cdot Q[\,A\,] \longrightarrow \infty = H(Q|P).$$

Now suppose that $Q \ll P$ with density $\varphi = dQ/dP$. Then $Z := \log\varphi$ satisfies $e^Z \in \mathcal{L}^1(P)$ and

$$H(Q|P) = E_Q[\,Z\,] - \log E[\,e^Z\,].$$

For the first identity we need an approximation argument. Let $Z_n = (-n)\vee(\log\varphi)\wedge n$. We split the expectation $E[\,e^{Z_n}\,]$ according to the two sets $\{\varphi \geq 1\}$ and $\{\varphi < 1\}$. Using monotone convergence for the first integral and dominated convergence for the second yields

$$E[\,e^{Z_n}\,] \longrightarrow E[\,e^{\log\varphi}\,] = 1.$$

Since $x\log x \geq -1/e$, we have $\varphi Z_n \geq -1/e$ uniformly in n, and Fatou's lemma yields

$$\liminf_{n\uparrow\infty} E_Q[\,Z_n\,] = \liminf_{n\uparrow\infty} E[\,\varphi Z_n\,] \geq E[\,\varphi\log\varphi\,] = H(Q|P).$$

Putting both facts together shows

$$\liminf_{n\uparrow\infty}\big(E_Q[\,Z_n\,] - \log E[\,e^{Z_n}\,]\big) \geq H(Q|P),$$

and the inequality $\leq$ in (3.15) follows. □

Remark 3.30. The preceding lemma shows that the relative entropy is monotone with respect to an increase of the underlying σ-algebra: Let P and Q be two probability measures on a measurable space $(\Omega, \mathcal{F})$, and denote by $H(Q|P)$ their relative entropy. Suppose that $\mathcal{F}_0$ is a σ-field such that $\mathcal{F}_0 \subset \mathcal{F}$ and denote by $H_0(Q|P)$ the relative entropy of Q with respect to P considered as probability measures on the smaller space $(\Omega, \mathcal{F}_0)$. Then the relation $\mathcal{L}^\infty(\Omega, \mathcal{F}_0, P) \subset \mathcal{L}^\infty(\Omega, \mathcal{F}, P)$ implies

$$H_0(Q|P) \leq H(Q|P);$$

in general this inequality is strict. ◇

Lemma 3.31. *For all $\alpha \geq 0$, the set*

$$\Phi_\alpha := \big\{\,\varphi \in L^1(\Omega, \mathcal{F}, P) \mid \varphi \geq 0,\; E[\,\varphi\,] = 1,\; E[\,\varphi\log\varphi\,] \leq \alpha\,\big\}$$

is weakly sequentially compact in $L^1(\Omega, \mathcal{F}, P)$.

Proof. Let $L^p := L^p(\Omega, \mathcal{F}, P)$. The set of all P-densities,

$$\mathcal{D} := \big\{\,\varphi \in L^1 \mid \varphi \geq 0,\; E[\,\varphi\,] = 1\,\big\},$$

is clearly convex and closed in L^1. Hence, this set is also weakly closed in L^1 by Theorem A.59. Moreover, Lemma 3.29 states that for $\varphi \in \mathcal{D}$

$$E[\,\varphi\log\varphi\,] = \sup_{Z\in L^\infty}\big(E[\,Z\varphi\,] - \log E[\,e^Z\,]\big).$$

In particular,

$$\varphi \mapsto E[\varphi \log \varphi]$$

is a weakly lower semicontinuous functional on $\mathcal{D}$, and so Φ_α is weakly closed. In addition, Φ_α is bounded in L^1 and uniformly integrable, due to the criterion of de la Vallée Poussin; see, e.g., Lemma 3 in §6 of Chapter II of [187]. Applying the Dunford–Pettis theorem and the Eberlein–Šmulian theorem as stated in Appendix A.7 concludes the proof. □

Proof of Theorem 3.28. In view of Theorem 3.24 and inequality (3.13) (whose proof extends to all $m \in \mathbb{R}^d$), it remains to prove that

$$\inf_{E_Q[Y]=m} H(Q|P) \le \sup_{\lambda \in \mathbb{R}^d} [\lambda \cdot m - \log Z(\lambda)] \tag{3.16}$$

for those m which do not belong to $\operatorname{ri}\Gamma(\mu)$, where $\mu := P \circ Y^{-1}$. The right-hand side of (3.16) is just the Fenchel–Legendre transform at m of the convex function $\log Z$ and, thus, denoted $(\log Z)^*(m)$.

First, we consider the case in which m is not contained in the closure $\overline{\Gamma}(\mu)$ of the convex hull of the support of μ. Proposition A.1, the separating hyperplane theorem, yields some $\xi \in \mathbb{R}^d$ such that

$$\xi \cdot m > \sup\{\, \xi \cdot x \mid x \in \overline{\Gamma}(\mu)\,\} \ge \sup\{\, \xi \cdot x \mid x \in \operatorname{supp}\mu\,\}.$$

By taking $\lambda_n := n\xi$, it follows that

$$\lambda_n \cdot m - \log Z(\lambda_n) \ge n\big(\xi \cdot m - \sup_{y \in \operatorname{supp}\mu} \xi \cdot y\big) \longrightarrow +\infty \quad \text{as } n \uparrow \infty.$$

Hence, the right-hand side of (3.16) is infinite if $m \notin \overline{\Gamma}(\mu)$.

It remains to prove (3.16) for $m \in \overline{\Gamma}(\mu) \setminus \operatorname{ri}\Gamma(\mu)$ with $(\log Z)^*(m) < \infty$. Recall from (1.22) that $\operatorname{ri}\Gamma(\mu) = \operatorname{ri}\overline{\Gamma}(\mu)$. Pick some $m_1 \in \operatorname{ri}\Gamma(\mu)$ and let

$$m_n := \frac{1}{n} m_1 + \left(1 - \frac{1}{n}\right) m.$$

Then $m_n \in \operatorname{ri}\Gamma(\mu)$ by (1.21). By the convexity of $(\log Z)^*$, we have

$$\begin{aligned}\limsup_{n\uparrow\infty} (\log Z)^*(m_n) &\le \limsup_{n\uparrow\infty} \left(\frac{1}{n}(\log Z)^*(m_1) + \frac{n-1}{n}(\log Z)^*(m)\right) \\ &= (\log Z)^*(m).\end{aligned} \tag{3.17}$$

We also know that to each m_n there corresponds a $\lambda_n \in \mathbb{R}^d$ such that

$$m_n = E_{\lambda_n}[Y] \quad \text{and} \quad H(P_{\lambda_n}|P) = (\log Z)^*(m_n). \tag{3.18}$$

From (3.17) and (3.18) we conclude that

$$\limsup_{n\uparrow\infty} H(P_{\lambda_n}|P) = \limsup_{n\uparrow\infty}(\log Z)^*(m_n) \le (\log Z)^*(m) < \infty.$$

In particular, $H(P_{\lambda_n}|P)$ is uniformly bounded in n, and Lemma 3.31 implies that – after passing to a suitable subsequence if necessary – the densities dP_{λ_n}/dP converge weakly in $L^1(\Omega, \mathcal{F}, P)$ to a density φ. Let $dP_\infty = \varphi\, dP$. By the weak lower semicontinuity of

$$\frac{dQ}{dP} \longmapsto H(Q|P),$$

which follows from Lemma 3.29, we may conclude that $H(P_\infty|P) \le (\log Z)^*(m)$.

The theorem will be proved once we can show that $E_\infty[\,Y\,] = m$. To this end, let $\gamma := \sup_n (\log Z)^*(m_n)$, which is a finite non-negative number by (3.17). Taking

$$Z := \alpha\, \mathrm{I}_{\{|Y|\ge c\}}|Y|$$

on the right-hand side of (3.15) yields

$$\gamma \ge \alpha E_{\lambda_n}\big[\,|Y|\cdot \mathrm{I}_{\{|Y|\ge c\}}\,\big] - \log E\big[\exp\big(\alpha|Y|\mathrm{I}_{\{|Y|\ge c\}}\big)\big] \quad \text{for all } n \le \infty.$$

Note that the rightmost expectation is finite due to condition (3.9) and Lemma 3.13. By taking α large so that $\gamma/\alpha < \varepsilon/2$ for some given $\varepsilon > 0$, and by choosing c such that

$$\log E\big[\,\exp\big(\alpha|Y|\mathrm{I}_{\{|Y|\ge c\}}\big)\big] < \frac{\alpha\varepsilon}{2},$$

we obtain that

$$\sup_{n\le\infty} E_{\lambda_n}\big[\,|Y|\cdot \mathrm{I}_{\{|Y|\ge c\}}\,\big] \le \varepsilon.$$

But

$$E_{\lambda_n}\big[\,|Y|\cdot \mathrm{I}_{\{|Y|<c\}}\,\big] \longrightarrow E_\infty\big[\,|Y|\cdot \mathrm{I}_{\{|Y|<c\}}\,\big]$$

by the weak convergence of $dP_{\lambda_n}/dP \to dP_\infty/dP$, and so taking $\varepsilon \downarrow 0$ yields

$$m = \lim_{n\uparrow\infty} E_{\lambda_n}[\,Y\,] = E_\infty[\,Y\,],$$

as desired. □

3.3 Optimal contingent claims

In this section we study the problem of maximizing the expected utility

$$E[\,u(X)\,]$$

under a given budget constraint in a broader context. The random variables X will vary in a general convex class $\mathcal{X} \subset L^0(\Omega, F, P)$ of *admissible* payoff profiles. In the setting of our financial market model, this will allow us to explain the demand for *non-linear* payoff profiles provided by financial derivatives.

In order to formulate the budget constraint in this general context, we introduce a linear pricing rule of the form

$$\Phi(X) = E^*[\, X \,] = E[\, \varphi X \,]$$

where P^* is a probability measure on $(\Omega, \mathcal{F})$, which is equivalent to P with density φ. For a given initial wealth $w \in \mathbb{R}$, the corresponding budget set is defined as

$$\mathcal{B} := \{\, X \in \mathcal{X} \cap L^1(P^*) \mid E^*[\, X \,] \le w \,\}. \tag{3.19}$$

Our optimization problem can now be stated as follows:

$$\textit{Maximize } E[\, u(X) \,] \textit{ among all } X \in \mathcal{B}. \tag{3.20}$$

Note, however, that we will need some extra conditions which guarantee that the expectations $E[\, u(X) \,]$ make sense and are bounded from above.

Remark 3.32. In general, our optimization problem would not be well posed without the assumption $P^* \approx P$. Note first that it should be rephrased in terms of a class $\mathcal{X}$ of measurable functions on $(\Omega, \mathcal{F})$ since we can no longer pass to equivalence classes with respect to P. If P is not absolutely continuous with respect to P^* then there exists $A \in \mathcal{F}$ such that $P[\, A \,] > 0$ and $P^*[\, A \,] = 0$. For $X \in \mathcal{L}^1(P^*)$ and $c > 0$, the random variable $\widetilde{X} := X + c\,\mathrm{I}_A$ would satisfy $E^*[\, \widetilde{X} \,] = E^*[\, X \,]$ and $E[\, u(\widetilde{X}) \,] > E[\, u(X) \,]$. Similarly, if $P^*[\, A \,] > 0$ and $P[\, A \,] = 0$ then

$$\hat{X} := X + c - \frac{c}{P^*[\, A \,]}\,\mathrm{I}_A$$

would have the same price as X but higher expected utility. In particular, the expectations in (3.20) would be unbounded in both cases if $\mathcal{X}$ is the class of *all* measurable functions on $(\Omega, \mathcal{F})$ and if the function u is not bounded from above. ◇

Remark 3.33. If a solution X^* with $E[\, u(X^*) \,] < \infty$ exists then it is *unique*, since $\mathcal{B}$ is convex and u is strictly concave. Moreover, if $\mathcal{X} = L^0(\Omega, \mathcal{F}, P)$ or $\mathcal{X} = L^0_+(\Omega, \mathcal{F}, P)$ then X^* satisfies

$$E^*[\, X^* \,] = w$$

since $E^*[\, X^* \,] < w$ would imply that $X := X^* + w - E^*[\, X^*]$ is a strictly better choice, due to the strict monotonicity of u. ◇

Let us first consider the unrestricted case $\mathcal{X} = L^0(\Omega, F, P)$ where any finite random variable on (Ω, F, P) is admissible. The following heuristic argument identifies a candidate X^* for the maximization of the expected utility. Suppose that a solution X^* exists. For any $X \in L^\infty(P)$ and any $\lambda \in \mathbb{R}$,

$$X_\lambda := X^* + \lambda\big(X - E^*[\, X \,]\big)$$

satisfies the budget constraint $E^*[\, X_\lambda \,] = w$. A formal computation yields

$$\begin{aligned} 0 &= \frac{d}{d\lambda}\Big|_{\lambda=0} E[\, u(X_\lambda)\,] \\ &= E\big[\, u'(X^*)(X - E^*[\, X\,])\big] \\ &= E[\, u'(X^*)X\,] - E\big[\, X E[\, u'(X^*)\,]\,\varphi\big] \\ &= E\big[\, X(u'(X^*) - c\,\varphi)\,\big] \end{aligned}$$

where $c := E[\, u'(X^*)\,]$. The identity

$$E[\, X\, u'(X^*)\,] = c\, E[\, X\,\varphi\,]$$

for all bounded measurable X implies $u'(X^*) = c\,\varphi$ P-almost surely. Thus, if we denote by

$$I := (u')^{-1}$$

the inverse function of the strictly decreasing function u', then X^* should be of the form

$$X^* = I(c\,\varphi).$$

We will now formulate a set of assumptions on our utility function u which guarantee that $X^* := I(c\,\varphi)$ is indeed a maximizer of the expected utility, as suggested by the preceding argument.

Theorem 3.34. *Suppose $u : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable utility function which is bounded from above, and whose derivative satisfies*

$$\lim_{x\downarrow-\infty} u'(x) = +\infty. \tag{3.21}$$

Assume moreover that $c > 0$ is a constant such that

$$X^* := I(c\,\varphi) \in L^1(P^*).$$

Then X^* is the unique maximizer of the expected utility $E[\, u(X)\,]$ among all those $X \in L^1(P^*)$ for which $E^*[\, X\,] \le E^*[\, X^*\,]$. In particular, X^* solves our optimization problem (3.20) *for $\mathcal{X} = L^0(\Omega, \mathcal{F}, P)$ if c can be chosen such that $E^*[\, X^*\,] = w$.*

Proof. Uniqueness follows from Remark 3.33. Since u is bounded from above, its derivative satisfies

$$\lim_{x\uparrow\infty} u'(x) = 0,$$

in addition to (3.21). Hence, $(0, \infty)$ is contained in the range of u', and it follows that $I(c\,\varphi)$ is P-a.s. well-defined for all $c > 0$.

To show the optimality of $X^* = I(c\,\varphi)$, note that the concavity of u implies that for any $X \in L^1(P^*)$

$$u(X) \le u(X^*) + u'(X^*)(X - X^*) = u(X^*) + c\,\varphi(X - X^*).$$

Taking expectations with respect to P yields

$$E[\,u(X)\,] \le E[\,u(X^*)\,] + c\,E^*[\,X - X^*\,].$$

Hence, X^* is indeed a maximizer in the class $\{\,X \in L^1(P^*) \mid E^*[\,X\,] \le E^*[\,X^*\,]\}$. □

Example 3.35. Let $u(x) = 1 - e^{-\alpha x}$ be an exponential utility function with constant absolute risk aversion $\alpha > 0$. In this case,

$$I(y) = -\frac{1}{\alpha}\log\frac{y}{\alpha}.$$

It follows that

$$\begin{aligned} E^*[\,I(c\,\varphi)\,] &= -\frac{1}{\alpha}\log\frac{c}{\alpha} - \frac{1}{\alpha}\cdot E[\,\varphi\log\varphi\,] \\ &= -\frac{1}{\alpha}\log\frac{c}{\alpha} - \frac{1}{\alpha}\cdot H(P^*|P), \end{aligned}$$

where $H(P^*|P)$ denotes the relative entropy of P^* with respect to P; see Definition 3.20. Hence, the utility maximization problem can be solved for any $w \in \mathbb{R}$ if and only if the relative entropy $H(P^*|P)$ is finite. In this case, the optimal profile is given by

$$X^* = -\frac{1}{\alpha}\log\varphi + w + \frac{1}{\alpha}H(P^*|P),$$

and the maximal value of expected utility is

$$E[\,u(X^*)\,] = 1 - \exp\big(-\alpha w - H(P^*|P)\big),$$

corresponding to the certainty equivalent

$$w + \frac{1}{\alpha}H(P^*|P).$$

Let us now return to the financial market model considered in Section 3.1, and let P^* be the entropy-minimizing risk-neutral measure constructed in Corollary 3.25. The density of P^* is of the form

$$\varphi = \frac{e^{-\alpha\xi^*\cdot Y}}{E[\,e^{-\alpha\xi^*\cdot Y}\,]},$$

where $\xi^* \in \mathbb{R}^d$ denotes a maximizer of the expected utility $E[\,u(\xi\cdot Y)\,]$; see Example 3.12. In this case, the optimal profile takes the form

$$X^* = \xi^*\cdot Y + w = \frac{\overline{\xi}^*\cdot\overline{S}}{1+r},$$

i.e., X^* is the discounted payoff of the portfolio $\overline{\xi}^* = (\xi^0, \xi^*)$, where $\xi^0 = w - \xi^*\cdot\pi$ is determined by the budget constraint $\overline{\xi}\cdot\overline{\pi} = w$. Thus, the optimal profile is given by a *linear* profile in the given primary assets $S^0, \dots, S^d$: No derivatives are needed at this point. ◇

In most situations it will be natural to restrict the discussion to payoff profiles which are non-negative. For the rest of this section we will make this restriction, and so the utility function u may be defined only on $[0, \infty)$. In several applications we will also use an upper bound given by an $\mathcal{F}$-measurable random variable $W : \Omega \longrightarrow [0, \infty]$. We include the case $W \equiv +\infty$ and define the convex class of admissible payoff profiles as

$$\mathcal{X} := \{ X \in L^0(P) \mid 0 \le X \le W \ P\text{-a.s.} \}$$

Thus, our goal is to maximize the expected utility $E[\, u(X)\,]$ among all $X \in \mathcal{B}$ where the budget set $\mathcal{B}$ is defined in terms of $\mathcal{X}$ and P^* as in (3.19), i.e.,

$$\mathcal{B} = \Big\{ X \in L^1(P^*) \mid 0 \le X \le W \ P\text{-a.s. and } E^*[\, X\,] \le w \Big\}.$$

We first formulate a general existence result:

Proposition 3.36. *Let u be any utility function on $[0, \infty)$, and suppose that W is P-a.s. finite and satisfies $E[\, u(W)\,] < \infty$. Then there exists a unique $X^* \in \mathcal{B}$ which maximizes the expected utility $E[\, u(X)\,]$ among all $X \in \mathcal{B}$.*

Proof. Take a sequence (X_n) in $\mathcal{B}$ with $E^*[\, X_n\,] \le w$ and such that $E[\, u(X_n)\,]$ converges to the supremum of the expected utility. Since $\sup_n |X_n| \le W < \infty$ P-almost surely, we obtain from Lemma 1.69 a sequence

$$\widetilde{X}_n \in \operatorname{conv}\{ X_n, X_{n+1}, \dots \}$$

of convex combinations which converge almost-surely to some $\widetilde{X}$. Clearly, every $\widetilde{X}_n$ is contained in $\mathcal{B}$. Fatou's lemma implies

$$E^*[\, \widetilde{X}\,] \le \liminf_{n \uparrow \infty} E^*[\, \widetilde{X}_n\,] \le w,$$

and so $\widetilde{X} \in \mathcal{B}$. Each $\widetilde{X}_n$ can be written as $\sum_{i=1}^m \alpha_i^n X_{n_i}$ for indices $n_i \ge n$ and coefficients $\alpha_i^n \ge 0$ summing up to 1. Hence,

$$u(\widetilde{X}_n) \ge \sum_{i=1}^m \alpha_i^n u(X_{n_i}),$$

and it follows that

$$E[\, u(\widetilde{X}_n)\,] \ge \inf_{m \ge n} E[\, u(X_m)\,].$$

By dominated convergence,

$$E[\, u(\widetilde{X})\,] = \lim_{n \uparrow \infty} E[\, u(\widetilde{X}_n)\,],$$

and the right-hand side is equal to the supremum of the expected utility. □

Remark 3.37. The argument used to prove the preceding proposition works just as well in the following general setting. Let $U : \mathscr{B} \to \mathbb{R}$ be a concave functional on a set $\mathscr{B}$ of random variables defined on a probability space $(\Omega, \mathscr{F}, P)$ and with values in $\mathbb{R}^n$. Assume that

- $\mathscr{B}$ is convex and closed under P-a.s. convergence,
- There exists a random variable $W \in L^0_+(\Omega, \mathscr{F}, P)$ with $|X^i| \le W < \infty$ P-a.s. for each $X = (X^1, \dots, X^n) \in \mathscr{B}$,
- $\sup_{X \in \mathscr{B}} U(X) < \infty$,
- U is upper semicontinuous with respect to P-a.s. convergence.

Then there exists an $X^* \in \mathscr{B}$ which maximizes U on $\mathscr{B}$, and X^* is unique if U is strictly concave. As a special case, this includes the utility functionals

$$U(X) = \inf_{Q \in \mathscr{Q}} E_Q[\, u(X)\,],$$

appearing in a robust Savage representation of preferences on n-dimensional asset profiles, where u is a utility function on $\mathbb{R}^n$ and $\mathscr{Q}$ is a set of probability measures equivalent to P; see Section 2.5. ◇

We turn now to a characterization of the optimal profile X^* in terms of the inverse of the derivative u' of u in case where u is continuously differentiable on $(0, \infty)$. Let

$$a := \lim_{x \uparrow \infty} u'(x) \ge 0 \quad \text{and} \quad b := u'(0+) = \lim_{x \downarrow 0} u'(x) \le +\infty.$$

We define

$$I^+ : (a, b) \longrightarrow (0, \infty).$$

as the continuous, bijective, and strictly decreasing inverse function of u' on (a, b), and we extend I^+ to the full half axis $[0, \infty]$ by setting

$$I^+(y) := \begin{cases} 0 & \text{for } y \ge b, \\ +\infty & \text{for } y \le a. \end{cases} \tag{3.22}$$

With this convention, $I^+ : [0, \infty] \to [0, \infty]$ is continuous.

Remark 3.38. If u is a utility function defined on all of $\mathbb{R}$, the function I^+ is the inverse of the restriction of u' to $[0, \infty)$. Thus, I^+ is simply the positive part of the function $I = (u')^{-1}$. For instance, in the case of an exponential utility function $u(x) = 1 - e^{-\alpha x}$, we have $a = 0$, $b = \alpha$, and

$$I^+(y) = \left(\frac{1}{\alpha} \log \frac{y}{\alpha} \right)^- = \big(I(y)\big)^+, \quad y \ge 0. \tag{3.23}$$

◇

Theorem 3.39. *Assume that* $X^* \in \mathcal{B}$ *is of the form*

$$X^* = I^+(c\,\varphi) \wedge W$$

for some constant $c > 0$ *such that* $E^*[\,X^*\,] = w$. *If* $E[\,u(X^*)\,] < \infty$ *then* X^* *is the unique maximizer of the expected utility* $E[\,u(X)\,]$ *among all* $X \in \mathcal{B}$.

Proof. In a first step, we consider the function

$$v(y, \omega) := \sup_{0 \le x \le W(\omega)} \big(u(x) - xy\big) \tag{3.24}$$

defined for $y \in \mathbb{R}$ and $\omega \in \Omega$. Clearly, for each ω with $W(\omega) < \infty$ the supremum above is attained in a unique point $x^*(y) \in [0, W(\omega)]$, which satisfies

$$\begin{aligned} x^*(y) = 0 \quad &\Longleftrightarrow \quad u'(x) < y \quad \text{for all } x \in \big(0, W(\omega)\big), \\ x^*(y) = W(\omega) \quad &\Longleftrightarrow \quad u'(x) > y \quad \text{for all } x \in \big(0, W(\omega)\big). \end{aligned}$$

Moreover, $y = u'(x^*(y))$ if $x^*(y)$ is an interior point of the interval $[0, W(\omega)]$. It follows that

$$x^*(y) = I^+(y) \wedge W(\omega),$$

or

$$X^* = x^*(c\,\varphi) \quad \text{on } \{W < \infty\}. \tag{3.25}$$

If $W(\omega) = +\infty$, then the supremum in (3.24) is not attained if and only if $u'(x) > y$ for all $x \in (0, \infty)$. By our convention (3.22), this holds if and only if $y \le a$ and hence $I^+(y) = +\infty$. But our assumptions on X^* imply that $I^+(c\,\varphi) < \infty$ P-a.s. on $\{W = \infty\}$, and hence that

$$X^* = x^*(c\,\varphi) \quad P\text{-a.s. on } \{W = \infty\}. \tag{3.26}$$

Putting (3.24), (3.25), and (3.26) together yields

$$u(X^*) - X^* c\,\varphi = v(c\varphi, \cdot) \quad P\text{-a.s.}$$

Applied to an arbitrary $X \in \mathcal{B}$, this shows that

$$u(X^*) - c\,\varphi X^* \ge u(X) - c\,\varphi X \quad P\text{-a.s.}$$

Taking expectations gives

$$E[\,u(X^*)\,] \ge E[\,u(X)\,] + c \cdot E^*[\,X^* - X\,] \ge E[\,u(X)\,].$$

Hence, X^* maximizes the expected utility on $\mathcal{B}$. Uniqueness follows from Remark 3.33. □

In the following examples, we study the application of the preceding theorem to CARA and HARA utility functions. For simplicity we consider only the case $W \equiv \infty$. The extension to a non-trivial bound W is straightforward.

Example 3.40. For an exponential utility function $u(x) = 1-e^{-\alpha x}$ we have by (3.23)

$$\varphi I^+(y\,\varphi) = \frac{1}{\alpha}\,\varphi\Big(\log\Big(\frac{y\,\varphi}{\alpha}\Big)\Big)^- = \frac{1}{y}\,h\Big(\frac{y\,\varphi}{\alpha}\Big),$$

where $h(x) = (x\log x)^-$. Since h is bounded by e^{-1}, it follows that $\varphi I^+(y\,\varphi)$ belongs to $\mathcal{L}^1(P)$ for all $y > 0$. Thus,

$$g(y) := E^*[\,I^+(y\,\varphi)\,] = \frac{1}{y}\,E\Big[\,h\Big(\frac{y\,\varphi}{\alpha}\Big)\Big]$$

decreases continuously from $+\infty$ to 0 as y increases from 0 to ∞, and there exists a unique c with $g(c) = w$. The corresponding profile

$$X^* := I^+(c\,\varphi)$$

maximizes the expected utility $E[\,u(X)\,]$ among all $X \geq 0$. Let us now return to the special situation of the financial market model of Section 3.1, and take P^* as the entropy-minimizing risk-neutral measure of Corollary 3.25. Then the optimal profile X^* takes the form

$$X^* = (\xi^*\cdot Y - K)^+,$$

where ξ^* is the maximizer of the expected utility $E[\,u(\xi\cdot Y)\,]$, and where K is given by

$$K = \frac{1}{\alpha}\log\frac{c}{\alpha} - \frac{1}{\alpha}\log E[\,e^{-\alpha\xi^*\cdot Y}\,] = \frac{1}{\alpha}\log\frac{c}{\alpha} + \frac{1}{\alpha}H(P^*|P).$$

Note that X^* is a *linear* combination of the primary assets only in the case where $\xi^*\cdot Y \geq K$ P-almost surely. In general, X^* is a *basket call option* on the attainable asset $w + (1+r)\xi^*\cdot Y \in \mathcal{V}$ with *strike price* $w + (1+r)K$. Thus, a demand for *derivatives* appears. ◇

Example 3.41. If u is a HARA utility function of index $\gamma \in [0,1)$ then $u'(x) = x^{\gamma-1}$, hence

$$I^+(y) = y^{-\frac{1}{1-\gamma}}$$

and

$$I^+(y\,\varphi) = y^{-\frac{1}{1-\gamma}}\cdot\varphi^{-\frac{1}{1-\gamma}}.$$

In the *logarithmic* case $\gamma = 0$, we assume that the relative entropy $H(P|P^*)$ of P with respect to P^* is finite. Then

$$X^* = \frac{w}{\varphi} = w\,\frac{dP}{dP^*}$$

is the unique maximizer, and the maximal value of expected utility is

$$E[\,\log X^*\,] = \log w + H(P|P^*).$$

If $\gamma \in (0,1)$ and

$$E\big[\,\varphi^{-\frac{\gamma}{1-\gamma}}\,\big] = E^*\big[\,\varphi^{-\frac{1}{1-\gamma}}\,\big] < \infty,$$

then the unique optimal profile is given by

$$X^* = w\,\Big(E\big[\,\varphi^{-\frac{\gamma}{1-\gamma}}\,\big]\Big)^{-1}\varphi^{-\frac{1}{1-\gamma}},$$

and the maximal value of expected utility is equal to

$$E[\,u(X^*)\,] = \frac{1}{\gamma}w^{\gamma}\Big(E\big[\,\varphi^{-\frac{\gamma}{1-\gamma}}\,\big]\Big)^{1-\gamma}. \qquad \diamond$$

The following corollary gives a simple condition on W which guarantees the existence of the maximizer X^* in Theorem 3.39.

Corollary 3.42. *If $E[\,u(W)\,] < \infty$ and if $0 < w < E^*[\,W\,] < \infty$, then there exists a unique constant $c > 0$ such that*

$$X^* = I^+(c\,\varphi) \wedge W$$

satisfies $E^*[\,X^*\,] = w$. In particular, X^* is the unique maximizer of the expected utility $E[\,u(X)\,]$ among all $X \in \mathcal{B}$.

Proof. For any $\beta \in (0,\infty)$,

$$y \mapsto I^+(y) \wedge \beta$$

is a continuous decreasing function with $\lim_{y\uparrow b} I^+(y) \wedge \beta = 0$ and $I^+(y) \wedge \beta = \beta$ for all $y \le u'(\beta)$. Hence, dominated convergence implies that the function

$$g(y) := E^*[\,I^+(y\,\varphi) \wedge W\,],$$

is continuous and decreasing with

$$\lim_{y\uparrow\infty} g(y) = 0 < w < E^*[\,W\,] = \lim_{y\downarrow 0} g(y).$$

Moreover, g is even strictly decreasing on $\{\,y \mid 0 < g(y) < E^*[\,W\,]\}$. Hence, there exists a unique c with $g(c) = w$, and Theorem 3.39 yields the optimality of the corresponding X^*. $\square$

Let us now extend the discussion to the case where preferences themselves are uncertain. This additional uncertainty can be modelled by incorporating the choice of a utility function into the description of possible scenarios; for an axiomatic discussion see [130]. More precisely, we assume that preferences are described by a measurable function u on $[0,\infty) \times \Omega$ such that $u(\cdot,\omega)$ is a utility function on $[0,\infty)$ which is continuously differentiable on $(0, \infty)$. For each $\omega \in \Omega$, the inverse of $u'(\cdot, \omega)$ is extended as above to a function

$$I^{+}(\cdot, \omega) : [0, \infty] \longrightarrow [0, \infty].$$

Using exactly the same arguments as above, we obtain the following extension of Corollary 3.42 to the case of *random preferences*:

Corollary 3.43. *If $E[\, u(W, \cdot)\,] < \infty$ and if $0 < w < E^*[\, W\,] < \infty$, then there exists a unique constant $c > 0$ such that*

$$X^*(\omega) := I^{+}\big(c\,\varphi(\omega), \omega\big) \wedge W(\omega)$$

is the unique maximizer of the expected utility

$$E[\, u(X, \cdot)\,] = \int u(X(\omega), \omega)\, P(d\omega)$$

among all $X \in \mathcal{B}$.

So far, we have discussed the structure of asset profiles which are optimal with respect to a fixed utility function u. Let us now introduce an optimization problem with respect to the uniform order $\succcurlyeq_{\text{uni}}$ as discussed in Section 2.4. The partial order $\succcurlyeq_{\text{uni}}$ can be viewed as a reflexive and transitive relation on the space of financial positions

$$\mathcal{X} := L^1_+(\Omega, \mathcal{F}, P)$$

by letting

$$\begin{aligned} X \succcurlyeq_{\text{uni}} Y :&\iff \mu_X \succcurlyeq_{\text{uni}} \mu_Y \\ &\iff E[\, u(X)\,] \geq E[\, u(Y)\,] \quad \text{for all utility functions } u, \end{aligned} \tag{3.27}$$

where μ_X and μ_Y denote the distributions of X and Y under P. Note that $X \succcurlyeq_{\text{uni}} Y \succcurlyeq_{\text{uni}} X$ if and only if X and Y have the same distribution; see Remark 2.59. Thus, the relation $\succcurlyeq_{\text{uni}}$ is antisymmetric on the level of distributions but not on the level of financial positions.

Let us now fix a position $X_0 \in \mathcal{X}$ such that $E^*[\, X_0\,] < \infty$, and let us try to minimize the cost among all positions $X \in \mathcal{X}$ which are uniformly at least as attractive as X_0:

$$\text{Minimize } E^*[\, X\,] \text{ among all } X \succcurlyeq_{\text{uni}} X_0.$$

In order to describe the minimal cost and the minimizing profile, let us denote by F_φ and F_{X_0} the distribution functions and by q_φ and q_{X_0} quantile functions of φ and X_0; see Appendix A.3.

Theorem 3.44. *For any $X \in \mathcal{X}$ such that $X \succcurlyeq_{\text{uni}} X_0$,*

$$E^*[\, X\,] \geq \int_0^1 q_\varphi(1-s)\, q_{X_0}(s)\, ds. \tag{3.28}$$

The lower bound is attained by $X^* = f(\varphi)$*, where* f *is the decreasing function on* $[0,\infty)$ *defined by*

$$f(x) := q_{X_0}\big(1 - F_\varphi(x)\big)$$

if x *is a continuity point of* F_φ*, and by*

$$f(x) := \frac{1}{F_\varphi(x) - F_\varphi(x-)} \int_{F_\varphi(x-)}^{F_\varphi(x)} q_{X_0}(1-t)\,dt$$

otherwise.

The proof will use the following lemma, which yields another characterization of the relation $\succcurlyeq_{\text{uni}}$.

Lemma 3.45. *For two probability measures* μ *and* ν *on* $\mathbb{R}$*, the following conditions are equivalent:*

(a) $\mu \succcurlyeq_{\text{uni}} \nu$.

(b) *For all decreasing functions* $h : (0,1) \to [0,\infty)$*,*

$$\int_0^1 h(t) q_\mu(t)\,dt \ge \int_0^1 h(t) q_\nu(t)\,dt, \tag{3.29}$$

where q_μ *and* q_ν *are quantile functions of* μ *and* ν*.*

(c) *The relation* (3.29) *holds for all bounded decreasing functions* $h : (0,1) \to [0,\infty)$*.*

Proof. The relation $\mu \succcurlyeq_{\text{uni}} \nu$ is equivalent to

$$\int_0^y q_\mu(t)\,dt \ge \int_0^y q_\nu(t)\,dt \quad \text{for all } y \in [0,1];$$

see Theorem 2.58. The implication (c) ⇒(a) thus follows by taking $h = I_{(0,t]}$. For the proof of (a) ⇒(b), we may assume without loss of generality that h is left-continuous. Then there exists a positive Radon measure η on $(0,1]$ such that $h(t) = \eta([t,1])$. Fubini's theorem yields

$$\begin{aligned}\int_0^1 h(t)\,q_\mu(t)\,dt &= \int_0^1 \int_0^y q_\mu(t)\,dt\,\eta(dy)\\ &\ge \int_0^1 \int_0^y q_\nu(t)\,dt\,\eta(dy)\\ &= \int_0^1 h(t)\,q_\nu(t)\,dt.\end{aligned}$$

□

Proof of Theorem 3.44. Using the first Hardy–Littlewood inequality in Theorem A.24, we see that

$$E^*[\,X\,] = E[\,X\varphi\,] \geq \int_0^1 q_\varphi(1-t)\,q_X(t)\,dt,$$

where q_X is a quantile function for X. Taking $h(t) := q_\varphi(1-t)$ and using Lemma 3.45 thus yields (3.28).

Let us now turn to the identification of the optimal profile. Note that the function f defined in the assertion satisfies

$$f(q_\varphi) = E_\lambda[\,g \mid q_\varphi\,] \tag{3.30}$$

where g is defined by $g(t) = q_{X_0}(1-t)$, and where $E_\lambda[\,\cdot \mid q_\varphi\,]$ denotes the conditional expectation with respect to q_φ under the Lebesgue measure λ on (0, 1). Let us show that $X^* = f(\varphi)$ satisfies $X^* \succcurlyeq_{\text{uni}} X_0$. Indeed, for any utility function u

$$\begin{aligned} E[\,u(X^*)\,] &= E\big[\,u\big(f(\varphi)\big)\,\big] = \int_0^1 u\big(f(q_\varphi)\big)\,dt \\ &\geq \int_0^1 u\big(q_{X_0}(1-t)\big)\,dt = \int_0^1 u\big(q_{X_0}(t)\big)\,dt \\ &= E[\,u(X_0)\,], \end{aligned}$$

where we have applied Lemma A.19 and Jensen's inequality for conditional expectations. Moreover, X^* attains the lower bound in (3.28):

$$\begin{aligned} E^*[\,X^*\,] &= E[\,f(\varphi)\,\varphi\,] = \int_0^1 f\big(q_\varphi(t)\big)\,q_\varphi(t)\,dt \\ &= \int_0^1 q_{X_0}(1-t)\,q_\varphi(t)\,dt = \int_0^1 q_{X_0}(t)\,q_\varphi(1-t)\,dt, \end{aligned}$$

due to (3.30). □

Remark 3.46. The solution X^* has the same expectation under P as X_0. Indeed, (3.30) shows that

$$E[\,X^*\,] = E[\,f(\varphi)\,] = \int_0^1 f(q_\varphi(t))\,dt = \int_0^1 q_{X_0}(1-t)\,dt = E[\,X_0\,]. \qquad \diamond$$

Remark 3.47. The lower bound in (3.28) may be viewed as a "reservation price" for X_0 in the following sense. Let X_0 be a financial position, and let $\mathcal{X}$ be any class of financial positions such that $X \in \mathcal{X}$ is available at price $\pi(X)$. For a given relation $\succeq$ on $\mathcal{X} \cup \{X_0\}$,

$$\pi_R(X_0) := \inf\{\,\pi(X) \mid X \in \mathcal{X},\ X \succeq X_0\,\}$$

is called the *reservation price* of X_0 with respect to $\mathcal{X}$, π, and $\succeq$.

If $\mathcal{X}$ is the space of constants with $\pi(c) = c$, and if the relation $\succeq$ is of von Neumann–Morgenstern type with some utility function u, then $\pi_R(X_0)$ reduces to the certainty equivalent of X_0 with respect to u; see (2.9).

In the context of the optimization problem (3.20), where

$$X \succeq X_0 \;:\Longleftrightarrow\; E[\,u(X)\,] \geq E[\,u(X_0)\,],$$

the reservation price is given by $E^*[\,X^*\,]$, where X^* is the utility maximizer in the budget set defined by $w := E^*[\,X_0\,]$.

In the context of the financial market model of Chapter 1, we can take $\mathcal{X}$ as the space $\mathcal{V}$ of attainable claims with

$$V \succeq X_0 \;:\Longleftrightarrow\; V \geq X_0 \;P\text{-a.s.}$$

and $\pi(V) = \overline{\xi} \cdot \overline{\pi}$ for $V = \overline{\xi} \cdot \overline{S}$. In this case, the reservation price $\pi_R(X_0)$ coincides with the upper bound $\pi_{\sup}(X_0)$ of the arbitrage-free prices for X_0; see Theorem 1.31. ◇

3.4 Microeconomic equilibrium

The aim of this section is to provide a brief introduction to the theory of market equilibrium. Prices of assets will no longer be given in advance. Instead, they will be derived from "first principles" in a microeconomic setting where different agents demand asset profiles in accordance with their preferences and with their budget constraints. These budget constraints are determined by a given price system. The role of equilibrium prices consists in adjusting the constraints in such a way that the resulting overall demand is matched by the overall supply of assets.

Consider a finite set $\mathcal{A}$ of *economic agents* and a convex set $\mathcal{X} \subset L^0(\Omega, \mathcal{F}, P)$ of *admissible claims*. At time $t = 0$, each agent $a \in \mathcal{A}$ has an initial endowment whose discounted payoff at time $t = 1$ is described by an admissible claim

$$W_a \in \mathcal{X}, \quad a \in \mathcal{A}.$$

The aggregated claim

$$W := \sum_{a \in \mathcal{A}} W_a$$

is also called the *market portfolio*. Agents may want to exchange their initial endowment W_a against some other admissible claim $X_a \in \mathcal{X}$. This could lead to a new allocation $(X_a)_{a \in \mathcal{A}}$ if the resulting total demand matches the overall supply:

Definition 3.48. A collection $(X_a)_{a \in \mathcal{A}} \subset \mathcal{X}$ is called a *feasible allocation* if it satisfies the *market clearing condition*

$$\sum_{a \in \mathcal{A}} X_a = W \quad P\text{-a.s.} \tag{3.31}$$

The budget constraints will be determined by a linear pricing rule of the form

$$\Phi(X) := E[\varphi X], \quad X \in \mathcal{X},$$

where φ is a *price density*, i.e., an integrable function on $(\Omega, \mathcal{F})$ such that $\varphi > 0$ P-a.s. and $E[|W_a|\varphi] < \infty$ for all $a \in \mathcal{A}$. To any such φ we can associate a normalized *price measure* $P^\varphi \approx P$ with density $\varphi E[\varphi]^{-1}$.

Remark 3.49. In the context of our one-period model of a financial market with d risky assets $S^1, \dots, S^d$ and a risk-free asset $S^0 \equiv 1 + r$, P^φ is a risk-neutral measure if the pricing rule Φ is consistent with the given price vector $\overline{\pi} = (\pi^0, \pi)$, where $\pi^0 = 1$. In this section, the pricing rule will be derived as an equilibrium price measure, given the agents' preferences and endowments. In particular, this will amount to an endogenous derivation of the price vector π. In a situation where the structure of the equilibrium is already partially known in the sense that it is consistent with the given price vector π, the construction of a microeconomic equilibrium yields a specific choice of a martingale measure P^*, i.e., of a specific extension of π from the space $\mathcal{V}$ of attainable payoffs to a larger space of admissible claims. ◇

The preferences of agent $a \in \mathcal{A}$ are described by a utility function u_a. Given the price density φ, an agent $a \in \mathcal{A}$ may want to exchange the endowment W_a for an admissible claim X_a^φ which maximizes the expected utility

$$E[u_a(X)]$$

among all X in the agent's *budget set*

$$\begin{aligned} B_a(\varphi) &:= \{ X \in \mathcal{X} \mid E[\varphi X] \le E[\varphi W_a] \} \\ &= \{ X \in \mathcal{X} \mid E^\varphi[X] \le E^\varphi[W_a] \}. \end{aligned}$$

In this case, we will say that X_a^φ solves the *utility maximization problem* of agent $a \in \mathcal{A}$ with respect to the price density φ. The key problem is whether φ can be chosen in such a way that the requested profiles X_a^φ, $a \in \mathcal{A}$, form a feasible allocation.

Definition 3.50. A price density φ^* together with a feasible allocation $(X_a^*)_{a\in\mathcal{A}}$ is called an *Arrow–Debreu equilibrium* if each X_a^* solves the utility maximization problem of agent $a \in \mathcal{A}$ with respect to φ^*.

Thus, the price density φ^* appearing in an Arrow–Debreu equilibrium decentralizes the crucial problem of implementing the global feasibility constraint (3.31). This is achieved by adjusting the budget sets in such a way that the resulting demands respect the market clearing condition, even though the individual demand is determined without any regard to this global constraint.

Example 3.51. Assume that each agent $a \in \mathcal{A}$ has an exponential utility function with parameter $\alpha_a > 0$, and let us consider the unconstrained case

$$\mathcal{X} = L^0(\Omega, \mathcal{F}, P).$$

In this case, there is a unique equilibrium, and it is easy to describe it explicitly. For a given pricing measure $P^* \approx P$ such that $W_a \in L^1(P^*)$ for all $a \in \mathcal{A}$, the utility maximization problem for agent $a \in \mathcal{A}$ can be solved if and only if $H(P^*|P) < \infty$, and in this case the optimal demand is given by

$$X_a^* = -\frac{1}{\alpha_a} \log \varphi^* + w_a^* + \frac{1}{\alpha_a} H(P^*|P)$$

where

$$w_a^* := E^*[\, W_a \,];$$

see Example 3.35. The market clearing condition (3.31) takes the form

$$W = -\frac{1}{\alpha} \log \varphi^* + \sum_{a \in \mathcal{A}} w_a^* + \frac{1}{\alpha} H(P^*|P)$$

where α is defined via

$$\frac{1}{\alpha} = \sum_{a \in \mathcal{A}} \frac{1}{\alpha_a}. \tag{3.32}$$

Thus, a normalized equilibrium price density must have the form

$$\varphi^* = \frac{e^{-\alpha W}}{E[\, e^{-\alpha W} \,]}, \tag{3.33}$$

and this shows *uniqueness*. As to *existence*, let us assume that

$$E[\, |W_a| e^{-\alpha W} \,] < \infty, \quad a \in \mathcal{A};$$

this condition is satisfied if, e.g., the random variables W_a are bounded from below. Define $P^* \approx P$ via (3.33). Then

$$H(P^*|P) = -\alpha \, E^*[\, W \,] - \log E[\, e^{-\alpha W} \,] < \infty,$$

and the optimal profile for agent $a \in \mathcal{A}$ with respect to the pricing measure P^* takes the form

$$X_a^* = w_a^* + \frac{\alpha}{\alpha_a} \big(W - E^*[\, W \,] \big). \tag{3.34}$$

Since

$$\sum_{a \in \mathcal{A}} w_a^* = E^*[\, W \,],$$

the allocation $(X_a^*)_{a \in \mathcal{A}}$ is feasible, and so we have constructed an Arrow–Debreu equilibrium. Thus, the agents share the market portfolio in a *linear* way, and in inverse proportion to their risk aversion.

Let us now return to our financial market model of Section 3.1. We assume that the initial endowment of agent $a \in \mathcal{A}$ is given by a portfolio $\overline{\eta}_a \in \mathbb{R}^{d+1}$ so that the discounted payoff at time $t = 1$ is

$$W_a = \frac{\overline{\eta}_a \cdot \overline{S}}{1+r}, \quad a \in \mathcal{A}.$$

In this case, the market portfolio is given by $W = \overline{\eta} \cdot \overline{S}/(1+r)$ with $\overline{\eta} := \sum_a \overline{\eta}_a = (\eta^0, \eta)$. The optimal claim for agent $a \in \mathcal{A}$ in (3.34) takes the form

$$X_a^* = \overline{\eta}_a \cdot \overline{\pi} + \frac{\alpha}{\alpha_a} \eta \cdot \left(\frac{S}{1+r} - \pi \right),$$

where $\overline{\pi} = (1, \pi)$ and

$$\pi^i = E^*\left[\frac{S^i}{1+r} \right] \quad \text{for } i = 1, \dots, d.$$

Thus, we could have formulated the equilibrium problem within the smaller space $\mathcal{X} = \mathcal{V}$ of attainable payoffs, and the resulting equilibrium allocation would have been the same. In particular, the extension of $\mathcal{X}$ from $\mathcal{V}$ to the general space $L^0(\Omega, \mathcal{F}, P)$ of admissible claims does not create a demand for derivatives in our present example. ◇

From now on we assume that the set of admissible claims is given by

$$\mathcal{X} = L^0_+(\Omega, \mathcal{F}, P),$$

and that the preferences of agent $a \in \mathcal{A}$ are described by a utility function $u_a : [0, \infty) \to \mathbb{R}$ which is continuously differentiable on $(0, \infty)$. In particular, the initial endowments W_a are assumed to be non-negative. Moreover, we assume

$$P[\, W_a > 0 \,] \neq 0 \quad \text{for all } a \in \mathcal{A}.$$

and

$$E[\, W \,] < \infty. \tag{3.35}$$

A function $\varphi \in L^1(\Omega, \mathcal{F}, P)$ such that $\varphi > 0$ P-a.s. is a price density if

$$E[\, \varphi W \,] < \infty\,;$$

note that this condition is satisfied as soon as φ is bounded, due to our assumption (3.35). Given a price density φ, each agent faces exactly the optimization problem discussed in Section 3.3 in terms of the price measure $P^\varphi \approx P$. Thus, if $(X_a^*)_{a \in \mathcal{A}}$ is an equilibrium allocation with respect to the price density φ^*, feasibility implies $0 \le X_a^* \le W$, and so it follows as in the proof of Corollary 3.42 that

$$X_a^* = I_a^+(c_a\, \varphi^*), \quad a \in \mathcal{A}, \tag{3.36}$$

with positive constants $c_a > 0$. Note that the market clearing condition

$$W = \sum_{a \in \mathcal{A}} X_a^* = \sum_{a \in \mathcal{A}} I_a^+(c_a \varphi^*)$$

will determine φ^* as a *decreasing* function of W, and thus the optimal profiles X_a^* will be *increasing* functions of W.

Before we discuss the *existence* of an Arrow–Debreu equilibrium, let us first illustrate the structure of such equilibria by the following simple examples. In particular, they show that an equilibrium allocation will typically involve non-linear derivatives of the market portfolio W.

Example 3.52. Let us consider the constrained version of the preceding example where agents $a \in \mathcal{A}$ have exponential utility functions with parameters $\alpha_a > 0$. Define

$$\underline{w} := \sup\{\, c \mid W \geq c \; P\text{-a.s.}\,\} \geq 0,$$

and let P^* be the measure defined via (3.33). For any agent $a \in \mathcal{A}$ such that

$$w_a^* := E^*[\, W_a \,] \geq \frac{\alpha}{\alpha_a}\big(E^*[\, W \,] - \underline{w}\big), \tag{3.37}$$

the unrestricted optimal profile

$$X_a^* = w_a^* + \frac{\alpha}{\alpha_a}\big(W - E^*[\, W \,]\big)$$

satisfies $X_a^* \geq 0$ P-a.s. Thus, if *all* agents satisfy the requirement (3.37) then the unrestricted equilibrium computed in Example 3.51 is a forteriori an Arrow–Debreu equilibrium in our present context. In this case, there is no need for non-linear derivatives of the market portfolio.

If some agents do *not* satisfy the requirement (3.37) then the situation becomes more involved, and the equilibrium allocation will need derivatives such as call options. Let us illustrate this effect in the simple setting where there are only two agents $a \in \mathcal{A} = \{1, 2\}$. Suppose that agent 1 satisfies condition (3.37), while agent 2 does not. For $c \geq 0$, we define the measure $P^c \approx P$ in terms of the density

$$\varphi^c := \begin{cases} \dfrac{1}{Z_1}\, e^{-\alpha_1 W} & \text{on } \{W \leq c\}, \\[2ex] \dfrac{1}{Z_2}\, e^{-\alpha W} & \text{on } \{W \geq c\}, \end{cases}$$

where α is given by (3.32), and where the constants Z_1 and Z_2 are determined by the continuity condition

$$\log Z_2 - \log Z_1 = c(\alpha_1 - \alpha)$$

and by the normalization $E[\varphi^c] = 1$. Note that $P^0 = P^*$ with P^* as in (3.33). Consider the equation

$$\frac{\alpha}{\alpha_2} E^c[(W-c)^+] = w_2^c := E^c[W_2]. \tag{3.38}$$

Both sides are continuous in c. As c increases from 0 to $+\infty$, the left-hand side decreases from $\frac{\alpha}{\alpha_2} E^*[W]$ to 0, while w_2^c goes from $w_2^0 < \frac{\alpha}{\alpha_2} E^*[W]$ to $E^\infty[W_2] > 0$. Thus, there exists a solution c of (3.38). Let us now check that

$$X_2^c := \frac{\alpha}{\alpha_2}(W-c)^+, \quad X_1^c := W - X_2^c$$

defines an equilibrium allocation with respect to the pricing measure P^c. Clearly, X_1^c and X_2^c are non-negative and satisfy $X_1^c + X_2^c = W$. The budget condition for agent 2 is satisfied due to (3.38), and this implies the budget condition

$$E^c[X_1^c] = E^c[W] - w_2^c = w_1^c$$

for agent 1. Both are optimal since

$$X_a^c = I_a^+(\lambda_a \varphi^c)$$

with

$$\lambda_1 := \alpha_1 Z_1 \quad \text{and} \quad \lambda_2 := \alpha_2 Z_2 e^{\alpha c}.$$

Thus, agent 2 demands $\frac{\alpha}{\alpha_2}$ shares of a call option on the market portfolio W with strike c, agent 1 demands the remaining part of W, and so the market is cleared.

In the general case of a finite set $\mathcal{A}$ of agents, the equilibrium price measure $\hat{P}$ has the following structure. There are levels $0 := c_0 < \cdots < c_N = \infty$ with $1 \le N \le |\mathcal{A}|$ such that the price density $\hat{\varphi}$ is given by

$$\hat{\varphi} = \frac{1}{Z_i} e^{-\beta_i W} \quad \text{on } \{ W \in [c_{i-1}, c_i] \}$$

for $i = 1, \dots, N$, where

$$\beta_i := \Big(\sum_{\alpha \in \mathcal{A}_i} \frac{1}{\alpha_a} \Big)^{-1},$$

and where $\mathcal{A}_i$ $(i = 1, \dots, N)$ are the increasing sets of agents which are active at the i^{th} layer in the sense that $X_a > 0$ on $\{W \in (c_{i-1}, c_i]\}$. At each layer $(c_{i-1}, c_i]$, the active agents are sharing the market portfolio in inverse proportions to their risk aversion. Thus, the optimal profile $\hat{X}_a$ of any agent $a \in \mathcal{A}$ is given by an increasing piecewise linear function in W, and thus it can be implemented by a linear combination of call options with strikes c_i. More precisely, an agent $a \in \mathcal{A}_i$ takes β_i/α_a shares of the *spread*

$$(W - c_{i-1})^+ - (W - c_i)^+,$$

i.e., the agent goes long on a call option with strike c_{i-1} and short on a call option with strike c_i. ◇

Example 3.53. Assume that all agents $a \in \mathcal{A}$ have preferences described by HARA utility functions so that

$$I_a^+(y) = y^{-\frac{1}{1-\gamma_a}}, \quad a \in \mathcal{A}$$

with $0 \le \gamma_a < 1$. For a given price density φ, the optimal claims take the form

$$X_a = I_a^+(c_a \varphi) = b_a \, \varphi^{-\frac{1}{1-\gamma_a}} \tag{3.39}$$

with constants $b_a > 0$. If $\gamma_a = \gamma$ for all $a \in \mathcal{A}$, then the market clearing condition (3.31) implies

$$W = \sum_{a \in \mathcal{A}} X_a = \Big(\sum_{a \in \mathcal{A}} b_a\Big)\varphi^{-\frac{1}{1-\gamma}},$$

i.e., the equilibrium price density φ^* takes the form

$$\varphi^* = \frac{1}{Z}\, W^{\gamma-1},$$

where Z is the normalizing constant, and so the agents demand *linear* shares of the market portfolio W. If risk aversion varies among the agents then the structure of the equilibrium becomes more complex, and it will involve non-linear *derivatives* of the market portfolio. Let us number the agents so that $\mathcal{A} = \{1, \dots, n\}$ and $\gamma_1 \ge \dots \ge \gamma_n$. Condition (3.39) implies

$$X_i = d_i \, X_n^{\beta_i}$$

with some constants d_i, and where

$$\beta_i := \frac{1-\gamma_n}{1-\gamma_i}$$

satisfies $\beta_1 \ge \dots \ge \beta_n = 1$ with at least one strict inequality. Thus, each X_i is a convex increasing function of X_n. In equilibrium, X_n is a concave function of W determined by the condition

$$\sum_{i=1}^{n} d_i X_n^{\beta_i} = W, \tag{3.40}$$

and the price density φ^* takes the form

$$\varphi^* = \frac{1}{Z}\, X_n^{\gamma_n - 1}.$$

As an illustration, we consider the special case “Bernoulli vs. Cramer”, where $\mathcal{A} = \{1, 2\}$ with $u_1(x) = \sqrt{x}$ and $u_2(x) = \log x$, i.e., $\gamma_1 = \frac{1}{2}$ and $\gamma_2 = 0$; see Example 2.40. The solutions of (3.40) can be parameterized with $c \ge 0$ such that

$$X_2^c = 2\sqrt{c}\big(\sqrt{W+c} - \sqrt{c}\big) \in [0, W]$$

and

$$X_1^c = W - X_2^c.$$

The corresponding price density takes the form

$$\varphi^c = \frac{1}{Z(c)} \frac{1}{\sqrt{W+c} - \sqrt{c}},$$

where $Z(c)$ is the normalizing constant. Now assume that $W^{-1} \in L^1(P)$, and let P^∞ denote the measure with density $W^{-1}\big(E[\,W^{-1}\,]\big)^{-1}$. As c increases from 0 to ∞, $E^c[\,X_2^c\,]$ increases continuously from 0 to $E^\infty[\,W\,]$, while $E^c[\,W_2\,]$ goes continuously from $E^0[\,W_2\,] > 0$ to $E^\infty[\,W_2\,] < E^\infty[\,W\,]$; here we use our assumption that $P[\,W_a > 0\,] \neq 0$ for all $a \in \mathcal{A}$. Thus, there is a $c \in (0, \infty)$ such that

$$E^c[\,X_2^c\,] = E^c[\,W_2\,],$$

and this implies that the budget constraint is satisfied for both agents. With this choice of the parameter c, (X_1^c, X_2^c) is an equilibrium allocation with respect to the pricing measure P^c: Agent 2 demands the concave profile X_2^c, agent 1 demands the convex profile X_1^c, both in accordance with their budget constraints, and the market is cleared. ◇

Let us now return to our general setting, and let us prove the existence of an Arrow–Debreu equilibrium. Consider the following condition:

$$\limsup_{x \downarrow 0} x\, u_a'(x) < \infty \quad \text{and} \quad E\left[\, u_a'\left(\frac{W}{|\mathcal{A}|}\right)\right] < \infty, \qquad a \in \mathcal{A}. \tag{3.41}$$

Remark 3.54. Condition (3.41) is clearly satisfied if

$$u_a'(0) := \lim_{x \downarrow 0} u_a'(x) < \infty, \qquad a \in \mathcal{A}. \tag{3.42}$$

But it also includes HARA utility functions u_a with parameter $\gamma_a \in [0, 1)$ if we assume

$$E[\,W^{\gamma_a - 1}\,] < \infty, \qquad a \in \mathcal{A},$$

in addition to our assumption $E[\,W\,] < \infty$. ◇

Theorem 3.55. *Under assumptions* (3.35) *and* (3.41)*, there exists an Arrow–Debreu equilibrium.*

In a first step, we are going to show that an equilibrium allocation maximizes a suitable weighted average

$$U^\lambda(X) := \sum_{a \in \mathcal{A}} \lambda_a\, E[\,u_a(X_a)\,]$$

of the individual utility functionals over all feasible allocations $X = (X_a)_{a\in\mathcal{A}}$. The weights are non-negative, and without loss of generality we can assume that they are normalized so that the vector $\lambda := (\lambda_a)_{a\in\mathcal{A}}$ belongs to the convex compact set

$$\Lambda = \Big\{ \lambda \in [0,1]^{\mathcal{A}} \;\Big|\; \sum_{a\in\mathcal{A}} \lambda_a = 1 \Big\}.$$

In a second step, we will use a fixed-point argument to obtain a weight vector and a corresponding price density such that the maximizing allocation satisfies the individual budget constraints.

Definition 3.56. A feasible allocation $(X_a)_{a\in\mathcal{A}}$ is called λ-*efficient* for $\lambda \in \Lambda$ if it maximizes U^λ over all feasible allocations.

In view of (3.36), part (b) of the following lemma shows that the equilibrium allocation $(X_a^*)_{a\in\mathcal{A}}$ in an Arrow–Debreu equilibrium is λ-efficient for the vector $\lambda = (c \cdot c_a^{-1})_{a\in\mathcal{A}}$, where $c^{-1} := \sum_a c_a^{-1}$. Thus, the existence proof for an Arrow–Debreu equilibrium is reduced to the construction of a suitable vector $\lambda^* \in \Lambda$.

Lemma 3.57. (a) *For any $\lambda \in \Lambda$ there exists a unique λ-efficient allocation $(X_a^\lambda)_{a\in\mathcal{A}}$.*

(b) *A feasible allocation $(X_a)_{a\in\mathcal{A}}$ is λ-efficient if and only if it satisfies the* first order conditions

$$\lambda_a\, u_a'(X_a) \le \varphi, \quad \textit{with equality on } \{X_a > 0\} \tag{3.43}$$

with respect to some price density φ. In this case, $(X_a)_{a\in\mathcal{A}}$ coincides with $(X_a^\lambda)_{a\in\mathcal{A}}$, and the price density can be chosen as

$$\varphi^\lambda := \max_{a\in\mathcal{A}} \lambda_a\, u_a'(X_a^\lambda). \tag{3.44}$$

(c) *For each $a \in \mathcal{A}$, X_a^λ maximizes $E[\, u_a(X)\,]$ over all $X \in \mathcal{X}$ such that*

$$E[\,\varphi^\lambda X\,] \le E[\,\varphi^\lambda X_a^\lambda\,].$$

Proof. (a): Existence and uniqueness follow from the general argument in Remark 3.37 applied to the set $\mathcal{B}$ of all feasible allocations and to the functional U^λ. Note that

$$U^\lambda(X) \le \max_{a\in\mathcal{A}} E[\, u_a(W)\,]$$

for any feasible allocation, and that the right-hand side is finite due to our assumption (3.35). Moreover, by dominated convergence, U^λ is indeed continuous on $\mathcal{B}$ with respect to P-a.s. convergence.

(b): Let us first show sufficiency. If $X = (X_a)_{a\in\mathcal{A}}$ is a feasible allocation satisfying the first order conditions, and $Y = (Y_a)_{a\in\mathcal{A}}$ is another feasible allocation then

$$\begin{aligned} U^\lambda(X) - U^\lambda(Y) &= \sum_{a\in\mathcal{A}} \lambda_a E[\, u_a(X_a) - u_a(Y_a)\,] \\ &\ge \sum_{a\in\mathcal{A}} \lambda_a E[\, u_a'(X_a)(X_a - Y_a)\,] \\ &\ge E\Big[\,\varphi\Big(\sum_{a\in\mathcal{A}} X_a - \sum_{a\in\mathcal{A}} Y_a\Big)\Big] = 0, \end{aligned}$$

using concavity of u_a in the second step and the first order conditions in the third. This shows that X is λ-efficient.

Turning to necessity, consider the λ-efficient allocation $(X_a^\lambda)_{a\in\mathcal{A}}$ for $\lambda \in \Lambda$ and another feasible allocation $(X_a)_{a\in\mathcal{A}}$. For $\varepsilon \in (0,1]$, let $Y_a^\varepsilon := \varepsilon X_a + (1-\varepsilon)X_a^\lambda$. Since $(Y_a^\varepsilon)_{a\in\mathcal{A}}$ is feasible, λ-efficiency of $(X_a^\lambda)_{a\in\mathcal{A}}$ yields

$$\begin{aligned} 0 &\ge \frac{1}{\varepsilon}\sum_{a\in\mathcal{A}} \lambda_a\, E[\, u_a(Y_a^\varepsilon) - u_a(X_a^\lambda)\,] \\ &\ge \frac{1}{\varepsilon}\sum_{a\in\mathcal{A}} \lambda_a E[\, u_a'(Y_a^\varepsilon)(Y_a^\varepsilon - X_a^\lambda)\,] \\ &= \sum_{a\in\mathcal{A}} \lambda_a E[\, u_a'(Y_a^\varepsilon)(X_a - X_a^\lambda)\,]. \end{aligned} \tag{3.45}$$

Let us first assume (3.42); in part (d) of the proof we show how to modify the argument under condition (3.41). Using dominated convergence and (3.42), we may let $\varepsilon \downarrow 0$ in the above inequality to conclude

$$\sum_{a\in\mathcal{A}} E[\,\varphi_a^\lambda X_a\,] \le \sum_{a\in\mathcal{A}} E[\,\varphi_a^\lambda X_a^\lambda\,] \le E[\,\varphi^\lambda W\,], \tag{3.46}$$

where

$$\varphi_a^\lambda := \lambda_a u_a'(X_a^\lambda).$$

Note that φ^λ is a price density since by (3.42)

$$0 < \varphi^\lambda \le \max\{\,\lambda_a u_a'(0) \mid a \in \mathcal{A}\,\} < \infty.$$

Take a feasible allocation $(X_a)_{a\in\mathcal{A}}$ such that

$$\sum_{a\in\mathcal{A}} \varphi_a^\lambda X_a = \varphi^\lambda\, W\,; \tag{3.47}$$

for example, we can enumerate $\mathcal{A} := \{1, \dots, |\mathcal{A}|\}$ and take $X_a := W I_{\{T=a\}}$ where

$$T(\omega) := \min\{\, a \mid \varphi_a^\lambda(\omega) = \varphi^\lambda(\omega)\,\}.$$

In view of (3.46), we see that

$$\sum_{a\in\mathcal{A}} E[\,\varphi_a^\lambda\, X_a^\lambda\,] = E[\,\varphi^\lambda\, W\,]. \tag{3.48}$$

This implies $\varphi_a^\lambda = \varphi^\lambda$ on $\{X_a^\lambda > 0\}$, which is equivalent to the first order condition (3.43) with respect to φ^λ.

(c): In order to show optimality of X_a^λ, we may assume without loss of generality that $P[\,X_a^\lambda > 0\,] > 0$, and hence $\lambda_a > 0$. Thus, the first order condition with respect to φ^λ takes the form

$$X_a^\lambda = I_a^+(\lambda_a^{-1}\varphi^\lambda),$$

due to our convention (3.22). By Corollary 3.42, X_a^λ solves the optimization problem for agent $a \in \mathcal{A}$ under the constraint

$$E[\,\varphi^\lambda\, X\,] \le E[\,\varphi^\lambda\, X_a^\lambda\,].$$

(d): If (3.42) is replaced by (3.41), then we first need an additional argument in order to pass from (3.45) to (3.46). Note first that by Fatou's lemma,

$$\begin{aligned}\liminf_{\varepsilon\downarrow 0} \sum_{a\in\mathcal{A}} \lambda_a E[\,u_a'(Y_a^\varepsilon)\, X_a\,] &\ge \sum_{a\in\mathcal{A}} \lambda_a \liminf_{\varepsilon\downarrow 0} E[\,u_a'(Y_a^\varepsilon)\, X_a\,]\\ &\ge \sum_{a\in\mathcal{A}} \lambda_a E[\,u_a'(X_a^\lambda)\, X_a\,].\end{aligned}$$

On the other hand, since

$$\kappa := \max_{a\in\mathcal{A}} \sup_{0<x\le 1} x\, u_a'(x) < \infty$$

by (3.41), we have $x u_a'(x) \le \kappa + x u_a'(1) \le \kappa(1+x)$ for all $x \ge 0$. This implies

$$u_a'(X_a)X_a \le V := \kappa(1+W) \in L^1(P), \tag{3.49}$$

and also

$$u_a'(Y_a^\varepsilon)\, X_a^\lambda \le u_a'\big((1-\varepsilon)\, X_a^\lambda\big) X_a^\lambda \le (1-\varepsilon)^{-1}\, V,$$

since $Y_a^\varepsilon \ge (1-\varepsilon)X_a^\lambda$. Thus, dominated convergence implies

$$E[\,u_a'(Y_a^\varepsilon)\, X_a^\lambda\,] \longrightarrow E[\,u_a'(X_a^\lambda)\, X_a^\lambda\,], \quad \varepsilon \downarrow 0,$$

and this concludes the proof of (3.46).

By (3.49), we have

$$\varphi_a^\lambda\, X_a^\lambda := \lambda_a\, u_a'(X_a^\lambda)\, X_a^\lambda \in L^1(P).$$

Hence $E[\,\varphi^\lambda\, W\,] < \infty$ follows by taking in (3.46) a feasible allocation $(X_a)_{a\in\mathcal{A}}$ which is as in (3.47). We furthermore get (3.48), which yields as in part (b) the first order conditions (3.43).

It remains to show that φ^λ is integrable in order to conclude that φ^λ is a price density. Our assumption (3.41) implies

$$F := \max_{a\in\mathcal{A}} u_a'\left(\frac{W}{|\mathcal{A}|}\right) \in L^1(P), \tag{3.50}$$

and so it is enough to show that $F \geq \varphi^\lambda$. Since $X_a^\lambda = I_a^+\big(\varphi^\lambda/\lambda_a\big)$, feasibility and $\lambda_a \leq 1$ imply

$$W \leq \sum_{a\in\mathcal{A}} I_a^+(\varphi^\lambda) \leq |\mathcal{A}| \max_{a\in\mathcal{A}} I_a^+(\varphi^\lambda),$$

hence

$$\begin{aligned} F &\geq \max_{a\in\mathcal{A}} u_a'\big(\max_{b\in\mathcal{A}} I_b^+(\varphi^\lambda)\big) \\ &\geq u_{a_0}'\big(I_{a_0}^+(\varphi^\lambda)\big) = \varphi^\lambda \quad \text{on } \big\{\max_{a\in\mathcal{A}} I_a^+(\varphi^\lambda) = I_{a_0}^+(\varphi^\lambda)\big\}. \end{aligned}$$ □

After these preliminaries, we are now in a position to prove the existence of an Arrow–Debreu equilibrium. Note that for each $\lambda \in \Lambda$ the λ-efficient allocation $(X_a^\lambda)_{a\in\mathcal{A}}$ and the price density φ^λ would form an Arrow–Debreu equilibrium if

$$E[\,\varphi^\lambda\, W_a\,] = E[\,\varphi^\lambda\, X_a^\lambda\,] \quad \text{for all } a \in \mathcal{A}. \tag{3.51}$$

If this is not the case, then we can replace λ by the vector $g(\lambda) = \big(g_a(\lambda)\big)_{a\in\mathcal{A}}$ defined by

$$g_a(\lambda) := \lambda_a + \frac{1}{E[\,V\,]} \cdot E[\,\varphi^\lambda(W_a - X_a^\lambda)\,],$$

where V is given by (3.49). Note that $g(\lambda) \in \Lambda$: Since the first order conditions (3.43) together with (3.49) imply

$$E[\,\varphi^\lambda\, X_a^\lambda\,] = \lambda_a\, E[\,u_a'(X_a^\lambda)\, X_a^\lambda\,] \leq \lambda_a\, E[\,V\,],$$

we have $g_a(\lambda) \geq 0$, and $\sum_a g_a(\lambda) = 1$ follows by feasibility. Thus, we increase the weights of agents which were allocated less than they could afford. Clearly, any fixed point of the map $g : \Lambda \to \Lambda$ will satisfy condition (3.51) and thus yield an Arrow–Debreu equilibrium.

Proof of Theorem 3.55. (a): The set Λ is convex and compact. Thus, the existence of a fixed point of the map $g : \Lambda \to \Lambda$ follows from Brouwer's fixed point theorem as soon as we can verify that g is *continuous*; see, for instance, Corollary 16.52 in [2] for a proof of Brouwer's fixed point theorem. Suppose that the sequence $(\lambda_n) \subset \Lambda$ converges to $\lambda \in \Lambda$. In part (c) we show that $X_n := X^{\lambda_n}$ and $\varphi_n := \varphi^{\lambda_n}$ converge P-a.s. to X^λ and φ^λ, respectively. We will show next that we may apply the dominated convergence theorem, so that

$$\lim_{n\uparrow\infty} E[\,\varphi_n\, W_a\,] = E[\,\varphi^\lambda\, W_a\,]$$

and

$$\lim_{n\uparrow\infty} E[\,\varphi_n\, X_n\,] = E[\,\varphi^\lambda\, X^\lambda\,]$$

and this will prove the continuity of g. To verify the assumptions of the dominated convergence theorem, note that

$$W_a\,\varphi_n \le W\,\varphi_n \le W F,$$

where F is as in (3.50). Moreover,

$$W F \le |\mathcal{A}| F\, \mathrm{I}_{\{W \le |\mathcal{A}|\}} + \max_{a\in\mathcal{A}} u_a'(1)\cdot W \in L^1(P).$$

Thus, $\varphi_n\, W_a$ and $\varphi_n\, X_n$ are bounded by $W F \in L^1(P)$.

(b): By our convention (3.22), the map $f : \Lambda \times [0,\infty] \to [0,\infty]$ defined by

$$f(\lambda, y) = \sum_{a\in\mathcal{A}} I_a^+(\lambda_a^{-1}\, y)$$

is continuous. If we fix $\lambda \in \Lambda$, then the function $f(\lambda,\cdot)$ is continuous on $[0,\infty]$ and strictly decreasing on $(a(\lambda), b(\lambda))$ where

$$a(\lambda) := \max_{a\in\mathcal{A}} \lim_{x\uparrow\infty} \lambda_a\, u_a'(x) \ge 0 \quad \text{and} \quad b(\lambda) = \max_{a\in\mathcal{A}} \lambda_a\, u_a'(0+) \le +\infty.$$

Moreover, $f(\lambda, y) = \infty$ for $y \le a(\lambda)$ and $f(\lambda, y) = 0$ for $y \ge b(\lambda)$. Hence, for each $w \in (0,\infty)$ there exists exactly one solution $y^\lambda \in (a(\lambda), b(\lambda))$ of the equation

$$f(\lambda, y^\lambda) = w.$$

Recall that $[0,\infty]$ can be regarded as a compact topological space. To see that y^λ depends continuously on $\lambda \in \Lambda$, take a sequence $\lambda_n \to \lambda$ and a subsequence (λ_{n_k}) such that the solutions $y_k = y^{\lambda_{n_k}}$ of $f(\lambda_{n_k}, y) = w$ converge to some limit $y_\infty \in [a(\lambda), b(\lambda)]$. By continuity of f,

$$f(\lambda, y_\infty) = \lim_{k\uparrow\infty} f(\lambda_{n_k}, y_k) = w,$$

and so y_∞ must coincide with y^λ.

(c): Recall that

$$X_a^\lambda = I_a^+(\lambda_a^{-1}\varphi^\lambda) \tag{3.52}$$

for any $a \in \mathcal{A}$. By feasibility,

$$W = \sum_{a\in\mathcal{A}} X_a^\lambda = f(\lambda, \varphi^\lambda).$$

Thus, φ^{λ_n} converges P-a.s. to φ^λ as $\lambda_n \to \lambda$ due to part (b), and so X^{λ_n} converges P-a.s. to X^λ due to (3.52). This completes the proof in (a) that the map g is continuous. □

Remark 3.58. In order to simplify the exposition, we have restricted the discussion of equilibrium prices to contingent claims with payoff at time $t = 1$. We have argued in terms of discounted payoffs, and so we have implicitly assumed that the interest rate r has already been fixed. From an economic point of view, also the interest rate should be determined by an equilibrium argument. This requires an intertemporal extension of our setting, which distinguishes between deterministic payoffs y at time $t = 0$ and nominal contingent payoffs Y at time $t = 1$. Thus, we replace $\mathcal{X} = L^0_+$ by the space

$$\overline{\mathcal{Y}} := \left\{ \overline{Y} = (y, Y) \mid y \in [0, \infty),\ Y \in L^0_+ \right\}.$$

A pricing rule is given by a linear functional on $\overline{\mathcal{Y}}$ of the form

$$\overline{\Phi}(\overline{Y}) := \varphi_0 \cdot y + E[\,\varphi\, Y\,],$$

where $\varphi_0 \in (0, \infty)$ and φ is a price density as before. Any such price system specifies an interest rate for transferring income from time $t = 0$ to time $t = 1$. Indeed, comparing the *forward price* $c \cdot E[\,\varphi\,]$ for the fixed amount c to be delivered at time 1 with the *spot price* $c \cdot \varphi_0$ for the amount c made available at time 0, we see that the implicit interest rate is given by

$$1 + r = \frac{E[\,\varphi\,]}{\varphi_0}.$$

If we describe the preferences of agent $a \in \mathcal{A}$ by a utility functional of the form

$$U_a(\overline{Y}) = u_{a,0}(y) + E[u_{a,1}(Y)]$$

with smooth utility functions $u_{a,0}$ and $u_{a,1}$, then we can show along the lines of the preceding discussion that an Arrow–Debreu equilibrium exists in this extended setting. Thus, we obtain an equilibrium allocation $(\overline{Y}^*_a)_{a\in\mathcal{A}}$ and an equilibrium price system $\overline{\Phi}^* = (\varphi^*_0, \varphi^*)$ such that each $\overline{Y}^*_a$ maximizes the functional U_a in the agent's budget set determined by an initial endowment in $\overline{\mathcal{Y}}$ and by the pricing rule $\overline{\Phi}^*$. In particular, we have then specified an equilibrium interest rate r^*. Normalizing the price system to $\varphi^*_0 = 1$ and defining P^* as a probability measure with density $\varphi^*/E[\,\varphi^*\,]$, we see that the price at time $t = 0$ of a contingent claim with nominal payoff $Y \geq 0$ at time $t = 1$ is given as the expectation

$$E^*\left[\,\frac{Y}{1 + r^*}\,\right]$$

of the discounted claim with respect to the measure P^*. ◇

Let us now extend the discussion to situations where agents are heterogeneous not only in their utility functions but also in their expectations. Thus, we assume that the preferences of agent $a \in \mathcal{A}$ are described by a Savage functional of the form

$$U_a(X) := E_{Q_a}[\,u_a(X)\,],$$

where Q_a is a probability measure on $(\Omega, \mathcal{F})$ which is equivalent to P. In addition to our assumption

$$\limsup_{x\downarrow 0} x\, u_a'(x) < \infty, \quad a \in \mathcal{A}, \tag{3.53}$$

we assume that

$$E_{Q_a}[\, W\,] < \infty \quad \text{and} \quad E_{Q_a}\left[\, u_a'\left(\frac{W}{|\mathcal{A}|}\right)\right] < \infty, \quad a \in \mathcal{A}, \tag{3.54}$$

As before, a feasible allocation $(X_a^*)_{a\in\mathcal{A}}$ together with a price density φ^* is called an Arrow–Debreu equilibrium if each X_a^* maximizes the functional U_a on the budget set of agent $a \in \mathcal{A}$, which is determined by φ^*.

Theorem 3.59. *Under assumptions* (3.35)*,* (3.53)*, and* (3.54)*, there exists an Arrow–Debreu equilibrium.*

Proof. For any $\lambda \in \Lambda$, the general argument of Remark 3.37 yields the existence of a λ-efficient allocation $(X_a^\lambda)_{a\in\mathcal{A}}$, i.e., of a feasible allocation which maximizes the functional

$$U^\lambda(X) := \sum_{a\in\mathcal{A}} \lambda_a U_a(X_a)$$

over all feasible allocations $X = (X_a)_{a\in\mathcal{A}}$. Since

$$U_a(X_a^\lambda) = E[\, \varphi_a\, u_a(X_a^\lambda)\,],$$

$(X_a^\lambda)_{a\in\mathcal{A}}$can be viewed as a λ-efficient allocation in the model where agents have random utility functions of the form

$$\widetilde{u}_a(x, \omega) = u_a(x)\, \varphi_a(\omega),$$

while their expectations are homogeneous and given by P. In view of Corollary 3.43, it follows as before that X^λ satisfies the first order conditions

$$X_a^\lambda = I_a^+(\lambda_a^{-1}\varphi_a^{-1}\, \varphi^\lambda), \quad a \in \mathcal{A},$$

with

$$\varphi^\lambda = \max_{a\in\mathcal{A}} \lambda_a\, u_a'(X_a^\lambda)\, \varphi_a,$$

and that X_a^λ satisfies

$$U_a(X_a^\lambda) \geq E[\, u_a(Y_a)\, \varphi_a\,] \geq U_a(Y_a)$$

for all Y_a in the budget set of agent $a \in \mathcal{A}$. The remaining arguments are essentially the same as in the proof of Theorem 3.55. □

Chapter 4
Monetary measures of risk

In this chapter, we discuss the problem of quantifying the *risk* of a financial position. As in Chapter 2, such a position will be described by the corresponding payoff profile, that is, by a real-valued function X on some set of possible scenarios. In a probabilistic model, specified by a probability measure on scenarios, we could focus on the resulting distribution of X and try to measure the risk in terms of moments or quantiles. Note that a classical measure of risk such as the variance does not capture a basic asymmetry in the financial interpretation of X: Here it is the *downside risk* that matters. This asymmetry is taken into account by measures such as Value at Risk which are based on quantiles for the lower tail of the distribution, see Section 4.4 below. Value at Risk, however, fails to satisfy some natural consistency requirements. Such observations have motivated the systematic investigation of measures of risk that satisfy certain basic axioms.

From the point of view of an investor, we could simply turn around the discussion of Chapter 2 and measure the risk of a position X in terms of the loss functional

$$L(X) = -U(X).$$

Here U is a utility functional representing a given preference relation $\succ$ on financial positions. Assuming robust preferences, we are led to the notion of *robust shortfall risk* defined by

$$L(X) = \sup_{Q \in \mathcal{Q}} E_Q[\,\ell(-X)\,],$$

where $\ell(x) := -u(-x)$ is a convex increasing loss function and $\mathcal{Q}$ is a class of probability measures. The results of Section 2.5 show how such loss functionals can be characterized in terms of convexity and monotonicity properties of the preference relation. In particular, a financial position could be viewed as being *acceptable* if the robust shortfall risk of X does not exceed a given bound.

From the point of view of a supervising agency, however, a specific monetary purpose comes into play. In this perspective a risk measure is viewed as a capital requirement: We are looking for the minimal amount of capital which, if added to the position and invested in a risk-free manner, makes the position acceptable. This monetary interpretation is captured by an additional axiom of *cash invariance*. Together with convexity and monotonicity, it singles out the class of *convex measures of risk*. These measures can be represented in the form

$$\rho(X) = \sup_{Q} \big(E_Q[\,-X\,] - \alpha(Q) \big),$$

where α is a penalty function defined on probability measures on Ω. Under the additional condition of *positive homogeneity*, we obtain the class of *coherent* risk measures. Here we are back to the situation in Proposition 2.85, and the representation takes the form

$$\rho(X) = \sup_{Q \in \mathcal{Q}} E_Q[-X],$$

where $\mathcal{Q}$ is some class of probability measures on Ω.

The axiomatic approach to such monetary risk measures was initiated by P. Artzner, F. Delbaen, J. Eber, and D. Heath [11], and it will be developed in the first three sections. In Section 4.4 we discuss some coherent risk measures related to Value at Risk. These risk measures only involve the distribution of a position under a given probability measure. In Section 4.5 we characterize the class of convex risk measures which share this property of law-invariance. Section 4.6 discusses the role of concave distortions, and in Section 4.7 the resulting risk measures are characterized by a property of comonotonicity. In Section 4.8 we discuss measures of risk which arise naturally in the context of a financial market model. In Section 4.9 we analyze the structure of monetary measures of risk which are induced by our notion of robust shortfall risk.

4.1 Risk measures and their acceptance sets

Let Ω be a fixed set of scenarios. A *financial position* is described by a mapping $X : \Omega \longrightarrow \mathbb{R}$ where $X(\omega)$ is the discounted net worth of the position at the end of the trading period if the scenario $\omega \in \Omega$ is realized. Our aim is to quantify the risk of X by some number $\rho(X)$, where X belongs to a given class $\mathcal{X}$ of financial positions. Throughout this section, $\mathcal{X}$ will be a linear space of bounded functions containing the constants. We do not assume that a probability measure is given on Ω.

Definition 4.1. A mapping $\rho : \mathcal{X} \to \mathbb{R}$ is called a *monetary measure of risk* if it satisfies the following conditions for all $X, Y \in \mathcal{X}$.

- *Monotonicity:* If $X \leq Y$, then $\rho(X) \geq \rho(Y)$.
- *Cash invariance:* If $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$.

The financial meaning of monotonicity is clear: The downside risk of a position is reduced if the payoff profile is increased. Cash invariance is also called *translation invariance*. It is motivated by the interpretation of $\rho(X)$ as a capital requirement, i.e., $\rho(X)$ is the amount which should be added to the position X in order to make it acceptable from the point of view of a supervising agency. Thus, if the amount m is added to the position and invested in a risk-free manner, the capital requirement is reduced by the same amount. In particular, cash invariance implies

$$\rho\big(X + \rho(X)\big) = 0, \tag{4.1}$$

and

$$\rho(m) = \rho(0) - m \quad \text{for all } m \in \mathbb{R}.$$

For most purposes it would be no loss of generality to assume that a given monetary risk measure satisfies the condition of

- *Normalization:* $\rho(0) = 0$.

In some situations, however, it will be convenient not to insist on normalization.

Remark 4.2. We are using the convention that X describes the worth of a financial position after discounting. For instance, the discounting factor can be chosen as $1/(1+r)$ where r is the return of a risk-free investment. Instead of measuring the risk of the discounted position X, one could consider directly the nominal worth

$$\widetilde{X} = (1+r)X.$$

The corresponding risk measure $\widetilde{\rho}(\widetilde{X}) := \rho(X)$ is again monotone. Cash invariance is replaced by the following property:

$$\widetilde{\rho}\big(\widetilde{X} + (1+r)m\big) = \widetilde{\rho}(\widetilde{X}) - m, \tag{4.2}$$

i.e., the risk is reduced by m if an additional amount m is invested in a risk-free manner. Conversely, any $\widetilde{\rho} : \mathcal{X} \to \mathbb{R}$ which is monotone and satisfies (4.2) defines a monetary measure of risk via $\rho(X) := \widetilde{\rho}\big((1+r)X\big)$. ◇

Lemma 4.3. *Any monetary measure of risk ρ is Lipschitz continuous with respect to the supremum norm $\|\cdot\|$:*

$$|\rho(X) - \rho(Y)| \le \|X - Y\|.$$

Proof. Clearly, $X \le Y + \|X - Y\|$, and so $\rho(Y) - \|X - Y\| \le \rho(X)$ by monotonicity and cash invariance. Reversing the roles of X and Y yields the assertion. □

From now on we concentrate on monetary measures of risk which have an additional convexity property.

Definition 4.4. A monetary risk measure $\rho : \mathcal{X} \to \mathbb{R}$ is called a *convex measure of risk* if it satisfies

- *Convexity:* $\rho(\lambda X + (1-\lambda)Y) \le \lambda\rho(X) + (1-\lambda)\rho(Y)$, for $0 \le \lambda \le 1$.

Consider the collection of possible future outcomes that can be generated with the resources available to an investor: One investment strategy leads to X, while a second strategy leads to Y. If one *diversifies*, spending only the fraction λ of the resources on the first possibility and using the remaining part for the second alternative, one obtains $\lambda X + (1-\lambda)Y$. Thus, the axiom of convexity gives a precise meaning to the idea that diversification should not increase the risk. If ρ is convex and normalized, then

$$\begin{aligned} \rho(\lambda X) &\le \lambda\rho(X) \quad \text{for } 0 \le \lambda \le 1, \\ \rho(\lambda X) &\ge \lambda\rho(X) \quad \text{for } \lambda \ge 1. \end{aligned}$$

Definition 4.5. A convex measure of risk ρ is called a *coherent risk measure* if it satisfies

- *Positive Homogeneity:* If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda\rho(X)$.

If a monetary measure of risk ρ is positively homogeneous, then it is normalized, i.e., $\rho(0) = 0$. Under the assumption of positive homogeneity, convexity is equivalent to

- *Subadditivity:* $\rho(X + Y) \leq \rho(X) + \rho(Y)$.

This property allows to decentralize the task of managing the risk arising from a collection of different positions: If separate risk limits are given to different "desks", then the risk of the aggregate position is bounded by the sum of the individual risk limits.

In many situations, however, risk may grow in a *non-linear* way as the size of the position increases. For this reason we will not insist on positive homogeneity. Instead, our focus will be on convex measures of risk.

A monetary measure of risk ρ induces the class

$$\mathcal{A}_\rho := \{ X \in \mathcal{X} \mid \rho(X) \leq 0 \}$$

of positions which are acceptable in the sense that they do not require additional capital. The class $\mathcal{A}_\rho$ will be called the *acceptance set* of ρ. The following two propositions summarize the relations between monetary measures of risk and their acceptance sets.

Proposition 4.6. *Suppose that ρ is a monetary measure of risk with acceptance set $\mathcal{A} := \mathcal{A}_\rho$.*

(a) *$\mathcal{A}$ is non-empty, and satisfies the following two conditions:*

$$\inf\{ m \in \mathbb{R} \mid m \in \mathcal{A} \} > -\infty. \tag{4.3}$$

$$X \in \mathcal{A},\ Y \in \mathcal{X}, Y \geq X \quad \Longrightarrow \quad Y \in \mathcal{A}. \tag{4.4}$$

Moreover, $\mathcal{A}$ has the following closure property: For $X \in \mathcal{A}$ and $Y \in \mathcal{X}$,

$$\left\{ \lambda \in [0,1] \mid \lambda X + (1-\lambda)Y \in \mathcal{A} \right\} \text{ is closed in } [0,1]. \tag{4.5}$$

(b) *ρ can be recovered from $\mathcal{A}$:*

$$\rho(X) = \inf\{ m \in \mathbb{R} \mid m + X \in \mathcal{A} \}. \tag{4.6}$$

(c) *ρ is a* convex *risk measure if and only if $\mathcal{A}$ is convex.*

(d) *ρ is positively homogeneous if and only if $\mathcal{A}$ is a cone. In particular, ρ is coherent if and only if $\mathcal{A}$ is a convex cone.*

Proof. The first two properties in (a) are straightforward. As to (4.5), the function $\lambda \mapsto \rho(\lambda X + (1-\lambda)Y)$ is continuous by Lemma 4.3. Hence, the set of $\lambda \in [0, 1]$ such that $\rho(\lambda X + (1-\lambda)Y) \le 0$ is closed.

(b): Cash invariance implies that for $X \in \mathcal{X}$,

$$\begin{aligned}\inf\{ m \in \mathbb{R} \mid m + X \in \mathcal{A}_\rho \} &= \inf\{ m \in \mathbb{R} \mid \rho(m + X) \le 0 \} \\ &= \inf\{ m \in \mathbb{R} \mid \rho(X) \le m \} \\ &= \rho(X).\end{aligned}$$

(c): $\mathcal{A}$ is clearly convex if ρ is a convex measure of risk. The converse will follow from Proposition 4.7 together with (4.8).

(d): Clearly, positive homogeneity of ρ implies that $\mathcal{A}$ is a cone. The converse follows as in (c). □

Conversely, one can take a given class $\mathcal{A} \subset \mathcal{X}$ of acceptable positions as the primary object. For a position $X \in \mathcal{X}$, we can then define the capital requirement as the minimal amount m for which $m + X$ becomes acceptable:

$$\rho_{\mathcal{A}}(X) := \inf\{ m \in \mathbb{R} \mid m + X \in \mathcal{A} \}. \tag{4.7}$$

Note that, with this notation, (4.6) takes the form

$$\rho_{\mathcal{A}_\rho} = \rho. \tag{4.8}$$

Proposition 4.7. *Assume that $\mathcal{A}$ is a non-empty subset of $\mathcal{X}$ which satisfies* (4.3) *and* (4.4). *Then the functional $\rho_{\mathcal{A}}$ has the following properties:*

(a) *$\rho_{\mathcal{A}}$ is a monetary measure of risk.*

(b) *If $\mathcal{A}$ is a convex set, then $\rho_{\mathcal{A}}$ is a convex measure of risk.*

(c) *If $\mathcal{A}$ is a cone, then $\rho_{\mathcal{A}}$ is positively homogeneous. In particular, $\rho_{\mathcal{A}}$ is a coherent measure of risk if $\mathcal{A}$ is a convex cone.*

(d) *$\mathcal{A}$ is a subset of $\mathcal{A}_{\rho_{\mathcal{A}}}$. If $\mathcal{A}$ satisfies the closure property* (4.5) *then $\mathcal{A} = \mathcal{A}_{\rho_{\mathcal{A}}}$.*

Proof. (a): It is straightforward to verify that $\rho_{\mathcal{A}}$ satisfies cash invariance and monotonicity. We show next that $\rho_{\mathcal{A}}$ takes only finite values. To this end, fix some Y in the non-empty set $\mathcal{A}$. For $X \in \mathcal{X}$ given, there exists a finite number m with $m + X > Y$, because X and Y are both bounded. Then

$$\rho_{\mathcal{A}}(X) - m = \rho_{\mathcal{A}}(m + X) \le \rho_{\mathcal{A}}(Y) \le 0,$$

and hence $\rho_{\mathcal{A}}(X) \le m < \infty$. Note that (4.3) is equivalent to $\rho_{\mathcal{A}}(0) > -\infty$. To show that $\rho_{\mathcal{A}}(X) > -\infty$ for arbitrary $X \in \mathcal{X}$, we take m' such that $X + m' \le 0$ and conclude by monotonicity and cash invariance that $\rho_{\mathcal{A}}(X) \ge \rho_{\mathcal{A}}(0) + m' > -\infty$.

(b): Suppose that $X_1,\ X_2 \in \mathcal{X}$ and that $m_1,\ m_2 \in \mathbb{R}$ are such that $m_i + X_i \in \mathcal{A}$. If $\lambda \in [0, 1]$, then the convexity of $\mathcal{A}$ implies that $\lambda(m_1 + X_1) + (1-\lambda)(m_2 + X_2) \in \mathcal{A}$. Thus, by the cash invariance of $\rho_{\mathcal{A}}$,

$$\begin{aligned}0 &\geq \rho_{\mathcal{A}}\big(\lambda(m_1 + X_1) + (1-\lambda)(m_2 + X_2)\big)\\ &= \rho_{\mathcal{A}}\big(\lambda X_1 + (1-\lambda)X_2\big) - \big(\lambda m_1 + (1-\lambda)m_2\big),\end{aligned}$$

and the convexity of $\rho_{\mathcal{A}}$ follows.

(c): As in the proof of convexity, we obtain that $\rho_{\mathcal{A}}(\lambda X) \leq \lambda\rho_{\mathcal{A}}(X)$ for $\lambda \geq 0$ if $\mathcal{A}$ is a cone. To prove the converse inequality, let $m < \rho_{\mathcal{A}}(X)$. Then $m + X \notin \mathcal{A}$ and hence $\lambda m + \lambda X \notin \mathcal{A}$ for $\lambda \geq 0$. Thus $\lambda m < \rho_{\mathcal{A}}(\lambda X)$, and (c) follows.

(d): The inclusion $\mathcal{A} \subseteq \mathcal{A}_{\rho_{\mathcal{A}}}$ is obvious. Now assume that $\mathcal{A}$ satisfies (4.5). We have to show that $X \notin \mathcal{A}$ implies that $\rho_{\mathcal{A}}(X) > 0$. To this end, take $m > \|X\| = \sup_\omega |X(\omega)|$. By assumption, there exists an $\varepsilon \in (0, 1)$ such that $\varepsilon m + (1-\varepsilon)X \notin \mathcal{A}$. Thus,

$$0 \leq \rho_{\mathcal{A}}\big(\varepsilon\, m + (1-\varepsilon)X\big) = \rho_{\mathcal{A}}\big((1-\varepsilon)X\big) - \varepsilon\, m.$$

Since $\rho_{\mathcal{A}}$ is a monetary measure of risk, Lemma 4.3 shows that

$$\Big|\rho_{\mathcal{A}}\big((1-\varepsilon)X\big) - \rho_{\mathcal{A}}(X)\Big| \leq \varepsilon\,\|X\|.$$

Hence,

$$\rho_{\mathcal{A}}(X) \geq \rho_{\mathcal{A}}\big((1-\varepsilon)X\big) - \varepsilon\,\|X\| \geq \varepsilon\,\big(m - \|X\|\big) > 0. \qquad \square$$

In the following examples, we take $\mathcal{X}$ as the linear space of all bounded measurable functions on some measurable space $(\Omega, \mathcal{F})$, and we denote by $\mathcal{M}_1 = \mathcal{M}_1(\Omega, \mathcal{F})$ the class of all probability measures on $(\Omega, \mathcal{F})$.

Example 4.8. Consider the *worst-case risk measure* $\rho_{\max}$ defined by

$$\rho_{\max}(X) = -\inf_{\omega\in\Omega} X(\omega) \quad \text{for all } X \in \mathcal{X}.$$

The value $\rho_{\max}(X)$ is the least upper bound for the potential loss which can occur in any scenario. The corresponding acceptance set $\mathcal{A}$ is given by the convex cone of all non-negative functions in $\mathcal{X}$. Thus, $\rho_{\max}$ is a coherent measure of risk. It is the most conservative measure of risk in the sense that any normalized monetary risk measure ρ on $\mathcal{X}$ satisfies

$$\rho(X) \leq \rho\big(\inf_{\omega\in\Omega} X(\omega)\big) = \rho_{\max}(X).$$

Note that $\rho_{\max}$ can be represented in the form

$$\rho_{\max}(X) = \sup_{Q\in\mathcal{Q}} E_Q[\,-X\,], \tag{4.9}$$

where $\mathcal{Q}$ is the class $\mathcal{M}_1$ of all probability measures on $(\Omega, \mathcal{F})$. ◇

Example 4.9. Let $\mathcal{Q}$ be a set of probability measures on $(\Omega, \mathcal{F})$, and consider a mapping $\gamma : \mathcal{Q} \to \mathbb{R}$ with $\sup_Q \gamma(Q) < \infty$, which specifies for each $Q \in \mathcal{Q}$ some "floor" $\gamma(Q)$. Suppose that a position X is acceptable if

$$E_Q[\,X\,] \geq \gamma(Q) \quad \text{for all } Q \in \mathcal{Q}.$$

The set $\mathcal{A}$ of such positions satisfies (4.3) and (4.4), and it is convex. Thus, the associated monetary risk measure $\rho = \rho_{\mathcal{A}}$ is *convex*, and it takes the form

$$\rho(X) = \sup_{Q \in \mathcal{Q}} (\gamma(Q) - E_Q[\,X\,]).$$

Alternatively, we can write

$$\rho(X) = \sup_{Q \in \mathcal{M}_1} \big(E_Q[\,-X\,] - \alpha(Q)\big), \tag{4.10}$$

where the *penalty function* $\alpha : \mathcal{M}_1 \to (-\infty, \infty]$ is defined by $\alpha(Q) = -\gamma(Q)$ for $Q \in \mathcal{Q}$ and $\alpha(Q) = +\infty$ otherwise. Note that ρ is a *coherent* risk measure if $\gamma(Q) = 0$ for all $Q \in \mathcal{Q}$. ◇

Example 4.10. Consider a utility function u on $\mathbb{R}$, a probability measure $Q \in \mathcal{M}_1$, and fix some threshold $c \in \mathbb{R}$. Let us call a position X acceptable if its certainty equivalent is at least c, i.e., if its expected utility $E_Q[\,u(X)\,]$ is bounded from below by $u(c)$. Clearly, the set

$$\mathcal{A} := \{\, X \in \mathcal{X} \mid E_Q[\,u(X)\,] \geq u(c)\,\}.$$

is non-empty, convex, and satisfies (4.3) and (4.4). Thus, $\rho_{\mathcal{A}}$ is a convex measure of risk. As an obvious robust extension, we can define acceptability in terms of a whole class $\mathcal{Q}$ of probability measures on $(\Omega, \mathcal{F})$, i.e.,

$$\mathcal{A} := \bigcap_{Q \in \mathcal{Q}} \{\, X \in \mathcal{X} \mid E_Q[\,u(X)\,] \geq u(c_Q)\,\},$$

with constants c_Q such that $\sup_{Q \in \mathcal{Q}} c_Q < \infty$. The corresponding risk measures will be studied in more detail in Section 4.9. ◇

Example 4.11. Suppose now that we have specified a probabilistic model, i.e., a probability measure P on $(\Omega, \mathcal{F})$. In this context, a position X is often considered to be acceptable if the probability of a loss is bounded by a given level $\lambda \in (0, 1)$, i.e., if

$$P[\,X < 0\,] \leq \lambda.$$

The corresponding monetary risk measure $V@R_\lambda$, defined by

$$V@R_\lambda(X) = \inf\{\, m \in \mathbb{R} \mid P[\,m + X < 0\,] \leq \lambda\,\},$$

is called *Value at Risk* at level λ. Note that it is well defined on the space $\mathcal{L}^0(\Omega, \mathcal{F}, P)$ of all random variables which are P-a.s. finite, and that

$$V@R_\lambda(X) = E[-X] + \Phi^{-1}(1-\lambda)\sigma(X), \tag{4.11}$$

if X is a Gaussian random variable with variance $\sigma^2(X)$ and Φ^{-1} denotes the inverse of the distribution function Φ of $N(0,1)$. Clearly, $V@R_\lambda$ is positively homogeneous, but in general it is not convex, as shown by Example 4.41 below. In Section 4.4, Value at Risk will be discussed in detail. In particular, we will study some closely related coherent and convex measures of risk. ◇

Example 4.12. As in the preceding example, we fix a probability measure P on $(\Omega, \mathcal{F})$. For an asset with payoff $\widetilde{X} \in \mathcal{L}^2 = \mathcal{L}^2(\Omega, \mathcal{F}, P)$, price $\pi(\widetilde{X})$, and variance $\sigma^2(\widetilde{X}) \neq 0$, the *Sharpe ratio* is defined as

$$\frac{E[\widetilde{X}] - \pi(\widetilde{X})(1+r)}{\sigma(\widetilde{X})} = \frac{E[X]}{\sigma(X)},$$

where $X := \widetilde{X}(1+r)^{-1} - \pi(\widetilde{X})$ is the corresponding discounted net worth. Suppose that we find the position X acceptable if the Sharpe ratio is bounded from below by some constant $c > 0$. The resulting functional ρ_c on $\mathcal{L}^2$ defined by (4.7) for the class

$$\mathcal{A}_c := \{ X \in \mathcal{L}^2 \mid E[X] \geq c \cdot \sigma(X) \}$$

is given by

$$\rho_c(X) = E[-X] + c \cdot \sigma(X).$$

It is cash invariant and positively homogeneous, and it is convex since $\sigma(\cdot)$ is a convex functional on $\mathcal{L}^2$. But ρ_c is not a monetary risk measure, because it is not monotone. Indeed, if $X = e^Z$ and Z is a random variable with normal distribution $N(0, \sigma^2)$, then $X \geq 0$ but

$$\rho_c(X) = -e^{\sigma^2/2} + ce^{\sigma^2/2}\sqrt{e^{\sigma^2} - 1}$$

becomes positive for large enough σ. Note, however, that (4.11) shows that $\rho_c(X)$ coincides with $V@R_\lambda(X)$ if X is Gaussian and if $c = \Phi^{-1}(1-\lambda)$ with $0 < \lambda \leq 1/2$. Thus, both ρ_c and $V@R_\lambda$ have all the properties of a coherent risk measure if restricted to a Gaussian subspace $\widetilde{\mathcal{X}}$ of $\mathcal{L}^2$, i.e, a linear space consisting of normally distributed random variables. But neither ρ_c nor $V@R_\lambda$ can be coherent on the full space $\mathcal{L}^2$, since the existence of normal random variables on $(\Omega, \mathcal{F}, P)$ implies that $\mathcal{X}$ will also contain random variables as considered in Example 4.41. ◇

Example 4.13. Let $c : \mathcal{F} \to [0,1]$ be any set function which is normalized and monotone in the sense that $c(\emptyset) = 0$, $c(\Omega) = 1$, and $c(A) \leq c(B)$ if $A \subset B$. For instance, c can be given by $c(A) := \psi(P[A])$ for some probability measure P and an increasing function $\psi : [0,1] \to [0,1]$ such that $\psi(0) = 0$ and $\psi(1) = 1$. The

Choquet integral of a bounded measurable function $X \geq 0$ with respect to c is defined as

$$\int X\,dc := \int_0^\infty c(X > x)\,dx.$$

If c is a probability measure, Fubini's theorem implies that $\int X\,dc$ coincides with the usual integral. In the general case, the Choquet integral is a *nonlinear* functional of X, but we still have $\int \lambda X\,dc = \lambda \int X\,dc$ and $\int (X+m)\,dc = \int X\,dc + m$ for constants $\lambda, m \geq 0$. If $X \in \mathcal{X}$ is arbitrary, we take $m \in \mathbb{R}$ such that $X + m \geq 0$ and get

$$\int (X+m)\,dc - m = \int_{-m}^0 \big(c(X > x) - 1\big)\,dx + \int_0^\infty c(X > x)\,dx.$$

The right-hand side is independent of $m \geq -\inf X$, and so it makes sense to extend the definition of the Choquet integral by putting

$$\int X\,dc := \int_{-\infty}^0 \big(c(X > x) - 1\big)\,dx + \int_0^\infty c(X > x)\,dx$$

for all $X \in \mathcal{X}$. It follows that

$$\int \lambda X\,dc = \lambda \int X\,dc \quad \text{and} \quad \int (X+m)\,dc = \int X\,dc + m$$

for all $\lambda \geq 0$ and $m \in \mathbb{R}$. Moreover, we have

$$\int Y\,dc \geq \int X\,dc \quad \text{for } Y \geq X.$$

Thus, the Choquet integral of the loss,

$$\rho(X) := \int (-X)\,dc,$$

is a positively homogeneous monetary risk measure on $\mathcal{X}$. In Section 4.7, we will characterize these risk measures in terms of a property called "comonotonicity". We will also show that ρ is convex, and hence coherent, if and only if c is *submodular* or 2-*alternating*, i.e.,

$$c(A \cap B) + c(A \cup B) \leq c(A) + c(B) \quad \text{for } A, B \in \mathcal{F}.$$

In this case, ρ admits the representation

$$\rho(X) = \max_{Q \in \mathcal{Q}_c} E_Q[\,-X\,], \tag{4.12}$$

where $\mathcal{Q}_c$ is the *core* of c, defined as the class of all finitely additive and normalized set functions $Q : \mathcal{F} \to [0, 1]$ such that $Q[\,A\,] \leq c(A)$ for all $A \in \mathcal{F}$; see Theorem 4.88. ◇

In the next two sections, we are going to show how representations of the form (4.9), (4.12), or (4.10) for coherent or convex risk measures arise in a systematic manner.

4.2 Robust representation of convex risk measures

In this section, we consider the situation where $\mathcal{X}$ consists of all bounded measurable functions on the measurable space $(\Omega, \mathcal{F})$. Recall that $\mathcal{X}$ is a Banach space if endowed with the supremum norm $\|\cdot\|$. As in Section 2.5, we denote by $\mathcal{M}_1 := \mathcal{M}_1(\Omega, \mathcal{F})$ the set of all probability measures on $(\Omega, \mathcal{F})$ and by $\mathcal{M}_{1,f} := \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ the set of all finitely additive set functions $Q : \mathcal{F} \to [0, 1]$ which are normalized to $Q[\,\Omega\,] = 1$. By $E_Q[\,X\,]$ we denote the integral of X with respect to $Q \in \mathcal{M}_{1,f}$; see Appendix A.6. We do not assume that a probability measure on $(\Omega, \mathcal{F})$ is given a priori.

If ρ is a *coherent* measure of risk on $\mathcal{X}$, then we are in the context of Proposition 2.85, i.e., the functional J defined by $J(X) := -\rho(X)$ satisfies the four properties listed in Proposition 2.84. Hence, we have the following result:

Proposition 4.14. *A functional $\rho : \mathcal{X} \to \mathbb{R}$ is a coherent measure of risk if and only if there exists a subset $\mathcal{Q}$ of $\mathcal{M}_{1,f}$ such that*

$$\rho(X) = \sup_{Q \in \mathcal{Q}} E_Q[\,-X\,], \quad X \in \mathcal{X}. \tag{4.13}$$

Moreover, $\mathcal{Q}$ can be chosen as a convex set for which the supremum in (4.13) *is attained.*

Our first goal in this section is to obtain an analogue of this result for *convex* measures of risk. Applied to a coherent measure of risk, it will yield an alternative proof of Proposition 4.14, which does not depend on the discussion in Chapter 2, and it will provide a description of the maximal set $\mathcal{Q}$ in (4.13). Our second goal will be to obtain criteria which guarantee that a measure of risk can be represented in terms of σ-additive probability measures.

Let $\alpha : \mathcal{M}_{1,f} \to \mathbb{R} \cup \{+\infty\}$ be any functional such that

$$\inf_{Q \in \mathcal{M}_{1,f}} \alpha(Q) \in \mathbb{R}.$$

For each $Q \in \mathcal{M}_{1,f}$ the functional $X \mapsto E_Q[\,-X\,] - \alpha(Q)$ is convex, monotone, and cash invariant on $\mathcal{X}$, and these three properties are preserved when taking the supremum over $Q \in \mathcal{M}_{1,f}$. Hence,

$$\rho(X) := \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[\,-X\,] - \alpha(Q)\big) \tag{4.14}$$

defines a convex measure of risk on $\mathcal{X}$ such that

$$\rho(0) = -\inf_{Q \in \mathcal{M}_{1,f}} \alpha(Q).$$

The functional α will be called a *penalty function* for ρ on $\mathcal{M}_{1,f}$, and we will say that ρ is *represented* by α on $\mathcal{M}_{1,f}$.

Theorem 4.15. *Any convex measure of risk ρ on $\mathcal{X}$ is of the form*

$$\rho(X) = \max_{Q \in \mathcal{M}_{1,f}} \big(E_Q[-X] - \alpha_{\min}(Q) \big), \quad X \in \mathcal{X}, \tag{4.15}$$

where the penalty function $\alpha_{\min}$ is given by

$$\alpha_{\min}(Q) := \sup_{X \in \mathcal{A}_\rho} E_Q[-X] \quad \text{for } Q \in \mathcal{M}_{1,f}.$$

Moreover, $\alpha_{\min}$ is the minimal penalty function which represents ρ, i.e., any penalty function α for which (4.14) *holds satisfies $\alpha(Q) \geq \alpha_{\min}(Q)$ for all $Q \in \mathcal{M}_{1,f}$.*

Proof. In a first step, we show that

$$\rho(X) \geq \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[-X] - \alpha_{\min}(Q) \big) \quad \text{for all } X \in \mathcal{X}.$$

To this end, recall that $X' := \rho(X) + X \in \mathcal{A}_\rho$ by (4.1). Thus, for all $Q \in \mathcal{M}_{1,f}$

$$\alpha_{\min}(Q) \geq E_Q[-X'] = E_Q[-X] - \rho(X).$$

From here, our claim follows.

For X given, we will now construct some $Q_X \in \mathcal{M}_{1,f}$ such that

$$\rho(X) \leq E_{Q_X}[-X] - \alpha_{\min}(Q_X),$$

which, in view of the previous step, will prove our representation (4.15). By cash invariance it suffices to prove this for $X \in \mathcal{X}$ with $\rho(X) = 0$. Moreover, we may assume without loss of generality that $\rho(0) = 0$. Then X is not contained in the nonempty convex set

$$\mathcal{B} := \{ Y \in \mathcal{X} \mid \rho(Y) < 0 \}.$$

Since $\mathcal{B}$ is open is open due to Lemma 4.3, we may apply the separation argument in the form of Theorem A.54. It yields a non-zero continuous linear functional ℓ on $\mathcal{X}$ such that

$$\ell(X) \leq \inf_{Y \in \mathcal{B}} \ell(Y) =: b.$$

We claim that $\ell(Y) \geq 0$ if $Y \geq 0$. Monotonicity and cash invariance of ρ imply that $1 + \lambda Y \in \mathcal{B}$ for any $\lambda > 0$. Hence,

$$\ell(X) \leq \ell(1 + \lambda Y) = \ell(1) + \lambda \ell(Y) \quad \text{for all } \lambda > 0,$$

which could not be true if $\ell(Y) < 0$.

Our next claim is that $\ell(1) > 0$. Since ℓ does not vanish identically, there must be some Y such that $0 < \ell(Y) = \ell(Y^+) - \ell(Y^-)$. We may assume without loss of generality that $\|Y\| < 1$. Positivity of ℓ implies $\ell(Y^+) > 0$ and $\ell(1 - Y^+) \geq 0$. Hence $\ell(1) = \ell(1 - Y^+) + \ell(Y^+) > 0$.

By the two preceding steps and Theorem A.50, we conclude that there exists some $Q_X \in \mathcal{M}_{1,f}$ such that

$$E_{Q_X}[\,Y\,] = \frac{\ell(Y)}{\ell(1)} \quad \text{for all } Y \in \mathcal{X}.$$

Note that $\mathcal{B} \subset \mathcal{A}_\rho$, and so

$$\alpha_{\min}(Q_X) = \sup_{Y \in \mathcal{A}_\rho} E_{Q_X}[\,-Y\,] \geq \sup_{Y \in \mathcal{B}} E_{Q_X}[\,-Y\,] = -\frac{b}{\ell(1)}.$$

On the other hand, $Y + \varepsilon \in \mathcal{B}$ for any $Y \in \mathcal{A}_\rho$ and each $\varepsilon > 0$. This shows that $\alpha_{\min}(Q_X)$ is in fact equal to $-b/\ell(1)$. It follows that

$$E_{Q_X}[\,-X\,] - \alpha_{\min}(Q_X) = \frac{1}{\ell(1)}\big(b - \ell(X)\big) \geq 0 = \rho(X).$$

Thus, Q_X is as desired, and the proof of the representation (4.15) is complete.

Finally, let α be any penalty function for ρ. Then, for all $Q \in \mathcal{M}_{1,f}$ and $X \in \mathcal{X}$

$$\rho(X) \geq E_Q[\,-X\,] - \alpha(Q),$$

and hence

$$\begin{aligned} \alpha(Q) &\geq \sup_{X \in \mathcal{X}} \big(E_Q[\,-X\,] - \rho(X)\big) \\ &\geq \sup_{X \in \mathcal{A}_\rho} \big(E_Q[\,-X\,] - \rho(X)\big) \\ &\geq \alpha_{\min}(Q). \end{aligned} \tag{4.16}$$

Thus, α dominates $\alpha_{\min}$. □

Remark 4.16. (a) If we take $\alpha = \alpha_{\min}$ in (4.16), then all inequalities in (4.16) must be identities. Thus, we obtain an alternative formula for the minimal penalty function $\alpha_{\min}$:

$$\alpha_{\min}(Q) = \sup_{X \in \mathcal{X}} \big(E_Q[\,-X\,] - \rho(X)\big). \tag{4.17}$$

(b) Note that $\alpha_{\min}$ is convex and lower semicontinuous for the *total variation distance* on $\mathcal{M}_{1,f}$ as defined in Definition A.49, since it is the supremum of affine continuous functions on $\mathcal{M}_{1,f}$.

(c) Suppose ρ is defined via $\rho := \rho_{\mathcal{A}}$ for a given acceptance set $\mathcal{A} \subset \mathcal{X}$. Then $\mathcal{A}$ determines $\alpha_{\min}$:

$$\alpha_{\min}(Q) = \sup_{X \in \mathcal{A}} E_Q[\,-X\,] \quad \text{for all } Q \in \mathcal{M}_{1,f}.$$

This follows from the fact that $X \in \mathcal{A}$ implies $\varepsilon + X \in \mathcal{A}_\rho$ for all $\varepsilon > 0$. ◇

Remark 4.17. Equation (4.17) shows that the penalty function $\alpha_{\min}$ corresponds to the *Fenchel–Legendre transform*, or *conjugate function*, of the convex function ρ on the Banach space $\mathcal{X}$. More precisely,

$$\alpha_{\min}(Q) = \rho^*(\ell_Q), \tag{4.18}$$

where $\rho^* : \mathcal{X}' \to \mathbb{R} \cup \{+\infty\}$ is defined on the dual $\mathcal{X}'$ of $\mathcal{X}$ by

$$\rho^*(\ell) = \sup_{X \in \mathcal{X}} \big(\ell(X) - \rho(X)\big),$$

and where $\ell_Q \in \mathcal{X}'$ is given by $\ell_Q(X) = E_Q[-X]$ for $Q \in \mathcal{M}_{1,f}$. This suggests an alternative proof of Theorem 4.15. First note that, by Theorem A.50, $\mathcal{X}'$ can be identified with the space $ba := ba(\Omega, \mathcal{F})$ of finitely additive set functions with finite total variation. Moreover, ρ is lower semicontinuous with respect to the weak topology $\sigma(\mathcal{X}, \mathcal{X}')$, since any set $\{\rho \le c\}$ is convex, strongly closed due to Lemma 4.3, and hence weakly closed by Theorem A.59. Thus, the general duality theorem for conjugate functions as stated in Theorem A.61 yields

$$\rho^{**} = \rho,$$

where ρ^{**} denotes the conjugate function of ρ^*, i.e.,

$$\rho(X) = \sup_{\ell \in ba} \big(\ell(X) - \rho^*(\ell)\big). \tag{4.19}$$

In a second step, using the arguments in the second part of the proof of Theorem 4.15, we can now check that monotonicity and cash invariance of ρ imply that $\ell \le 0$ and $\ell(1) = -1$ for any $\ell \in \mathcal{X}' = ba$ such that $\rho^*(\ell) < \infty$. Identifying $-\ell$ with $Q \in \mathcal{M}_{1,f}$ and using equation (4.18), we see that (4.19) reduces to the representation

$$\rho(X) = \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[-X] - \alpha_{\min}(Q)\big).$$

Moreover, the supremum is actually attained: $\mathcal{M}_{1,f}$ is weak* compact in $\mathcal{X}' = ba$ due to the Banach–Alaoglu theorem stated in Theorem A.62, and so the upper semicontinuous functional $Q \mapsto E_Q[-X] - \alpha_{\min}(Q)$ attains its maximum on $\mathcal{M}_{1,f}$. ◇

The representation

$$\rho(X) = \sup_{Q \in \mathcal{Q}} E_Q[-X], \quad X \in \mathcal{X}, \tag{4.20}$$

of a *coherent* measure of risk ρ via some set $\mathcal{Q} \subset \mathcal{M}_{1,f}$ is a particular case of the representation theorem for convex measures of risk, since it corresponds to the penalty function

$$\alpha(Q) = \begin{cases} 0 & \text{if } Q \in \mathcal{Q} \\ +\infty & \text{otherwise.} \end{cases}$$

The following corollary shows that the minimal penalty function of a coherent measure of risk is always of this type.

Corollary 4.18. *The minimal penalty function* $\alpha_{\min}$ *of a coherent measure of risk* ρ *takes only the values* 0 *and* $+\infty$. *In particular,*

$$\rho(X) = \max_{Q \in \mathcal{Q}_{\max}} E_Q[-X], \quad X \in \mathcal{X},$$

for the convex set

$$\mathcal{Q}_{\max} := \{ Q \in \mathcal{M}_{1,f} \mid \alpha_{\min}(Q) = 0 \},$$

and $\mathcal{Q}_{\max}$ *is the largest set for which a representation of the form* (4.20) *holds.*

Proof. Recall from Proposition 4.6 that the acceptance set $\mathcal{A}_\rho$ of a coherent measure of risk is a cone. Thus, the minimal penalty function satisfies

$$\alpha_{\min}(Q) = \sup_{X \in \mathcal{A}_\rho} E_Q[-X] = \sup_{\lambda X \in \mathcal{A}_\rho} E_Q[-\lambda X] = \lambda\, \alpha_{\min}(Q)$$

for all $Q \in \mathcal{M}_{1,f}$ and $\lambda > 0$. Hence, $\alpha_{\min}$ can take only the values 0 and $+\infty$. □

The penalty function α arising in the representation (4.14) is not unique, and it is often convenient to represent a convex measure of risk by a penalty function that is not the minimal one. For instance, the minimal penalty function may be finite for certain finitely additive set functions while another α is concentrated only on probability measures as in the case of Example 4.8. Another situation of this type occurs for risk measures which are constructed as the supremum of a family of convex measures of risk:

Proposition 4.19. *Suppose that for every* i *in some index set* I *we are given a convex measure of risk* ρ_i *on* $\mathcal{X}$ *with associated penalty function* α_i. *If* $\sup_{i\in I} \rho_i(0) < \infty$ *then*

$$\rho(X) := \sup_{i \in I} \rho_i(X), \quad X \in \mathcal{X},$$

is a convex measure of risk that can be represented with the penalty function

$$\alpha(Q) := \inf_{i \in I} \alpha_i(Q), \quad Q \in \mathcal{M}_{1,f}.$$

Proof. The condition $\rho(0) = \sup_{i\in I} \rho_i(0) < \infty$ implies that ρ takes only finite values. Moreover,

$$\begin{aligned}\rho(X) &= \sup_{i \in I} \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[-X] - \alpha_i(Q)\big)\\ &= \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[-X] - \inf_{i \in I} \alpha_i(Q)\big),\end{aligned}$$

and the assertion follows. □

In the sequel, we are particularly interested in those convex measures of risk which admit a representation in terms of σ*-additive probability measures*. Such a risk measure ρ can be represented by a penalty function α which is infinite outside the set $\mathcal{M}_1 := \mathcal{M}_1(\Omega, \mathcal{F})$:

$$\rho(X) = \sup_{Q \in \mathcal{M}_1} \big(E_Q[\,-X\,] - \alpha(Q) \big). \tag{4.21}$$

In this case, one can no longer expect that the supremum above is attained. This is illustrated by Example 4.8 if X does not take on its infimum.

A representation (4.21) in terms of probability measures is closely related to certain continuity properties of ρ. We first examine a necessary condition of "continuity from above".

Lemma 4.20. *A convex measure of risk ρ which admits a representation* (4.21) *on $\mathcal{M}_1$ is continuous from above in the sense that*

$$X_n \searrow X \quad \Longrightarrow \quad \rho(X_n) \nearrow \rho(X). \tag{4.22}$$

Moreover, continuity from above is equivalent to lower semicontinuity with respect to bounded pointwise convergence: If (X_n) is a bounded sequence in $\mathcal{X}$ which converges pointwise to $X \in \mathcal{X}$, then

$$\rho(X) \le \liminf_{n \uparrow \infty} \rho(X_n). \tag{4.23}$$

Proof. First we show (4.23) under the assumption that ρ has a representation in terms of probability measures. Dominated convergence implies that $E_Q[\,X_n\,] \to E_Q[\,X\,]$ for each $Q \in \mathcal{M}_1$. Hence,

$$\begin{aligned} \rho(X) &= \sup_{Q \in \mathcal{M}_1} \big(\lim_{n \uparrow \infty} E_Q[\,-X_n\,] - \alpha(Q) \big) \\ &\le \liminf_{n \uparrow \infty} \sup_{Q \in \mathcal{M}_1} \big(E_Q[\,-X_n\,] - \alpha(Q) \big) \\ &= \liminf_{n \uparrow \infty} \rho(X_n). \end{aligned}$$

In order to show the equivalence of (4.23) and (4.22), let us first assume (4.23). By monotonicity, $\rho(X_n) \le \rho(X)$ for each n if $X_n \searrow X$, and so $\rho(X_n) \nearrow \rho(X)$ follows.

Now we assume continuity from above. Let (X_n) be a bounded sequence in $\mathcal{X}$ which converges pointwise to X. Define $Y_m := \sup_{n \ge m} X_n \in \mathcal{X}$. Then Y_m decreases P-a.s. to X. Since $\rho(X_n) \ge \rho(Y_n)$ by monotonicity, condition (4.22) yields that

$$\liminf_{n \uparrow \infty} \rho(X_n) \ge \lim_{n \uparrow \infty} \rho(Y_n) = \rho(X). \qquad \square$$

The following proposition gives a strong sufficient condition which guarantees that any penalty function for ρ is concentrated on the set $\mathcal{M}_1$ of probability measures. This condition is "continuity from below" rather than from above; we will see a class of examples in Section 4.9.

Proposition 4.21. *Let ρ be a convex measure of risk which is* continuous from below *in the sense that*

$$X_n \nearrow X \quad \Longrightarrow \quad \rho(X_n) \searrow \rho(X),$$

and suppose that α is any penalty function on $\mathcal{M}_{1,f}$ representing ρ. Then α is concentrated on the class $\mathcal{M}_1$ of probability measures, i.e.,

$$\alpha(Q) < \infty \quad \Longrightarrow \quad Q \text{ is } \sigma\text{-additive}.$$

Proof. Recall that Q is σ-additive if and only if $Q[\,A_n\,] \nearrow 1$ for any increasing sequence of events $A_n \in \mathcal{F}$ such that $\bigcup_n A_n = \Omega$. Thus, the assertion is implied by Lemma 4.22 below if we take $X_n := I_{A_n}$. □

Lemma 4.22. *Let ρ be a convex measure of risk on $\mathcal{X}$ which is represented by the penalty function α on $\mathcal{M}_{1,f}$, and consider the level sets*

$$\Lambda_c := \{\, Q \in \mathcal{M}_{1,f} \mid \alpha(Q) \le c \,\}, \quad \text{for } c > -\rho(0) = \inf_{Q \in \mathcal{M}_{1,f}} \alpha(Q).$$

For any sequence (X_n) in $\mathcal{X}$ such that $0 \le X_n \le 1$, the following two conditions are equivalent:

(a) $\rho(\lambda X_n) \longrightarrow \rho(\lambda)$ *for each* $\lambda \ge 1$.

(b) $\inf_{Q \in \Lambda_c} E_Q[\,X_n\,] \longrightarrow 1$ *for all* $c > -\rho(0)$.

Proof. (a) ⇒(b): In a first step, we show that for all $Y \in \mathcal{X}$

$$\inf_{Q \in \Lambda_c} E_Q[\,Y\,] \ge -\frac{c + \rho(\lambda Y)}{\lambda} \quad \text{for all } \lambda > 0. \tag{4.24}$$

Indeed, since α represents ρ, we have for $Q \in \Lambda_c$

$$c \ge \alpha(Q) \ge E_Q[\,-\lambda Y\,] - \rho(\lambda Y),$$

and dividing by $-\lambda$ yields (4.24).

Now consider a sequence (X_n) which satisfies (a). Then (4.24) shows that for all $\lambda \ge 1$

$$\liminf_{n \uparrow \infty} \inf_{Q \in \Lambda_c} E_Q[\,X_n\,] \ge -\lim_{n \uparrow \infty} \frac{c + \rho(\lambda X_n)}{\lambda} = 1 - \frac{c + \rho(0)}{\lambda}.$$

Taking $\lambda \uparrow \infty$ and assuming $X_n \le 1$ proves (b).

(b) ⇒(a): Clearly, for all n

$$\rho(\lambda) \le \rho(\lambda X_n) = \sup_{Q \in \mathcal{M}_{1,f}} \big(E_Q[\,-\lambda X_n\,] - \alpha(Q) \big).$$

Since $E_Q[\,-\lambda X_n\,] \le 0$ for all Q, only those Q can contribute to the supremum on the right-hand side for which

$$\alpha(Q) \le 1 - \rho(\lambda) = 1 + \lambda - \rho(0) =: c.$$

Hence, for all n

$$\rho(\lambda X_n) = \sup_{Q \in \Lambda_c} \big(E_Q[\,-\lambda X_n\,] - \alpha(Q) \big).$$

But condition (b) implies that $E_Q[\,-\lambda X_n\,]$ converges to $-\lambda$ uniformly in $Q \in \Lambda_c$, and so (a) follows. □

Remark 4.23. Let ρ be a convex measure of risk which is continuous from below. Then ρ is also continuous from above, as can be seen by combining Proposition 4.21 and Lemma 4.20. Thus, a straightforward argument yields that $\rho(X_n) \longrightarrow \rho(X)$ whenever (X_n) is a bounded sequence in $\mathcal{X}$ which converges pointwise to X. ◇

Example 4.24. Let us consider a utility function u on $\mathbb{R}$, a probability measure $Q \in \mathcal{M}_1(\Omega, \mathcal{F})$, and fix some threshold $c \in \mathbb{R}$. As in Example 4.10, we suppose that a position X is acceptable if its expected utility $E_Q[\,u(X)\,]$ is bounded from below by $u(c)$. Alternatively, we can introduce the convex increasing loss function $\ell(x) = -u(-x)$ and define the convex set of acceptable positions

$$\mathcal{A} := \{\, X \in \mathcal{X} \mid E_Q[\,\ell(-X)\,] \le x_0 \,\},$$

where $x_0 := -u(c)$. Let $\rho := \rho_{\mathcal{A}}$ denote the convex measure of risk induced by $\mathcal{A}$. In Section 4.9, we will show that ρ is continuous from below, and we will derive a formula for its minimal penalty function. ◇

Let us now continue the discussion in a topological setting. More precisely, we will assume for the rest of this section that Ω is a separable metric space and that $\mathcal{F}$ is the σ-field of Borel sets. As before, $\mathcal{X}$ is the linear space of all bounded measurable functions on $(\Omega, \mathcal{F})$. We denote by $C_b(\Omega)$ the subspace of bounded continuous functions on Ω, and we focus on the representation of convex risk measures viewed as functionals on $C_b(\Omega)$.

Proposition 4.25. *Let ρ be a convex measure of risk on $\mathcal{X}$ such that*

$$\rho(X_n) \searrow \rho(\lambda) \text{ for any sequence } (X_n) \text{ in } C_b(\Omega) \text{ that increases to a constant } \lambda > 0. \tag{4.25}$$

Then there exists a penalty function α on $\mathcal{M}_1$ such that

$$\rho(X) = \max_{Q \in \mathcal{M}_1} \big(E_Q[\,-X\,] - \alpha(Q) \big) \quad \text{for } X \in C_b(\Omega). \tag{4.26}$$

In fact, one can take

$$\alpha(Q) := \inf \big\{\, \alpha_{\min}(\widetilde{Q}) \mid E_{\widetilde{Q}}[\,\cdot\,] = E_Q[\,\cdot\,] \text{ on } C_b(\Omega) \,\big\}. \tag{4.27}$$

Proof. Let $\alpha_{\min}$ be the minimal penalty function of ρ on $\mathcal{M}_{1,f}$. We show that for any $\widetilde{Q}$ with $\alpha_{\min}(\widetilde{Q}) < \infty$ there exists $Q \in \mathcal{M}_1$ such that $E_{\widetilde{Q}}[\,X\,] = E_Q[\,X\,]$ for all $X \in C_b(\Omega)$. Take a sequence (Y_n) in $C_b(\Omega)$ which increases to some $Y \in C_b(\Omega)$, and choose $\delta > 0$ such that $X_n := 1 + \delta(Y_n - Y) \geq 0$ for all n. Clearly, (X_n) satisfies condition (a) of Lemma 4.22, and so $E_{\widetilde{Q}}[\,X_n\,] \to 1$, i.e.,

$$E_{\widetilde{Q}}[\,Y_n\,] \nearrow E_{\widetilde{Q}}[\,Y\,].$$

This continuity property of the linear functional $E_{\widetilde{Q}}[\,\cdot\,]$ on $C_b(\Omega)$ implies, via the Daniell–Stone representation theorem as stated in Appendix A.6, that it coincides on $C_b(\Omega)$ with the integral with respect to a σ-additive measure Q. Taking α as in (4.27) gives the result. □

Remark 4.26. If Ω is *compact* then *any* convex risk measure admits a representation (4.26) on the space $C_b(\Omega) = C(\Omega)$. In order to see that condition (4.25) is verified, recall *Dini's lemma:* On a compact set, a sequence of continuous functions X_n increasing to a continuous function X converges even uniformly. Indeed, the compact sets $K_n := \{X_n \geq X - \varepsilon\}$ satisfy $\bigcap_n K_n = \emptyset$, hence $K_{n_0} = \emptyset$ for some n_0. Since ρ is Lipschitz continuous on $C(\Omega)$ by Lemma 4.3, it satisfies condition (4.25).

Alternatively, we could argue as in Remark 4.17 and apply the general duality theorem for the Fenchel–Legendre transform to the convex functional ρ on the Banach space $C(\Omega)$. Just note that any continuous functional ℓ on $C(\Omega)$ which is positive and normalized is of the form $\ell(X) = E_Q[\,X\,]$ for some probability measure $Q \in \mathcal{M}_1$; see Theorem A.47. ◇

Definition 4.27. A convex risk measure ρ on $\mathcal{X}$ is called *tight* if there exists an increasing sequence $K_1 \subset K_2 \subset \cdots$ of compact subsets of Ω such that

$$\rho(\lambda I_{K_n}) \longrightarrow \rho(\lambda) \quad \text{for all } \lambda \geq 1.$$

Note that every convex measure of risk is tight if Ω is compact.

Proposition 4.28. *Suppose that the convex risk measure ρ on $\mathcal{X}$ is tight. Then* (4.25) *holds and the conclusion of Proposition* 4.25 *is valid. Moreover, if Ω is a Polish space and α is a penalty function on $\mathcal{M}_1$ such that*

$$\rho(X) = \sup_{Q \in \mathcal{M}_1} \big(E_Q[\,-X\,] - \alpha(Q) \big) \quad \text{for } X \in C_b(\Omega),$$

then the level sets $\Lambda_c = \{\, Q \in \mathcal{M}_1 \mid \alpha(Q) \leq c \,\}$ are relatively compact for the weak topology on $\mathcal{M}_1$.

Proof. First we show (4.25). Suppose $X_n \in C_b(\Omega)$ are such that $X_n \nearrow \lambda > 0$. We may assume without loss of generality that ρ is normalized. Convexity and normalization guarantee that condition (4.25) holds for all $\lambda > 0$ as soon as it holds for all $\lambda \geq c$ where c is an arbitrary constant larger than 1. Hence, the cash invariance of ρ implies that there is no loss of generality in assuming $X_n \geq 0$ for all n. We must show that $\rho(X_n) \leq \rho(\lambda) + 2\varepsilon$ eventually, where we take $\varepsilon \in (0, \lambda - 1)$.

By assumption, there exists a compact set K_N such that

$$\rho\big((\lambda - \varepsilon)\mathrm{I}_{K_N}\big) \leq \rho(\lambda - \varepsilon) + \varepsilon = \rho(\lambda) + 2\varepsilon.$$

By Dini's lemma as recalled in Remark 4.26, there exists some $n_0 \in \mathbb{N}$ such that $\lambda - \varepsilon \leq X_n$ on K_N for all $n \geq n_0$. Finally, monotonicity implies

$$\rho(X_n) \leq \rho\big((\lambda - \varepsilon)\mathrm{I}_{K_N}\big) \leq \rho(\lambda) + 2\varepsilon.$$

To prove the relative compactness of Λ_c, we will show that for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset \Omega$ such that for all $c > -\rho(0)$

$$\inf_{Q \in \widetilde{\Lambda}_c} Q[\, K_\varepsilon \,] \geq 1 - \varepsilon(c + \rho(0) + 1).$$

The relative compactness of Λ_c will then be an immediate consequence of Prohorov's characterization of weakly compact sets in $\mathcal{M}_1$, as stated in Theorem A.41. We fix a countable dense set $\{\omega_1, \omega_2, \dots\} \subset \Omega$ and a complete metric δ which generates the topology of Ω. For $r > 0$ we define continuous functions Δ_i^r on Ω by

$$\Delta_i^r(\omega) := 1 - \frac{\delta(\omega, \omega_i) \wedge r}{r}.$$

The function Δ_i^r is dominated by the indicator function of the closed metric ball

$$\overline{B}_r(\omega_i) := \{\, \omega \in \Omega \mid \delta(\omega, \omega_i) \leq r \,\}.$$

Let

$$X_n^r(\omega) := \max_{i \leq n} \Delta_i^r(\omega).$$

Clearly, X_n^r is continuous and satisfies $0 \leq X_n^r \leq 1$ as well as $X_n^r \nearrow 1$ for $n \uparrow \infty$.

According to (4.24), we have for all $\lambda > 0$

$$\inf_{Q \in \Lambda_c} Q\Big[\, \bigcup_{i=1}^{n} \overline{B}_r(\omega_i) \Big] \geq \inf_{Q \in \Lambda_c} E_Q[\, X_n^r \,] \geq -\frac{c + \rho(\lambda X_n^r)}{\lambda}.$$

Now we take $\lambda_k := 2^k/\varepsilon$ and $r_k := 1/k$. The first part of this proof and (4.25) yield the existence of $n_k \in \mathbb{N}$ such that

$$\rho\big(\lambda_k X_{n_k}^{r_k}\big) \leq \rho(\lambda_k) + 1 = -\lambda_k + 1,$$

and thus

$$\sup_{Q \in \Lambda_c} Q\Big[\, \bigcap_{i=1}^{n_k} \Omega \backslash \overline{B}_{r_k}(\omega_i) \Big] \leq \frac{c + 1}{\lambda_k} = \varepsilon 2^{-k}(c + 1).$$

We let

$$K_\varepsilon := \bigcap_{k=1}^{\infty} \bigcup_{i=1}^{n_k} \overline{B}_{r_k}(\omega_i).$$

Then, for each $Q \in \Lambda_c$

$$\begin{aligned} Q[\,K_\varepsilon\,] &= 1 - Q\Big[\,\bigcup_{k=1}^{\infty} \bigcap_{i=1}^{n_k} \Omega \backslash \overline{B}_{r_k}(\omega_i)\,\Big] \\ &\geq 1 - \sum_{k=1}^{\infty} \varepsilon 2^{-k}(c+1) \\ &= 1 - \varepsilon(c+1). \end{aligned}$$

The reader may notice that K_ε is closed, totally bounded and, hence, compact. A short proof of this fact goes as follows: Let (x_j) be a sequence in K_ε. We must show that (x_j) has a convergent subsequence. Since K_ε is covered by $\overline{B}_{r_k}(\omega_1), \dots, \overline{B}_{r_k}(\omega_{n_k})$ for each k, there exists some $i_k \leq n_k$ such that infinitely many x_j are contained in $\overline{B}_{r_k}(\omega_{i_k})$. A diagonalization argument yields a single subsequence $(x_{j'})$ which for each k is contained in some $\overline{B}_{r_k}(\omega_{i_k})$. Thus, $(x_{j'})$ is a Cauchy sequence with respect to the complete metric δ and, hence, converging to some element $\omega \in \Omega$. □

Remark 4.29. Note that the representation (4.26) does not necessarily extend from $C_b(\Omega)$ to the space $\mathcal{X}$ of all bounded measurable functions. Suppose in fact that Ω is compact but not finite, so that condition (4.25) holds as explained in Remark 4.26. There is a finitely additive $Q_0 \in \mathcal{M}_{1,f}$ which does not belong to $\mathcal{M}_1$; see Example A.52. The proof of Proposition 4.25 shows that there is some $\widetilde{Q} \in \mathcal{M}_1$ such that the coherent risk measure ρ defined by $\rho(X) := E_{Q_0}[\,-X\,]$ coincides with $E_{\widetilde{Q}}[\,-X\,]$ for $X \in C_b(\Omega)$. But ρ does not admit a representation of the form

$$\rho(X) = \sup_{Q \in \mathcal{M}_1} \big(E_Q[\,-X\,] - \alpha(Q)\big) \quad \text{for all } X \in \mathcal{X}.$$

In fact, this would imply

$$\alpha(Q) \geq E_{Q_0}[\,X\,] - E_Q[\,X\,]$$

for $Q \in \mathcal{M}_1$ and any $X \in \mathcal{X}$, hence $\alpha(Q) = \infty$ for any $Q \in \mathcal{M}_1$. ◇

4.3 Convex risk measures on L^∞

For the rest of this chapter, we fix a probability measure P on $(\Omega, \mathcal{F})$ and consider risk measures ρ such that

$$\rho(X) = \rho(Y) \quad \text{if } X = Y \ P\text{-a.s.} \tag{4.28}$$

Note that only the nullsets of P will matter in this section.

Lemma 4.30. *Let ρ be a convex measure of risk that satisfies* (4.28) *and which is represented by a penalty function α as in* (4.14). *Then $\alpha(Q) = +\infty$ for any $Q \in \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ which is not absolutely continuous with respect to P.*

Proof. If $Q \in \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ is not absolutely continuous with respect to P, then there exists $A \in \mathcal{F}$ such that $Q[\,A\,] > 0$ but $P[\,A\,] = 0$. Take any $X \in \mathcal{A}_\rho$, and define $X_n := X - n\,\mathrm{I}_A$. Then $\rho(X_n) = \rho(X)$, i.e., X_n is again contained in $\mathcal{A}_\rho$. Hence,

$$\alpha(Q) \ge \alpha_{\min}(Q) \ge E_Q[\,-X_n\,] = E_Q[\,-X\,] + n\,Q[\,A\,] \longrightarrow \infty$$

as $n \uparrow \infty$. □

In view of (4.28), we can identify $\mathcal{X}$ with the Banach space $L^\infty := L^\infty(\Omega, \mathcal{F}, P)$. Let us denote by

$$\mathcal{M}_1(P) := \mathcal{M}_1(\Omega, \mathcal{F}, P)$$

the set of all probability measures on $(\Omega, \mathcal{F})$ which are absolutely continuous with respect to P. The following theorem characterizes those convex risk measures on L^∞ that can be represented by a penalty function concentrated on probability measures, and hence on $\mathcal{M}_1(P)$, due to Lemma 4.30.

Theorem 4.31. *Suppose $\rho : L^\infty \to \mathbb{R}$ is a convex measure of risk. Then the following conditions are equivalent.*

(a) *ρ can be represented by some penalty function on $\mathcal{M}_1(P)$.*

(b) *ρ can be represented by the restriction of the minimal penalty function $\alpha_{\min}$ to $\mathcal{M}_1(P)$:*

$$\rho(X) = \sup_{Q \in \mathcal{M}_1(P)} \big(E_Q[\,-X\,] - \alpha_{\min}(Q)\big), \quad X \in L^\infty. \tag{4.29}$$

(c) *ρ is continuous from above: If $X_n \searrow X$ P-a.s. then $\rho(X_n) \nearrow \rho(X)$.*

(d) *ρ has the "Fatou property": For any bounded sequence (X_n) which converges P-a.s. to some X,*

$$\rho(X) \le \liminf_{n \uparrow \infty} \rho(X_n).$$

(e) *ρ is lower semicontinuous for the weak* topology $\sigma(L^\infty, L^1)$.*

(f) *The acceptance set $\mathcal{A}_\rho$ of ρ is weak* closed in L^∞, i.e., $\mathcal{A}_\rho$ is closed with respect to the topology $\sigma(L^\infty, L^1)$.*

Proof. (f)$\Rightarrow$(b): We fix some $X \in \mathcal{X}$ and let

$$m = \sup_{Q \in \mathcal{M}_1(P)} \big(E_Q[\,-X\,] - \alpha_{\min}(Q) \big). \tag{4.30}$$

In view of Theorem 4.15, we need to show that $m \geq \rho(X)$ or, equivalently, that $m + X \in \mathcal{A}_\rho$. Suppose by way of contradiction that $m + X \notin \mathcal{A}_\rho$. Since the non-empty convex set $\mathcal{A}_\rho$ is weak* closed by assumption, we may apply Theorem A.56 in the locally convex space $(L^\infty, \sigma(L^\infty, L^1))$ with $\mathcal{C} := \mathcal{A}_\rho$ and $\mathcal{B} := \{m + X\}$. We obtain a continuous linear functional ℓ on $(L^\infty, \sigma(L^\infty, L^1))$ such that

$$\beta := \inf_{Y \in \mathcal{A}_\rho} \ell(Y) > \ell(m + X) =: \gamma > -\infty. \tag{4.31}$$

By Proposition A.58, ℓ is of the form $\ell(Y) = E[\,YZ\,]$ for some $Z \in L^1$. In fact, $Z \geq 0$. To show this, fix $Y \geq 0$ and note that $\rho(\lambda Y) \leq \rho(0)$ for $\lambda \geq 0$, by monotonicity. Hence $\lambda Y + \rho(0) \in \mathcal{A}_\rho$ for all $\lambda \geq 0$. It follows that

$$-\infty < \gamma < \ell\big(\lambda Y + \rho(0)\big) = \lambda \ell(Y) + \ell\big(\rho(0)\big).$$

Taking $\lambda \uparrow \infty$ yields that $\ell(Y) \geq 0$ and in turn that $Z \geq 0$. Moreover, $P[\,Z > 0\,] > 0$ since ℓ is non-zero. Thus,

$$\frac{dQ_0}{dP} := \frac{Z}{E[\,Z\,]}$$

defines a probability measure $Q_0 \in \mathcal{M}_1(P)$. By (4.31), we see that

$$\alpha_{\min}(Q_0) = \sup_{Y \in \mathcal{A}_\rho} E_{Q_0}[\,-Y\,] = -\frac{\beta}{E[\,Z\,]}.$$

However,

$$E_{Q_0}[\,X\,] + m = \frac{\ell(m + X)}{E[\,Z\,]} = \frac{\gamma}{E[\,Z\,]} < \frac{\beta}{E[\,Z\,]} = -\alpha_{\min}(Q_0),$$

in contradiction to (4.30). Hence, $m+X$ must be contained in $\mathcal{A}_\rho$, and thus $m \geq \rho(X)$.

(b) $\Rightarrow$(a) is obvious, and (a) $\Rightarrow$(c)$\Leftrightarrow$(d) follows as in Lemma 4.20, replacing pointwise convergence by P-a.s. convergence.

(c) $\Rightarrow$(e): We have to show that $\mathcal{C} := \{\rho \leq c\}$ is weak* closed for $c \in \mathbb{R}$. To this end, let $\mathcal{C}_r := \mathcal{C} \cap \{\,X \in L^\infty \mid \|X\|_\infty \leq r\,\}$ for $r > 0$. If (X_n) is a sequence in $\mathcal{C}_r$ converging in L^1 to some random variable X, then there is a subsequence that converges P-a.s., and the Fatou property of ρ implies that $X \in \mathcal{C}_r$. Hence, $\mathcal{C}_r$ is closed in L^1, and Lemma A.64 implies that $\mathcal{C} := \{\rho \leq c\}$ is weak* closed.

(e) $\Rightarrow$(f) is obvious. □

Definition 4.32. A convex measure of risk ρ on L^∞ is called *sensitive* with respect to P if

$$\rho(-X) > \rho(0)$$

for all $X \in L^\infty_+$ such that $P[\,X > 0\,] > 0$.

Sensitivity is also called *relevance*.

The theorem shows that any convex measure of risk of L^∞ that is continuous from above arises in the following manner. We consider any probabilistic model $Q \in \mathcal{M}_1(P)$, but these models are taken more or less seriously as described by the penalty function. Thus, the value $\rho(X)$ is computed as the worst case, over all models $Q \in \mathcal{M}_1(P)$, of the expected loss $E_Q[-X]$, but reduced by $\alpha(Q)$. In the following example, the given model P is the one which is taken most seriously, and the penalty function $\alpha(Q)$ is proportional to the deviation of Q from P, measured by the relative entropy.

Example 4.33. Consider the penalty function $\alpha : \mathcal{M}_1(P) \to (0, \infty]$ defined by

$$\alpha(Q) := \frac{1}{\beta} H(Q|P),$$

where $\beta > 0$ is a given constant and

$$H(Q|P) = E_Q\Big[\log \frac{dQ}{dP}\Big]$$

is the relative entropy of $Q \in \mathcal{M}_1(P)$ with respect to P; see Definition 3.20. The corresponding *entropic risk measure* ρ is given by

$$\rho(X) = \sup_{Q \in \mathcal{M}_1(P)} \Big(E_Q[-X] - \frac{1}{\beta} H(Q|P) \Big).$$

The variational principle for the relative entropy as stated in Lemma 3.29 shows that

$$E_Q[-X] - \frac{1}{\beta} H(Q|P) \le \frac{1}{\beta} \log E[e^{-\beta X}],$$

and the upper bound is attained by the measure with the density $e^{-\beta X}/E[e^{-\beta X}]$. Thus, the entropic risk measure takes the form

$$\rho(X) = \frac{1}{\beta} \log E[e^{-\beta X}].$$

In particular, ρ is sensitive with respect to P. Note that α is in fact the minimal penalty function representing ρ, since Lemma 3.29 implies

$$\alpha_{\min}(Q) = \sup_{X \in L^\infty} \Big(E_Q[-X] - \frac{1}{\beta} \log E[e^{-\beta X}] \Big) = \frac{1}{\beta} H(Q|P).$$

A financial interpretation of the entropic risk measure in terms of shortfall risk will be discussed in Example 4.105. ◇

Theorem 4.31 takes the following form for *coherent* measures of risk; the proof is the same as the one for Corollary 4.18.

Corollary 4.34. *A coherent measure of risk on L^∞ can be represented by a set $\mathcal{Q} \subset \mathcal{M}_1(P)$ if and only if the equivalent conditions of Theorem* 4.31 *are satisfied. In this case, the maximal representing subset of $\mathcal{M}_1(P)$ is given by*

$$\mathcal{Q}_{\max} := \{ Q \in \mathcal{M}_1(P) \mid \alpha_{\min}(Q) = 0 \}.$$

Moreover, ρ is sensitive if and only if $\mathcal{Q}_{\max} \approx P$ in the sense that for any $A \in \mathcal{F}$

$$P[A] = 0 \quad \Longleftrightarrow \quad Q[A] = 0 \quad \textit{for all } Q \in \mathcal{Q}_{\max}.$$

Let us also state a characterization of those coherent measures of risk on L^∞ which are continuous from below.

Corollary 4.35. *For a coherent measure of risk ρ on L^∞ the following properties are equivalent:*

(a) *ρ is continuous from below: $X_n \nearrow X \Longrightarrow \rho(X_n) \searrow \rho(X)$.*

(b) *There exists a set $\mathcal{Q} \subset \mathcal{M}_1(P)$ representing ρ such that the supremum is attained:*
$$\rho(X) = \max_{Q \in \mathcal{Q}} E_Q[-X] \quad \textit{for all } X \in \mathcal{X}.$$

(c) *There exists a set $\mathcal{Q} \subset \mathcal{M}_1(P)$ representing ρ such that the set of densities*
$$\mathcal{D} := \left\{ \frac{dQ}{dP} \;\middle|\; Q \in \mathcal{Q} \right\}$$
is weakly compact in $L^1(\Omega, \mathcal{F}, P)$.

Proof. (c) ⇒(a): This follows from Dini's lemma; see Remark 4.26.

(a) ⇒(b) follows from Corollary 4.18 and Proposition 4.21.

(b) ⇒(c): Without loss of generality, we can assume that $\mathcal{D}$ is weakly closed in L^1. For any $X \in L^\infty$, the continuous linear functional J_X on L^1 defined by

$$J_X(Z) := E[XZ]$$

attains its infimum on $\mathcal{D}$. According to James' theorem as stated in Appendix A.7, this implies weak compactness of $\mathcal{D}$. □

We now give examples of coherent measures of risk which will be studied in more detail in Section 4.4.

Example 4.36. In our present context, where we require condition (4.28), the *worst-case risk measure* takes the form

$$\rho_{\max}(X) := -\operatorname{ess\,inf} X = \inf \left\{ m \in \mathbb{R} \mid X + m \geq 0 \ P\text{-a.s.} \right\}.$$

One can easily check that $\rho_{\max}$ is coherent and satisfies the Fatou property. Moreover, the acceptance set of $\rho_{\max}$ is equal to the positive cone L^∞_+ in L^∞, and this implies $\alpha_{\min}(Q) = 0$ for any $Q \in \mathcal{M}_1(P)$. Thus,

$$\rho_{\max}(X) = \sup_{Q \in \mathcal{M}_1(P)} E_Q[\,-X\,].$$

Note however that the supremum on the right cannot be replaced by a maximum in case $(\Omega, \mathcal{F}, P)$ cannot be reduced to a finite model. Indeed, let $X \in L^\infty$ be such that X does not attain its essential infimum. Then there can be no $Q \in \mathcal{M}_1(P)$ such that $E_Q[\,X\,] = \operatorname{ess\,inf} X = -\rho_{\max}(X)$. In this case, the preceding corollary shows that $\rho_{\max}$ is not continuous from below. ◇

Example 4.37. Let $\mathcal{Q}_\lambda$ be the class of all $Q \in \mathcal{M}_1(P)$ whose density dQ/dP is bounded by $1/\lambda$ for some fixed parameter $\lambda \in (0, 1)$. The corresponding coherent risk measure

$$AV@R_\lambda(X) := \sup_{Q \in \mathcal{Q}_\lambda} E_Q[\,-X\,] \tag{4.32}$$

will be called the *Average Value at Risk* at level λ. This terminology will become clear in Section 4.4, which contains a detailed study of $AV@R_\lambda$. Note that the set of densities dQ/dP for $Q \in \mathcal{Q}_\lambda$ is weakly closed in L^1. Moreover, it is weakly compact due to the Dunford–Pettis theorem; see Theorem A.67. Thus, the supremum in (4.32) is actually attained. An explicit construction of the maximizing measure will be given in the proof of Theorem 4.47. ◇

Example 4.38. We take for $\mathcal{Q}$ the class of all conditional distributions $P[\,\cdot \mid A\,]$ such that $A \in \mathcal{F}$ has $P[\,A\,] > \lambda$ for some fixed level $\lambda \in (0, 1)$. The coherent measure of risk induced by $\mathcal{Q}$,

$$WCE_\lambda(X) := \sup\{\, E[\,-X \mid A\,] \mid A \in \mathcal{F},\ P[\,A\,] > \lambda \,\}, \tag{4.33}$$

is called the *worst conditional expectation* at level λ. We will show in Section 4.4 that it coincides with the Average Value at Risk of Example 4.37 if the underlying probability space is rich enough. ◇

Remark 4.39. In analogy to Remark 4.17, the implication (e)⇒(a) in the Representation Theorem 4.31 can be viewed as a special case of the general duality in Theorem A.61 for the *Fenchel–Legendre transform* of the convex function ρ on L^∞, combined with the properties of a monetary risk measure. From this general point of view, it is now clear how to state representation theorems for convex risk measures on the Banach spaces $L^p(\Omega, \mathcal{F}, P)$ for $1 \le p < \infty$. More precisely, let $q \in (1, \infty]$ be such that $\frac{1}{p} + \frac{1}{q} = 1$, and define

$$\mathcal{M}_1^q(P) := \Big\{\, Q \in \mathcal{M}_1(P) \,\Big|\, \frac{dQ}{dP} \in L^q \,\Big\}.$$

A convex risk measure ρ on L^p is of the form

$$\rho(X) = \sup_{Q \in \mathcal{M}_1^q(P)} \big(E_Q[-X] - \alpha(Q) \big)$$

if and only if it is lower semicontinuous on L^p, i.e., the Fatou property holds in the form

$$X_n \longrightarrow X \text{ in } L^p \quad \Longrightarrow \quad \rho(X) \leq \liminf_{n \uparrow \infty} \rho(X_n). \qquad \diamond$$

4.4 Value at Risk

A common approach to the problem of measuring the risk of a financial position X consists in specifying a quantile of the distribution of X under the given probability measure P. For $\lambda \in (0, 1)$, a λ-*quantile* of a random variable X on $(\Omega, \mathcal{F}, P)$ is any real number q with the property

$$P[X \leq q] \geq \lambda \quad \text{and} \quad P[X < q] \leq \lambda,$$

and the set of all λ-quantiles of X is an interval $[q_X^-(\lambda), q_X^+(\lambda)]$, where

$$q_X^-(t) = \sup\{x \mid P[X < x] < t\} = \inf\{x \mid P[X \leq x] \geq t\}$$

is the lower and

$$q_X^+(t) = \inf\{x \mid P[X \leq x] > t\} = \sup\{x \mid P[X < x] \leq t\}$$

is the upper quantile function of X; see Appendix A.3. In this section, we will focus on the properties of $q_X^+(\lambda)$, viewed as a functional on a space of financial positions X.

Definition 4.40. Fix some level $\lambda \in (0, 1)$. For a financial position X, we define its *Value at Risk at level* λ as

$$V@R_\lambda(X) := -q_X^+(\lambda) = q_{-X}^-(1-\lambda) = \inf\{m \mid P[X + m < 0] \leq \lambda\}. \tag{4.34}$$

In financial terms, $V@R_\lambda(X)$ is the smallest amount of capital which, if added to X and invested in the risk-free asset, keeps the probability of a negative outcome below the level λ. However, Value at Risk only controls the probability of a loss; it does not capture the size of such a loss if it occurs. Clearly, $V@R_\lambda$ is a monetary measure of risk on $\mathcal{X} = L^0$, which is positively homogeneous. The following example shows that the acceptance set of $V@R_\lambda$ is typically not convex, and so $V@R_\lambda$ is *not* a convex measure of risk. Thus, $V@R_\lambda$ may penalize diversification instead of encouraging it.

Example 4.41. Consider an investment into two defaultable corporate bonds, each with return $\widetilde{r} > r$, where $r \geq 0$ is the return on a riskless investment. The discounted net gain of an investment $w > 0$ in the i^{th} bond is given by

$$X_i = \begin{cases} -w & \text{in case of default,} \\ \dfrac{w(\widetilde{r} - r)}{1 + r} & \text{otherwise.} \end{cases}$$

If a default of the first bond occurs with probability $p \leq \lambda$, then

$$P\Big[\, X_1 - \frac{w(\widetilde{r} - r)}{1 + r} < 0 \,\Big] = P[\, 1^{\text{st}} \text{ bond defaults} \,] = p \leq \lambda.$$

Hence,

$$V@R_\lambda(X_1) = -\frac{w(\widetilde{r} - r)}{1 + r} < 0.$$

This means that the position X_1 is acceptable in the sense that is does not carry a positive Value at Risk, regardless of the possible loss of the entire investment w.

Diversifying the portfolio by investing the amount $w/2$ into each of the two bonds leads to the position $Y := (X_1 + X_2)/2$. Let us assume that the two bonds default independently of each other, each of them with probability p. For realistic $\widetilde{r}$, the probability that Y is negative is equal to the probability that at least one of the two bonds defaults: $P[\, Y < 0 \,] = p(2 - p)$. If, for instance, $p = 0.009$ and $\lambda = 0.01$ then we have $p < \lambda < p(2 - p)$, hence

$$V@R_\lambda(Y) = \frac{w}{2} \cdot \Big(1 - \frac{\widetilde{r} - r}{1 + r} \Big).$$

Typically, this value is close to one half of the invested capital w. In particular, the acceptance set of $V@R_\lambda$ is not convex. This example also shows that $V@R$ may strongly discourage diversification: It penalizes quite drastically the increase of the probability that something goes wrong, without rewarding the significant reduction of the expected loss conditional on the event of default. Thus, optimizing a portfolio with respect to $V@R_\lambda$ may lead to a concentration of the portfolio in one single asset with a sufficiently small default probability, but with an exposure to large losses. ◇

In the remainder of this section, we will focus on monetary measures of risk which, in contrast to $V@R_\lambda$, are convex or even coherent on $\mathcal{X} := L^\infty$. In particular, we are looking for convex risk measures which come close to $V@R_\lambda$. A first guess might be that one should take the smallest convex measure of risk, continuous from above, which dominates $V@R_\lambda$. However, since $V@R_\lambda$ itself is not convex, the following proposition shows that such a smallest $V@R_\lambda$-dominating convex measure of risk does not exist.

Proposition 4.42. *For each $X \in \mathcal{X}$ and each $\lambda \in (0, 1)$,*

$$V@R_\lambda(X) = \min \big\{\, \rho(X) \mid \rho \textit{ is convex, continuous from above, and} \geq V@R_\lambda \,\big\}.$$

Proof. Let $q := -V@R_\lambda(X) = q_X^+(\lambda)$ so that $P[\, X < q\,] \le \lambda$. If $A \in \mathcal{F}$ satisfies $P[\, A\,] > \lambda$, then $P[\, A \cap \{X \ge q\}\,] > 0$. Thus, we may define a measure Q_A by

$$Q_A := P[\,\cdot \mid A \cap \{X \ge q\}\,].$$

It follows that $E_{Q_A}[\,-X\,] \le -q = V@R_\lambda(X)$.

Let $\mathcal{Q} := \{\, Q_A \mid P[\, A\,] > \lambda\,\}$, and use this set to define a coherent measure of risk ρ via

$$\rho(Y) := \sup_{Q \in \mathcal{Q}} E_Q[\,-Y\,].$$

Then $\rho(X) \le V@R_\lambda(X)$. Hence, the assertion will follow if we can show that $\rho(Y) \ge V@R_\lambda(Y)$ for each $Y \in \mathcal{X}$. Let $\varepsilon > 0$ and take $A := \{\, Y \le -V@R_\lambda(Y)+\varepsilon\,\}$. Clearly $P[\, A\,] > \lambda$, and so $Q_A \in \mathcal{Q}$. Moreover, $Q_A[\, A\,] = 1$, and we obtain

$$\rho(Y) \ge E_{Q_A}[\,-Y\,] \ge V@R_\lambda(Y) - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the result follows. □

For the rest of this section, we concentrate on the following risk measure which is defined in terms of Value at Risk, but does satisfy the axioms of a coherent risk measure.

Definition 4.43. The *Average Value at Risk at level* $\lambda \in (0, 1]$ of a position $X \in \mathcal{X}$ is given by

$$AV@R_\lambda(X) = \frac{1}{\lambda} \int_0^\lambda V@R_\gamma(X)\, d\gamma.$$

Sometimes, the Average Value at Risk is also called the "Conditional Value at Risk" or the "expected shortfall", and one writes $CV@R_\lambda(X)$ or $ES_\lambda(X)$. These terms are motivated by formulas (4.37) and (4.35) below, but they are potentially misleading: "Conditional Value at Risk" might also be used to denote the Value at Risk with respect to a conditional distribution, and "expected shortfall" might be understood as the expectation of the shortfall X^-. For these reasons, we prefer the term Average Value at Risk. Note that

$$AV@R_\lambda(X) = -\frac{1}{\lambda} \int_0^\lambda q_X(t)\, dt$$

by (4.34). In particular, the definition of $AV@R_\lambda(X)$ makes sense for any $X \in L^1(\Omega, \mathcal{F}, P)$ and we have, in view of Lemma A.19,

$$AV@R_1(X) = -\int_0^1 q_X^+(t)\, dt = E[\,-X\,].$$

Remark 4.44. Theorem 2.58 shows that the partial order $\succcurlyeq_{\text{uni}}$ on probability measures on $\mathbb{R}$ with finite mean can be characterized in terms of Average Value at Risk:

$$\mu \succcurlyeq_{\text{uni}} \nu \iff AV@R_\lambda(X_\mu) \le AV@R_\lambda(X_\nu) \quad \text{for all } \lambda \in (0, 1],$$

where X_μ and X_ν are random variables with distributions μ and ν. ◇

Remark 4.45. For $X \in L^\infty$, we have

$$\lim_{\lambda\downarrow 0} V@R_\lambda(X) = -\operatorname{ess\,inf} X = \inf\{\, m \mid P[\, X+m < 0\,] \le 0\,\}.$$

Hence, it makes sense to define

$$AV@R_0(X) := V@R_0(X) := -\operatorname{ess\,inf} X,$$

which is the worst-case risk measure on L^∞ introduced in Example 4.36. Recall that it is continuous from above but in general not from below. ◇

Lemma 4.46. *For $\lambda \in (0,1)$ and any λ-quantile q of X,*

$$AV@R_\lambda(X) = \frac{1}{\lambda} E[\,(q-X)^+\,] - q = \frac{1}{\lambda} \inf_{r\in\mathbb{R}} \big(E[\,(r-X)^+\,] - \lambda r \big). \tag{4.35}$$

Proof. Let q_X be a quantile function with $q_X(\lambda) = q$. By Lemma A.19,

$$\frac{1}{\lambda} E[\,(q-X)^+\,] - q = \frac{1}{\lambda} \int_0^1 (q - q_X(t))^+ \, dt - q = -\frac{1}{\lambda} \int_0^\lambda q_X(t)\, dt = AV@R_\lambda(X).$$

This proves the first identity. The second one follows from Lemma A.22. □

Theorem 4.47. *For $\lambda \in (0,1]$, $AV@R_\lambda$ is a coherent measure of risk which is continuous from below. It has the representation*

$$AV@R_\lambda(X) = \max_{Q\in\mathcal{Q}_\lambda} E_Q[\,-X\,], \quad X \in \mathcal{X}, \tag{4.36}$$

where $\mathcal{Q}_\lambda$ is the set of all probability measures $Q \ll P$ whose density dQ/dP is P-a.s. bounded by $1/\lambda$. Moreover, $\mathcal{Q}_\lambda$ is equal to the maximal set $\mathcal{Q}_{\max}$ of Corollary 4.34.

Proof. Since $\mathcal{Q}_1 = \{P\}$, the assertion is obvious for $\lambda = 1$. For $0 < \lambda < 1$, consider the coherent risk measure $\rho_\lambda(X) := \sup_{Q\in\mathcal{Q}_\lambda} E_Q[\,-X\,]$. First we assume that we are given some $X < 0$. We define a measure $\widetilde{P} \approx P$ by $d\widetilde{P}/dP = X/E[\,X\,]$. Then

$$\rho_\lambda(X) = \frac{E[\,-X\,]}{\lambda} \sup\big\{\, \widetilde{E}[\,\varphi\,] \,|\, 0 \le \varphi \le 1,\ E[\,\varphi\,] = \lambda \,\big\}.$$

Clearly, the condition $E[\,\varphi\,] = \lambda$ on the right can be replaced by $E[\,\varphi\,] \le \lambda$. Thus, we can apply the Neyman–Pearson lemma in the form of Theorem A.30 and conclude that the supremum is attained by

$$\varphi_0 = I_{\{X<q\}} + \kappa I_{\{X=q\}}$$

for a λ-quantile q of X and some $\kappa \in [0,1]$ for which $E[\,\varphi_0\,] = \lambda$. Hence,

$$\rho_\lambda(X) = \frac{E[\,-X\,]}{\lambda} \cdot \widetilde{E}[\,\varphi_0\,] = \frac{1}{\lambda} E[\,-X\varphi_0\,],$$

Since $dQ_0 = \lambda^{-1}\varphi_0\, dP$ defines a probability measure in $\mathcal{Q}_\lambda$, we conclude that

$$\begin{aligned}\rho_\lambda(X) &= \max_{Q\in\mathcal{Q}_\lambda} E_Q[-X] = E_{Q_0}[-X]\\ &= \frac{1}{\lambda}\big(E[-X;\ X<q] - q\lambda + qP[X<q]\big)\\ &= \frac{1}{\lambda}E[(q-X)^+] - q\\ &= AV@R_\lambda(X),\end{aligned}$$

where we have used (4.35) in the last step. This proves (4.36) for $X < 0$. For arbitrary $X \in L^\infty$, we use the cash invariance of both ρ_λ and $AV@R_\lambda$.

It remains to prove that $\mathcal{Q}_\lambda$ is the maximal set of Corollary 4.34. To this end, we show that

$$\sup_{X\in\mathcal{X}} \big(E_Q[-X] - AV@R_\lambda(X)\big) = +\infty$$

for $Q \notin \mathcal{Q}_\lambda$. We denote by φ the density dQ/dP. There exist $\lambda' \in (0,\lambda)$ and $k > 1/\lambda'$ such that $P[\varphi \wedge k \ge 1/\lambda'] > 0$. For $c > 0$ define $X^{(c)} \in \mathcal{X}$ by

$$X^{(c)} := -c(\varphi \wedge k)\mathrm{I}_{\{\varphi\ge 1/\lambda'\}}.$$

Since

$$P[X^{(c)} < 0] = P\Big[\varphi \ge \frac{1}{\lambda'}\Big] \le \lambda' < \lambda,$$

we have $V@R_\lambda(X^{(c)}) = 0$, and (4.35) yields that

$$AV@R_\lambda(X^{(c)}) = \frac{1}{\lambda}E[-X^{(c)}] = \frac{c}{\lambda}E\Big[\varphi\wedge k;\ \varphi \ge \frac{1}{\lambda'}\Big].$$

On the other hand,

$$E_Q[-X^{(c)}] = c\cdot E\Big[\varphi\cdot\varphi\wedge k;\ \varphi \ge \frac{1}{\lambda'}\Big] \ge \frac{c}{\lambda'}E\Big[\varphi\wedge k;\ \varphi \ge \frac{1}{\lambda'}\Big].$$

Thus, the difference between $E_Q[-X^{(c)}]$ and $AV@R_\lambda(X^{(c)})$ becomes arbitrarily large as $c \uparrow \infty$. □

Remark 4.48. The proof shows that for $\lambda \in (0,1)$ the maximum in (4.36) is attained by the measure $Q_0 \in \mathcal{Q}_\lambda$, whose density is given by

$$\frac{dQ_0}{dP} = \frac{1}{\lambda}\big(\mathrm{I}_{\{X<q\}} + \kappa\mathrm{I}_{\{X=q\}}\big),$$

where q is a λ-quantile of X, and where κ is defined as

$$\kappa := \begin{cases} 0 & \text{if } P[X=q]=0,\\ \dfrac{\lambda - P[X<q]}{P[X=q]} & \text{otherwise.}\end{cases}$$

◇

Corollary 4.49. *For all $X \in \mathcal{X}$,*

$$
\begin{aligned}
AV@R_\lambda(X) &\geq WCE_\lambda(X) \\
&\geq E[\,-X \mid -X \geq V@R_\lambda(X)\,] \qquad (4.37) \\
&\geq V@R_\lambda(X),
\end{aligned}
$$

where WCE_λ is the coherent risk measure defined in (4.33). *Moreover, the first two inequalities are in fact identities if*

$$
P\big[\,X \leq q_\lambda^+(X)\,\big] = \lambda, \qquad (4.38)
$$

which is the case if X has a continuous distribution.

Proof. If $P[\,A\,] \geq \lambda$, then the density $P[\,\cdot \mid A\,]$ with respect to P is bounded by $1/\lambda$. Therefore, Theorem 4.47 implies that $AV@R_\lambda$ dominates WCE_λ. Since

$$
P\big[\,-X \geq V@R_\lambda(X) - \varepsilon\,\big] > \lambda,
$$

we have

$$
WCE_\lambda(X) \geq E\big[\,-X \mid -X \geq V@R_\lambda(X) - \varepsilon\,\big],
$$

and the second inequality follows by taking the limit $\varepsilon \downarrow 0$. Moreover, (4.35) shows that

$$
AV@R_\lambda(X) = E\big[\,-X \mid -X \geq V@R_\lambda(X)\,\big]
$$

as soon as (4.38) holds. □

Remark 4.50. We will see in Corollary 4.62 that the two coherent risk measures $AV@R_\lambda$ and WCE_λ coincide if the underlying probability space is rich enough. If this is not the case, then the first inequality in (4.37) may be strict for some X; see [1]. Moreover, the functional

$$
E[\,-X \mid -X \geq V@R_\lambda(X)\,]
$$

does not define a convex measure of risk. Hence, the second inequality in (4.37) cannot reduce to an identity in general. ◇

Remark 4.51. We have seen in Proposition 4.42 that there is no smallest convex risk measure dominating $V@R_\lambda$. But if we restrict our attention to the class of convex risk measures that dominate $V@R_\lambda$ and only depend on the distribution of a random variable, then the situation is different. In fact, we will see in Theorem 4.61 that $AV@R_\lambda$ is the smallest risk measure in this class, provided that the underlying probability space is rich enough. In this sense, Average Value at Risk can be regarded as the best conservative approximation to Value at Risk. ◇

4.5 Law-invariant risk measures

Clearly, $V@R_\lambda$ and $AV@R_\lambda$ only involve the distribution of a position under the given probability measure P. In this section we study the class of all risk measures which share this property of law-invariance.

Definition 4.52. A monetary measure of risk ρ on $\mathcal{X} = L^\infty(\Omega, \mathcal{F}, P)$ is called *law-invariant* if $\rho(X) = \rho(Y)$ whenever X and Y have the same distribution under P.

Throughout this section, we assume that the probability space $(\Omega, \mathcal{F}, P)$ is rich enough in the sense that it supports a random variable with a continuous distribution. This condition is satisfied if and only if $(\Omega, \mathcal{F}, P)$ is atomless; see Proposition A.27.

Remark 4.53. Any law-invariant monetary risk measure ρ is monotone with respect to the partial order $\succcurlyeq_{\text{mon}}$ introduced in Definition 2.69. More precisely,

$$\mu \succcurlyeq_{\text{mon}} \nu \quad \Longrightarrow \quad \rho(X_\mu) \le \rho(X_\nu),$$

if X_μ and X_ν are random variables with distributions μ and ν. To prove this, let q_μ and q_ν be quantile functions for μ and ν and take a random variable U with a uniform distribution on $(0, 1)$. Then $\widetilde{X}_\mu := q_\mu(U) \ge q_\nu(U) =: \widetilde{X}_\nu$ by Theorem 2.70, and $\widetilde{X}_\mu$ and $\widetilde{X}_\nu$ have the same distribution as X_μ and X_ν by Lemma A.19. Hence, law-invariance and monotonicity of ρ imply $\rho(X_\mu) = \rho(\widetilde{X}_\mu) \le \rho(\widetilde{X}_\nu) = \rho(X_\nu)$. $\diamond$

We can now formulate our first structure theorem for law-invariant convex risk measures.

Theorem 4.54. *Let ρ be a convex measure of risk and suppose that ρ is continuous from above. Then ρ is law-invariant if and only if its minimal penalty function $\alpha_{\min}(Q)$ depends only on the law of $\varphi_Q := \frac{dQ}{dP}$ under P when $Q \in \mathcal{M}_1(P)$. In this case, ρ has the representation*

$$\rho(X) = \sup_{Q \in \mathcal{M}_1(P)} \Big(\int_0^1 q_{-X}(t) q_{\varphi_Q}(t)\, dt - \alpha_{\min}(Q) \Big),$$

and the minimal penalty function satisfies

$$\begin{aligned} \alpha_{\min}(Q) &= \sup_{X \in \mathcal{A}_\rho} \int_0^1 q_{-X}(t) q_{\varphi_Q}(t)\, dt \\ &= \sup_{X \in L^\infty} \Big(\int_0^1 q_{-X}(t) q_{\varphi_Q}(t) - \rho(X) \Big). \end{aligned} \tag{4.39}$$

For the proof, we will need the following lemma.

Lemma 4.55. *For $X \in L^\infty$ and $Y \in L^1$,*

$$\int_0^1 q_X(t) q_Y(t)\, dt = \sup_{\widetilde{X} \sim X} E[\, \widetilde{X} Y \,],$$

where $\widetilde{X} \sim X$ indicates that $\widetilde{X}$ is a random variable with the same law as X.

Proof. The upper Hardy–Littlewood inequality in Theorem A.24 yields "$\geq$". To prove the reverse inequality, let us first assume that Y has a continuous distribution. Then Lemma A.21 implies that $U := F_Y(Y)$ has a uniform distribution and that P-a.s. $Y = q_Y(U)$. Since $\widetilde{X} := q_X(U) \sim X$ by Lemma A.19, we obtain

$$E[\,\widetilde{X}Y\,] = E[\,q_X(U)q_Y(U)\,] = \int_0^1 q_X(t)q_Y(t)\,dt,$$

and hence "$\leq$".

In the general case, let D be the set of all y such that $P[\,Y = y\,] > 0$ and take a random variable $Z \in L^1_+$ with a continuous distribution. Such a random variable exists due to Proposition A.27. We claim that the law of

$$Y_n := Y + \frac{1}{n} Z \mathrm{I}_{\{Y \in D\}}$$

is continuous. Indeed, for any y,

$$P[\,Y_n = y\,] = P[\,Y = y,\; Y \notin D\,] + \sum_{x \in D} P[\,Y = x,\; Z = n(y - x)\,] = 0.$$

Thus, $U_n := F_{Y_n}(Y_n)$ has a uniform distribution on (0, 1), and $X_n := q_X(U_n)$ has the same distribution as X. By adding a suitable constant to X, we may assume without loss of generality that $X \geq 0$. Since $Y_n \geq Y$ we have $q_{Y_n} \geq q_Y$ a.e., and it follows from the first part of the proof that

$$\begin{aligned}
\int_0^1 q_X(t)q_Y(t)\,dt &\leq \liminf_{n\uparrow\infty} \int_0^1 q_X(t)q_{Y_n}(t)\,dt \\
&= \liminf_{n\uparrow\infty} \sup_{\widetilde{X}\sim X} E[\,\widetilde{X}Y_n\,] \\
&= \sup_{\widetilde{X}\sim X} E[\,\widetilde{X}Y\,],
\end{aligned}$$

where the last identity follows from the fact that

$$\left| E[\,\widetilde{X}Y_n\,] - E[\,\widetilde{X}Y\,] \right| \leq \frac{1}{n} \|Z\|_1 \|X\|_\infty,$$

for all $\widetilde{X} \sim X$. □

Proof of Theorem 4.54. Suppose first that ρ is law-invariant. Then $X \in \mathcal{A}_\rho$ implies that $\widetilde{X} \in \mathcal{A}_\rho$ for all $\widetilde{X} \sim X$. Hence,

$$\alpha_{\min}(Q) = \sup_{X \in \mathcal{A}_\rho} E[\,-X\varphi_Q\,] = \sup_{X \in \mathcal{A}_\rho} \sup_{\widetilde{X}\sim X} E[\,-\widetilde{X}\varphi_Q\,] = \sup_{X \in \mathcal{A}_\rho} \int_0^1 q_{-X}(t)q_{\varphi_Q}(t)\,dt,$$

where we have used Lemma 4.55 in the last step. It follows that $\alpha_{\min}(Q)$ depends only on the law of φ_Q. In order to check the second identity in (4.39), note that $\widetilde{X} := X + \rho(X)$ belongs to $\mathcal{A}_\rho$ for any $X \in L^\infty$ and that $q_{-X} - \rho(X)$ is a quantile function for $-\widetilde{X}$.

Conversely, let us assume that $\alpha_{\min}(Q)$ depends only on the law of φ_Q. Let us write $\widetilde{Q} \sim Q$ to indicate that φ_Q and $\varphi_{\widetilde{Q}}$ have the same law. Then Lemma 4.55 yields

$$\begin{aligned}\rho(X) &= \sup_{Q\in\mathcal{M}_1(P)} \Big(E_Q[\,-X\,] - \alpha_{\min}(Q)\Big)\\ &= \sup_{Q\in\mathcal{M}_1(P)} \sup_{\widetilde{Q}\sim Q} \Big(E[\,-X\varphi_{\widetilde{Q}}\,] - \alpha_{\min}(Q)\Big)\\ &= \sup_{Q\in\mathcal{M}_1(P)} \Big(\int_0^1 q_{-X}(t) q_{\varphi_Q}(t)\,dt - \alpha_{\min}(Q)\Big). \qquad \square\end{aligned}$$

Example 4.56. Let $u : \mathbb{R} \to \mathbb{R}$ be an increasing concave function, and suppose that a position $X \in L^\infty$ is acceptable if $E[\,u(X)\,] \geq c$, where c is a given constant in the interior of $u(\mathbb{R})$. We have seen in Example 4.10 that the corresponding acceptance set induces a convex risk measure ρ. Clearly, ρ is law-invariant, and it will be shown in Proposition 4.104 that ρ is continuous from below and, hence, from above. Moreover, the corresponding minimal penalty function can be computed as

$$\alpha_{\min}(Q) = \inf_{\lambda>0} \frac{1}{\lambda}\Big(x_0 + \int_0^1 \ell^*\big(\lambda \cdot q_{\varphi_Q}(t)\big)\,dt\Big),$$

where

$$\ell^*(y) = \sup_{x\in\mathbb{R}} \big(xy + u(-x)\big) = \sup_{x\in\mathbb{R}} \big(xy - \ell(x)\big)$$

is the Fenchel–Legendre transform of the convex increasing loss function $\ell(x) := -u(-x)$; see Theorem 4.106. ◇

The following theorem clarifies the crucial role of the risk measures $AV@R_\lambda$: they can be viewed as the building blocks for law-invariant convex measures of risk on L^∞. Recall that we assume that $(\Omega, \mathcal{F}, P)$ is atomless.

Theorem 4.57. *A convex measure of risk ρ is law-invariant and continuous from above if and only if*

$$\rho(X) = \sup_{\mu\in\mathcal{M}_1((0,1])} \Big(\int_{(0,1]} AV@R_\lambda(X)\,\mu(d\lambda) - \beta_{\min}(\mu)\Big), \tag{4.40}$$

where

$$\beta_{\min}(\mu) = \sup_{X\in\mathcal{A}_\rho} \int_{(0,1]} AV@R_\lambda(X)\,\mu(d\lambda).$$

Proof. Clearly, the right-hand side of (4.40) defines a law-invariant convex risk measure that is continuous from above. Conversely, let ρ be law-invariant and continuous from above. We will show that for $Q \in \mathcal{M}_1(P)$ there exists a measure $\mu \in \mathcal{M}_1((0,1])$ such that

$$\int_0^1 q_{-X}(t)q_\varphi(t)\,dt = \int_{(0,1]} AV@R_s(X)\,\mu(ds),$$

where $\varphi := \varphi_Q = \frac{dQ}{dP}$. Then the assertion will follow from Theorem 4.54. Since $q_{-X}(t) = V@R_{1-t}(X)$ and $q_\varphi(t) = q_\varphi^+(t)$ for a.e. $t \in (0,1)$,

$$\int_0^1 q_{-X}(t)q_\varphi(t)\,dt = \int_0^1 V@R_t(X)q_\varphi^+(1-t)\,dt.$$

Since q_φ^+ is increasing and right-continuous, we can write $q_\varphi^+(t) = \nu((1-t,1])$ for some positive locally finite measure ν on $(0,1]$. Moreover, the measure μ given by $\mu(dt) = t\,\nu(dt)$ is a probability measure on $(0,1]$:

$$\int_{(0,1]} t\,\nu(dt) = \int_0^1 \nu((s,1])\,ds = \int_0^1 q_\varphi^+(s)\,ds = E[\,\varphi\,] = 1.$$

Thus,

$$\begin{aligned}\int_0^1 q_{-X}(t)q_\varphi(t)\,dt &= \int_0^1 V@R_t(X)\int_{(t,1]}\frac{1}{s}\,\mu(ds)\,dt\\ &= \int_{(0,1]}\frac{1}{s}\int_0^s V@R_t(X)\,dt\,\mu(ds) \qquad (4.41)\\ &= \int_{(0,1]} AV@R_s(X)\,\mu(ds).\end{aligned}$$

Conversely, for any probability measure μ on $(0,1]$, the function q defined by $q(t) := \int_{(1-t,1]} s^{-1}\,\mu(ds)$ can be viewed as the quantile function of the density $\varphi := q(U)$ of a measure $Q \in \mathcal{M}_1(P)$, where U has a uniform distribution on $(0,1)$. Altogether, we obtain a one-to-one correspondence between laws of densities φ and probability measures μ on $(0,1]$. □

Theorem 4.57 takes the following form for coherent measures of risk.

Corollary 4.58. *A coherent risk measure ρ is continuous from above and law-invariant if and only if*

$$\rho(X) = \sup_{\mu\in\mathcal{M}}\int_{(0,1]} AV@R_\lambda(X)\,\mu(d\lambda)$$

for some set $\mathcal{M} \subset \mathcal{M}_1((0,1])$.

Randomness of a position is reduced in terms of P if we replace the position by its conditional expectation with respect to some σ-algebra $\mathcal{G} \subset \mathcal{F}$. Such a reduction of randomness is reflected by a convex risk measure if it is law-invariant:

Corollary 4.59. *Assume that ρ is a convex risk measure which is continuous from above and law-invariant. Then ρ is monotone with respect to the binary relation $\succcurlyeq_{\text{uni}}$ introduced in* (3.27)*:*

$$Y \succcurlyeq_{\text{uni}} X \quad \Longrightarrow \quad \rho(Y) \le \rho(X),$$

for $Y, X \in \mathcal{X}$. In particular,

$$\rho\big(E[\,X \mid \mathcal{G}\,]\big) \le \rho(X),$$

for $X \in \mathcal{X}$ and any σ-algebra $\mathcal{G} \subset \mathcal{F}$, and

$$\rho\big(E[\,X\,]\big) = \rho(0) - E[\,X\,] \le \rho(X).$$

Proof. The first inequality follows from Theorem 4.57 combined with Remark 4.44. The second inequality is a special case of the first one, since $E[\,X \mid \mathcal{G}\,] \succcurlyeq_{\text{uni}} X$ according to Theorem 2.58. The third follows from the second by taking $\mathcal{G} = \{\emptyset, \Omega\}$. □

Recall from Theorem 2.70 that $\mu \succcurlyeq_{\text{mon}} \nu$ implies $\mu \succcurlyeq_{\text{uni}} \nu$. Thus, the preceding conclusion for convex risk measures is stronger than the one of Remark 4.53 for monetary risk measures.

Remark 4.60. If $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \cdots \subset \mathcal{F}$ are σ-algebras, then

$$\rho\big(E[\,X \mid \mathcal{G}_n\,]\big) \longrightarrow \rho\big(E[\,X \mid \mathcal{G}_\infty\,]\big) \quad \text{as } n \uparrow \infty,$$

where ρ is as in Corollary 4.59 and $\mathcal{G}_\infty = \sigma\big(\bigcup_n \mathcal{G}_n\big)$. Indeed, Doob's martingale convergence theorem (see, e.g., Theorem 19.1 in [19]) states that $E[\,X \mid \mathcal{G}_n\,] \to E[\,X \mid \mathcal{G}_\infty\,]$ P-a.s. as $n \uparrow \infty$. Hence, the Fatou property and Corollary 4.59 show that

$$\begin{aligned}\rho\big(E[\,X \mid \mathcal{G}_\infty\,]\big) &= \rho\big(\lim_{n\uparrow\infty} E[\,X \mid \mathcal{G}_n\,]\big)\\ &\le \liminf_{n\uparrow\infty} \rho\big(E[\,X \mid \mathcal{G}_n\,]\big)\\ &\le \rho\big(E[\,X \mid \mathcal{G}_\infty\,]\big).\end{aligned}$$

◇

In contrast to Proposition 4.42, the following theorem shows that $AV@R_\lambda$ is the best conservative approximation to $V@R_\lambda$ in the class of all *law-invariant* convex measures of risk which are continuous from above.

Theorem 4.61. *$AV@R_\lambda$ is the smallest law-invariant convex measure of risk which is continuous from above and dominates $V@R_\lambda$.*

Proof. That $AV@R_\lambda$ dominates $V@R_\lambda$ was already stated in (4.37). Suppose now that ρ is another law-invariant convex risk measure which dominates $V@R_\lambda$ and which is continuous from above. We must show that for a given $X \in \mathcal{X}$

$$\rho(X) \ge AV@R_\lambda(X). \tag{4.42}$$

Take $\varepsilon > 0$, and let $A := \{ -X \geq V@R_\lambda(X) - \varepsilon \}$ and

$$Y := E[\, X \mid X I_{A^c} \,] = X \cdot I_{A^c} + E[\, X \mid A \,] \cdot I_A .$$

Since $Y > q_X^+(\lambda) + \varepsilon \geq E[\, X \mid A \,]$ on A^c, we get $P[\, Y < E[\, X \mid A \,]\,] = 0$. On the other hand, $P[\, Y \leq E[\, X \mid A \,]\,] \geq P[\, A \,] > \lambda$, and this implies that $V@R_\lambda(Y) = E[\, -X \mid A \,]$. Since ρ dominates $V@R_\lambda$, we have $\rho(Y) \geq E[\, -X \mid A \,]$. Thus,

$$\rho(X) \geq \rho(Y) = E[\, -X \mid -X \geq V@R_\lambda(X) - \varepsilon \,],$$

by Corollary 4.59. Taking $\varepsilon \downarrow 0$ yields

$$\rho(X) \geq E[\, -X \mid -X \geq V@R_\lambda(X) \,].$$

If the distribution of X is continuous, Corollary 4.49 states that the conditional expectation on the right equals $AV@R_\lambda(X)$, and we obtain (4.42). If the distribution of X is not continuous, we denote by D the set of all points x such that $P[\, X = x \,] > 0$ and take any bounded random variable $Z \geq 0$ with a continuous distribution. Such a random variable exists due to Proposition A.27. We have seen in the proof of Lemma 4.55 that $X_n := X + \frac{1}{n} Z I_{\{X \in D\}}$ has a continuous distribution. Moreover, X_n decreases to X. The inequality (4.42) holds for each X_n and extends to X by continuity from above. $\square$

Corollary 4.62. *$AV@R_\lambda$ and WCE_λ coincide under our assumption that the probability space is atomless.*

Proof. We know from Corollary 4.49 that $WCE_\lambda(X) = AV@R_\lambda(X)$ if X has a continuous distribution. Repeating the approximation argument at the end of the preceding proof yields $WCE_\lambda(X) = AV@R_\lambda(X)$ for each $X \in \mathcal{X}$. $\square$

4.6 Concave distortions

Let us now have a closer look at the coherent risk measures

$$\rho_\mu(X) := \int AV@R_\lambda(X)\, \mu(d\lambda), \tag{4.43}$$

which appear in the Representation Theorem 4.57 for law-invariant convex risk measures. We are going to characterize these risk measures ρ_μ in two ways, first as Choquet integrals with respect to some concave distortion of the underlying probability measure P, and then, in the next section, by a property of comonotonicity.

Again, we will assume throughout this section that the underlying probability space $(\Omega, \mathcal{F}, P)$ is atomless. Since $AV@R_\lambda$ is coherent, continuous from below, and law-invariant, any mixture ρ_μ for some probability measure μ on (0, 1] has the same properties. According to Remark 4.45, we may set $AV@R_0(X) = -\operatorname{ess\,inf} X$ so that we can extend the definition (4.43) to probability measures μ on the closed interval [0, 1]. However, ρ_μ will only be continuous from above and not from below if $\mu(\{0\}) > 0$, because $AV@R_0$ is not continuous from below.

Our first goal is to show that $\rho_\mu(X)$ can be identified with the Choquet integral of the loss $-X$ with respect to the set function $c_\psi(A) := \psi(P[\,A\,])$, where ψ is the concave function defined in the following lemma. Choquet integrals were introduced in Remark **??**. Recall that every concave function ψ admits a right-continuous right-hand derivative ψ'_+; see Proposition A.4.

Lemma 4.63. *The identity*

$$\psi'_+(t) = \int_{(t,1]} s^{-1}\,\mu(ds)\,, \quad 0 < t < 1, \tag{4.44}$$

defines a one-to-one correspondence between probability measures μ on [0, 1] *and increasing concave functions $\psi : [0,1] \to [0,1]$ with $\psi(0) = 0$ and $\psi(1) = 1$. Moreover, we have $\psi(0+) = \mu(\{0\})$.*

Proof. Suppose first that μ is given and ψ is defined by $\psi(1) = 1$ and (4.44). Then ψ is concave and increasing on (0, 1]. Moreover,

$$1 - \psi(0+) = \int_0^1 \psi'(t)\,dt = \int_{(0,1]} \frac{1}{s} \int_0^1 \mathrm{I}_{\{t<s\le 1\}}\,dt\,\mu(ds) = \mu((0,1]) \le 1.$$

Hence, we may set $\psi(0) := 0$ and obtain an increasing concave function on [0, 1].

Conversely, if ψ is given, then $\psi'_+(t)$ is a decreasing right-continuous function on (0, 1) and can be written as $\psi'_+(t) = \nu((t,1])$ for some locally finite positive measure ν on (0, 1]. We first define μ on (0, 1] by $\mu(dt) = t\,\nu(dt)$. Then (4.44) holds and, by Fubini's theorem,

$$\mu((0,1]) = \int_0^1 \int_{(0,1]} \mathrm{I}_{\{t<s\}}\,\nu(ds)\,dt = 1 - \psi(0+) \le 1.$$

Hence, setting $\mu(\{0\}) := \psi(0+)$ defines a probability measure μ on [0, 1]. □

Theorem 4.64. *For a probability measure μ on* [0, 1]*, let ψ be the concave function defined in Lemma* 4.63*. Then, for $X \in \mathcal{X}$,*

$$\begin{aligned}\rho_\mu(-X) &= \psi(0+)AV@R_0(-X) + \int_0^1 q_X(t)\psi'(1-t)\,dt \\ &= \int_{-\infty}^0 \big(\psi(P[\,X > x\,]) - 1\big)\,dx + \int_0^\infty \psi(P[\,X > x\,])\,dx.\end{aligned}$$

Proof. Using the fact that $V@R_\lambda(-X) = q_X^-(1-\lambda)$, we get as in (4.41) that

$$\int_{(0,1]} AV@R_\lambda(-X)\,\mu(d\lambda) = \int_0^1 q_X(t)\psi'(1-t)\,dt.$$

Hence, we obtain the first identity. For the second one, we will first assume $X \geq 0$. Then

$$q_X^+(t) = \sup\{\, x \geq 0 \mid F_X(x) \leq t \,\} = \int_0^\infty \mathrm{I}_{\{F_X(x)\leq t\}}\,dx,$$

where F_X is the distribution function of X. Using Fubini's theorem, we obtain

$$\begin{aligned}\int_0^1 q_X(t)\psi'(1-t)\,dt &= \int_0^\infty \int_0^1 \mathrm{I}_{\{F_X(x)\leq 1-t\}}\psi'(t)\,dt\,dx \\ &= \int_0^\infty \psi(1-F_X(x))\,dx - \psi(0+)\operatorname{ess\,sup} X,\end{aligned}$$

since $\int_0^y \psi'(t)\,dt = (\psi(y) - \psi(0+))\mathrm{I}_{\{y>0\}}$. This proves the second identity for $X \geq 0$, since $\psi(0+) = \mu(\{0\})$ and $\operatorname{ess\,sup} X = AV@R_0(-X)$. If $X \in L^\infty$ is arbitrary, we consider $X + C$, where $C := -\operatorname{ess\,inf} X$. The cash invariance of ρ_μ yields

$$\begin{aligned}C + \rho_\mu(-X) &= \int_0^\infty \psi(P[\, X > x - C\,])\,dx \\ &= \int_{-C}^0 \psi(P[\, X > x\,])\,dx + \int_0^\infty \psi(P[\, X > x\,])\,dx \\ &= C + \int_{-\infty}^0 \big(\psi(P[\, X > x\,]) - 1\big)\,dx + \int_0^\infty \psi(P[\, X > x\,])\,dx. \quad \square\end{aligned}$$

Example 4.65. Clearly, the risk measure $AV@R_\lambda$ is itself of the form ρ_μ where $\mu = \delta_\lambda$. For $\lambda > 0$, the corresponding concave distortion function is given by

$$\psi(t) = \Big(\frac{t}{\lambda}\Big) \wedge 1 = \frac{1}{\lambda}\,(t \wedge \lambda).$$

Thus, we obtain yet another representation of $AV@R_\lambda$:

$$AV@R_\lambda(-X) = \frac{1}{\lambda}\int_0^\infty P[\, X > x\,] \wedge \lambda\,dx \quad \text{for } X \in L_+^\infty. \qquad \diamond$$

Corollary 4.66. *If* $\mu(\{0\}) = 0$ *in Theorem* 4.64*, then*

$$\rho_\mu(X) = -\int_0^1 q_X(\varphi(t))\,dt,$$

where φ *is an inverse function of* ψ*, taken in the sense of Definition* A.14.

Proof. Due to Lemma A.15, the distribution of φ under the Lebesgue measure has the distribution function ψ and hence the density ψ'. Therefore

$$\int_0^1 q_X(\varphi(t))\,dt = \int_0^1 q_X(t)\psi'(t)\,dt = -\int_0^1 q_{-X}(1-t)\psi'(t)\,dt,$$

where we have used Lemma A.23 in the last step. An application of Theorem 4.64 concludes the proof. □

Let us continue with a brief discussion of the set function $c_\psi(A) = \psi(P[\,A\,])$.

Definition 4.67. Let $\psi : [0, 1] \to [0, 1]$ be an increasing function such that $\psi(0) = 0$ and $\psi(1) = 1$. The set function

$$c_\psi(A) := \psi(P[\,A\,]), \quad A \in \mathscr{F},$$

is called the *distortion* of the probability measure P with respect to the *distortion function* ψ.

Definition 4.68. A set function $c : \mathscr{F} \to [0, 1]$ is called *monotone* if

$$c(A) \le c(B) \quad \text{for } A \subset B$$

and *normalized* if

$$c(\emptyset) = 0 \quad \text{and} \quad c(\Omega) = 1.$$

A monotone set function is called *submodular* or 2-*alternating* if

$$c(A \cup B) + c(A \cap B) \le c(A) + c(B).$$

Clearly, any distortion c_ψ is normalized and monotone.

Proposition 4.69. *Let c_ψ be the distortion of P with respect to the distortion function ψ. If ψ is concave, then c_ψ is submodular. Moreover, if the underlying probability space is atomless, then also the converse implication holds.*

Proof. Suppose first that ψ is concave. Take $A, B \in \mathscr{F}$ with $P[\,A\,] \le P[\,B\,]$. We must show that $c := c_\psi$ satisfies

$$c(A) - c(A \cap B) \ge c(A \cup B) - c(B).$$

This is trivial if $r = 0$, where

$$r := P[\,A\,] - P[\,A \cap B\,] = P[\,A \cup B\,] - P[\,B\,].$$

For $r > 0$ the concavity of ψ yields via (A.1) that

$$\frac{c(A) - c(A \cap B)}{P[\,A\,] - P[\,A \cap B\,]} \ge \frac{c(A \cup B) - c(B)}{P[\,A \cup B\,] - P[\,B\,]}.$$

Multiplying both sides with r gives the result.

Now suppose that $c = c_\psi$ is submodular and assume that $(\Omega, \mathcal{F}, P)$ is atomless. We have to show that $\psi(y) \geq (\psi(x) + \psi(z))/2$ whenever $0 \leq x \leq z \leq 1$ and $y = (x+z)/2$. To this end, we will construct two sets $A, B \subset \mathcal{F}$ such that $P[\,A\,] = P[\,B\,] = y$, $P[\,A \cap B\,] = x$, and $P[\,A \cup B\,] = z$. Submodularity then gives $\psi(x) + \psi(z) \leq 2\psi(y)$ and in turn the concavity of ψ.

In order to construct the two sets A and B, take a random variable U with a uniform distribution on $[0, 1]$, which exists by Proposition A.27. Then

$$A := \{\,0 \leq U \leq y\,\} \quad \text{and} \quad B := \{\,z - y \leq U \leq z\,\}$$

are as desired. □

Let us now recall the notion of a Choquet integral, which was introduced in Example 4.13.

Definition 4.70. Let $c : \mathcal{F} \to [0, 1]$ be any set function which is normalized and monotone. The *Choquet integral* of a bounded measurable function X on $(\Omega, \mathcal{F})$ with respect to c is defined as

$$\int X\,dc := \int_{-\infty}^{0} (c(X > x) - 1)\,dx + \int_{0}^{\infty} c(X > x)\,dx.$$

Note that the Choquet integral coincides with the usual integral as soon as c is a σ-additive probability measure; see also Lemma 4.91 below.

With this definition, Theorem 4.64 allows us to identify the risk measure ρ_μ as the Choquet integral of the loss with respect to a concave distortion c_ψ of the underlying probability measure P:

Corollary 4.71. *For a probability measure μ on* $[0, 1]$*, let ψ be the concave distortion function defined in Lemma* 4.63*, and let c_ψ denote the distortion of P with respect to ψ. Then, for $X \in L^\infty$,*

$$\rho_\mu(X) = \int (-X)\,dc_\psi.$$

Combining Corollary 4.71 with Theorem 4.57, we obtain the following characterization of law-invariant convex risk measures in terms of concave distortions:

Corollary 4.72. *A convex risk measure ρ is law-invariant and continuous from above if and only if*

$$\rho(X) = \sup_{\psi} \Big(\int (-X)\,dc_\psi - \gamma_{\min}(\psi) \Big),$$

where the supremum is taken over the class of all concave distortion functions ψ and

$$\gamma_{\min}(\psi) := \sup_{X \in \mathcal{A}_\rho} \int (-X)\,dc_\psi.$$

As another consequence of Theorem 4.64, we obtain an explicit description of the maximal representing set $\mathcal{Q}_\mu \subset \mathcal{M}_1(P)$ for the coherent risk measure ρ_μ.

Theorem 4.73. *Let μ be a probability measure on* [0, 1], *and let ψ be the corresponding concave function defined in Lemma* 4.63. *Then ρ_μ can be represented as*

$$\rho_\mu(X) = \sup_{Q \in \mathcal{Q}_\mu} E_Q[-X],$$

where the set $\mathcal{Q}_\mu$ is given by

$$\mathcal{Q}_\mu := \Big\{ Q \in \mathcal{M}_1(P) \,\Big|\, \varphi := \frac{dQ}{dP} \text{ satisfies } \int_t^1 q_\varphi(s)\, ds \le \psi(1-t) \text{ for } t \in (0,1) \Big\}.$$

Moreover, $\mathcal{Q}_\mu$ is the maximal subset of $\mathcal{M}_1(P)$ that represents ρ_μ.

Proof. The risk measure ρ_μ is coherent and continuous from above. By Corollary 4.34, it can be represented by taking the supremum of expectations over the set $\mathcal{Q}_{\max} = \{ Q \in \mathcal{M}_1(P) \,|\, \alpha_{\min}(Q) = 0 \}$. Using (4.39) and Theorem 4.64, we see that a measure $Q \in \mathcal{M}_1(P)$ with density $\varphi = dQ/dP$ belongs to $\mathcal{Q}_{\max}$ if and only if

$$\begin{aligned} \int_0^1 q_X(s) q_\varphi(s)\, ds &\le \rho_\mu(-X) \\ &= \psi(0+) AV@R_0(-X) + \int_0^1 q_X(s) \psi'(1-s)\, ds \end{aligned} \tag{4.45}$$

for all $X \in L^\infty$. For constant random variables $X \equiv t$, we have $q_X = \mathrm{I}_{[t,1]}$ a.e., and so we obtain

$$\int_t^1 q_\varphi(s)\, ds \le \psi(0+) + \int_t^1 \psi'(1-s)\, ds = \psi(1-t)$$

for all $t \in (0,1)$. Hence $\mathcal{Q}_{\max} \subset \mathcal{Q}_\mu$. For the proof of the converse inclusion, we show that the density φ of a fixed measure $Q \in \mathcal{Q}_\mu$ satisfies (4.45) for any given $X \in L^\infty$. To this end, let ν be the positive finite measure on [0, 1] such that $q_X^+(s) = \nu([0,s])$. Using Fubini's theorem and the definition of $\mathcal{Q}_\mu$, we get

$$\begin{aligned} \int_0^1 q_X(s) q_\varphi(s)\, ds &= \int_{[0,1]} \int_t^1 q_\varphi(s)\, ds\, \nu(dt) \\ &\le \int_{[0,1]} \psi(1-t)\, \nu(dt) \\ &= \psi(0+)\nu([0,1]) + \int_0^1 \psi'(1-s) \int_{[0,s]} \nu(dt)\, ds, \end{aligned}$$

which coincides with the right-hand side of (4.45). □

Corollary 4.74. *In the context of Theorem* 4.73*, the following conditions are equivalent.*

(a) ρ_μ *is continuous from below.*

(b) $\mu(\{0\}) = 0$.

(c) $\rho_\mu(X) = \max\limits_{Q \in \mathcal{Q}_\mu} E_Q[-X]$ *for all* $X \in L^\infty$.

If these equivalent conditions are satisfied, then the maximum in (c) *is attained by the measure* $Q_X \in \mathcal{Q}_\mu$ *with density* $dQ_X/dP = f(X)$*, where* f *is the decreasing function defined by*

$$f(x) := \psi'(F_X(x))$$

if x *is a continuity point of* F_X*, and by*

$$f(x) := \frac{1}{F_X(x) - F_X(x-)} \int_{F_X(x-)}^{F_X(x)} \psi'(t)\, dt$$

otherwise. Moreover, with λ *denoting the Lebesgue measure on* $(0, 1)$*,*

$$\mathcal{Q}_\mu = \Big\{ Q \ll P \;\Big|\; P \circ \Big(\frac{dQ}{dP}\Big)^{-1} \succcurlyeq_{\text{uni}} \lambda \circ (\psi')^{-1} \Big\}. \tag{4.46}$$

Proof. The equivalence of conditions (a) and (c) has already been proved in Corollary 4.35. If (b) holds, then ρ_μ is continuous from below, due to Theorem 4.47 and monotone convergence. Let us now show that condition (a) is not satisfied if $\delta := \mu(\{0\}) > 0$. In this case, we can write

$$\rho_\mu = \delta AV@R_0 + (1-\delta)\rho_{\mu'},$$

where $\mu' := \mu(\,\cdot\,|(0, 1])$. Then $\rho_{\mu'}$ is continuous from below since $\mu'(\{0\}) = 0$, but $AV@R_0$ is not, and so ρ_μ does not satisfy (a); see Remark 4.45.

Let us now prove the remaining assertions. Since $\psi(0+) = \mu(\{0\}) = 0$, a measure Q with density $\varphi = dQ/dP$ belongs to $\mathcal{Q}_\mu$ if and only if $\int_t^1 q_\varphi(s)\, ds \le \int_t^1 \psi'(1-s)\, ds$ for all t. Since $\psi'(1-t)$ is a quantile function for the law of ψ' under λ, part (e) of Theorem 2.58 implies (4.46). The problem of identifying the maximizing measure Q_X is hence equivalent to minimizing $E[\varphi X]$ under the constraint that φ is a density function such that $P \circ \varphi^{-1} \succcurlyeq_{\text{uni}} \lambda \circ (\psi')^{-1}$. Let us first assume that $X \ge 0$. Then it follows from Theorem 3.44 that $f(X)$ minimizes $E[YX]$ among all $Y \in L^1_+$ such that $P \circ Y^{-1} \succcurlyeq_{\text{uni}} \lambda \circ (\psi')^{-1}$. Moreover, Remark 3.46 shows that $E[f(X)] = \int_0^1 \psi'(t)\, dt = 1$, and so $\varphi_X := f(X) \ge 0$ is the density of an optimal probability measure $Q_X \in \mathcal{Q}_\mu$. If X is not positive, then we may take a constant c such that $X + c \ge 0$ and apply the preceding argument. The formula for f then follows from the fact that $F_{X+c}(X + c) = F_X(X)$. □

Remark 4.75. As long as we are interested in a law-invariant risk assessment, we can represent a financial position $X \in L^\infty$ by its distribution function F_X or, equivalently, by the function

$$G_X(t) := 1 - F_X(t) = P[\, X > t\,].$$

If we only consider positions X with values in $[0, 1]$ then their proxies G_X vary in the class of right-continuous decreasing functions G on $[0, 1]$ such that $G(1) = 0$ and $G(0) \le 1$. Due to Theorem 4.64, a law-invariant coherent risk measure ρ_μ induces a functional U on the class of proxies via

$$U(G_X) := \rho_\mu(-X) = \int_0^1 \psi(G_X(t))\, dt.$$

Since ψ is increasing and concave, the functional U has the form of a von Neumann–Morgenstern utility functional on the probability space given by Lebesgue measure on the unit interval $[0, 1]$. As such, it can be characterized by the axioms in Section 2.3, and this is the approach taken in Yaari's "dual theory of choice" [198]. More generally, we can introduce a utility function u on $[0, 1]$ with $u(0) = 0$ and consider the functional

$$U(G_X) := \int_0^1 \psi(G_X(t))\, du(t)$$

introduced by Quiggin [161]. For $u(x) = x$ this reduces to the "dual theory", for $\psi(x) = x$ we recover the classical utility functionals

$$\begin{aligned} U(G_X) &= \int_0^1 G_X(t)\, du(t) \\ &= -\int_0^1 u(t)\, dG_X(t) \\ &= E[\, u(X)\,] \end{aligned}$$

discussed in Section 2.3. ◇

4.7 Comonotonic risk measures

In many situations, the risk of a combined position $X + Y$ will be strictly lower than the sum of the individual risks, because one position serves as a hedge against adverse changes in the other position. If, on the other hand, there is no way for X to work as a hedge for Y then we may want the risk simply to add up. In order to make this idea precise, we introduce the notion of comonotonicity. Our main goal in this section is to characterize the class of all convex risk measures that share this property of comonotonicity.

As in the first two sections of this chapter, we will denote by $\mathcal{X}$ the linear space of all bounded measurable functions on the measurable space $(\Omega, \mathcal{F})$.

Definition 4.76. Two measurable functions X and Y on $(\Omega, \mathcal{F})$ are called *comonotone* if

$$\big(X(\omega) - X(\omega')\big)\big(Y(\omega) - Y(\omega')\big) \geq 0 \quad \text{for all } (\omega, \omega') \in \Omega \times \Omega. \tag{4.47}$$

A monetary measure of risk ρ on $\mathcal{X}$ is called *comonotonic* if

$$\rho(X + Y) = \rho(X) + \rho(Y)$$

whenever $X, Y \in \mathcal{X}$ are comonotone.

Lemma 4.77. *If ρ is a comonotonic monetary risk measure on $\mathcal{X}$, then ρ is positively homogeneous.*

Proof. Note that (X, X) is a comonotone pair. Hence $\rho(2X) = 2\rho(X)$. An iteration of this argument yields $\rho(rX) = r\rho(X)$ for all rational numbers $r \geq 0$. Positive homogeneity now follows from the Lipschitz continuity of ρ; see Lemma 4.3. □

We will see below that every comonotonic monetary risk measure on $\mathcal{X}$ arises as the Choquet integral with respect to a certain set function on $(\Omega, \mathcal{F})$. In the sequel, $c : \mathcal{F} \to [0, 1]$ will always denote a set function that is normalized and monotone; see Definition 4.68. Unless otherwise mentioned, we will not assume that c enjoys any additivity properties. Recall from Definition 4.70 that the Choquet integral of $X \in \mathcal{X}$ with respect to c is defined as

$$\int X \, dc = \int_{-\infty}^{0} (c(X > x) - 1) \, dx + \int_{0}^{\infty} c(X > x) \, dx.$$

The proof of the following proposition was already given in Example 4.13.

Proposition 4.78. *The Choquet integral of the loss,*

$$\rho(X) := \int (-X) \, dc,$$

is a monetary risk measure on $\mathcal{X}$ which is positively homogeneous.

Definition 4.79. Let X be a measurable function on $(\Omega, \mathcal{F})$. An inverse function $r_X : (0, 1) \to \mathbb{R}$ of the increasing function $G_X(x) := 1 - c(X > x)$, taken in the sense of Definition A.14, is called a *quantile function* for X with respect to c.

If c is a probability measure, then $G_X(x) = c(X \leq x)$. Hence, the preceding definition extends the notion of a quantile function given in Definition A.20. The following proposition yields an alternative representation of the Choquet integral in terms of quantile functions with respect to c.

Proposition 4.80. *Let r_X be a quantile function with respect to c for $X \in \mathcal{X}$. Then*

$$\int X \, dc = \int_0^1 r_X(t) \, dt.$$

Proof. We have $\int (X+m)\,dc = \int X\,dc + m$, and one easily checks that $r_{X+m} = r_X + m$ a.e. for all $m \in \mathbb{R}$ and each quantile function r_{X+m} of $X+m$. Thus, we may assume without loss of generality that $X \geq 0$. In this case, Remark A.16 and Lemma A.15 imply that the largest quantile function r_X^+ is given by

$$r_X^+(t) = \sup\{\, x \geq 0 \mid G_X(x) \leq t \,\} = \int_0^\infty \mathrm{I}_{\{G_X(x)\leq t\}}\,dx.$$

Since $r_X = r_X^+$ a.e. on (0, 1), Fubini's theorem implies

$$\begin{aligned}\int_0^1 r_X(t)\,dt &= \int_0^1 \int_0^\infty \mathrm{I}_{\{G_X(x)\leq t\}}\,dx\,dt \\ &= \int_0^\infty (1 - G_X(x))\,dx \\ &= \int X\,dc.\end{aligned}$$ □

The preceding proposition yields the following generalization of Corollary 4.66 when applied to a continuous distortion of a probability measure as defined in Definition 4.67.

Corollary 4.81. *Let $c_\psi(A) = \psi(P[\,A\,])$ be the distortion of the probability measure P with respect to the continuous distortion function ψ. If φ is an inverse function for the increasing function ψ in the sense of Definition* A.14*, then the Choquet integral with respect to c_ψ satisfies*

$$\int X\,dc_\psi = \int_0^1 q_X\big(1 - \varphi(t)\big)\,dt,$$

where q_X is a quantile function for $X \in \mathcal{X}$, taken with respect to P.

Proof. Due to the continuity of ψ, we have $\psi(a) \leq t$ if and only if $a \leq \varphi^+(t) = \inf\{x \mid \psi(x) > t\}$. Thus, we can compute the lower quantile function of X with respect to c_ψ:

$$\begin{aligned}r_X^-(t) &= \inf\{\, x \in \mathbb{R} \mid 1 - c_\psi(X > x) \geq t\} \\ &= \inf\{\, x \in \mathbb{R} \mid \psi(P[\,X > x\,]) \leq 1 - t\} \\ &= \inf\{\, x \in \mathbb{R} \mid P[\,X > x\,] \leq \varphi^+(1-t)\} \\ &= q_X^-(1 - \varphi^+(1-t)).\end{aligned}$$

Next note that $\varphi^+(t) = \varphi(t)$ for a.e. t. Moreover, φ has the continuous distribution function ψ under the Lebesgue measure, and so we can replace q_X^- by the arbitrary quantile function q_X. □

Theorem 4.82. *A monetary risk measure ρ on $\mathcal{X}$ is comonotonic if and only if there exists a normalized monotone set function c on $(\Omega, \mathcal{F})$ such that*

$$\rho(X) = \int (-X)\, dc, \quad X \in \mathcal{X}.$$

In this case, c is given by $c(A) = \rho(-I_A)$.

The preceding theorem implies in view of Corollary 4.71 that all mixtures

$$\rho_\mu = \int_{[0,1]} AV@R_\lambda\, \mu(d\lambda)$$

are comonotonic. We will see in Theorem 4.87 below that these are in fact all convex risk measures that are law-invariant and comonotonic. The proof of Theorem 4.82 requires a further analysis of comonotone random variables.

Lemma 4.83. *Two measurable functions X and Y on $(\Omega, \mathcal{F})$ are comonotone if and only if there exists a third measurable function Z on $(\Omega, \mathcal{F})$ and increasing functions f and g on $\mathbb{R}$ such that $X = f(Z)$ and $Y = g(Z)$.*

Proof. Clearly, $X := f(Z)$ and $Y := g(Z)$ are comonotone for given Z, f, and g. Conversely, suppose that X and Y are comonotone and define Z by $Z := X + Y$. We show that $z := Z(\omega)$ has a unique decomposition as $z = x + y$, where $(x, y) = (X(\omega'), Y(\omega'))$ for some $\omega' \in \Omega$. Having established this, we can put $f(z) := x$ and $g(z) := y$. The existence of the decomposition as $z = x + y$ follows by taking $x := X(\omega)$ and $y := Y(\omega)$, so it remains to show that these are the only possible values x and y. To this end, let us suppose that $X(\omega) + Y(\omega) = z = X(\omega') + Y(\omega')$ for some $\omega' \in \Omega$. Then

$$X(\omega) - X(\omega') = -\big(Y(\omega) - Y(\omega')\big),$$

and comonotonicity implies that this expression vanishes. Hence $x = X(\omega')$ and $y = Y(\omega')$.

Next, we check that both f and g are increasing functions on $Z(\Omega)$. So let us suppose that

$$X(\omega_1) + Y(\omega_1) = z_1 \le z_2 = X(\omega_2) + Y(\omega_2).$$

This implies

$$X(\omega_1) - X(\omega_2) \le -\big(Y(\omega_1) - Y(\omega_2)\big).$$

Comonotonicity thus yields that $X(\omega_1) - X(\omega_2) \le 0$ and $Y(\omega_1) - Y(\omega_2) \le 0$, whence $f(z_1) \le f(z_2)$ and $g(z_1) \le g(z_2)$. Thus, f and g are increasing on $Z(\Omega)$, and it is straightforward to extend them to increasing functions defined on $\mathbb{R}$. □

Lemma 4.84. *If $X, Y \in \mathcal{X}$ is a pair of comonotone functions, and r_X, r_Y, r_{X+Y} are quantile functions with respect to c, then*

$$r_{X+Y}(t) = r_X(t) + r_Y(t) \quad \text{for a.e. } t.$$

Proof. Write $X = f(Z)$ and $Y = g(Z)$ as in Lemma 4.83. The same argument as in the proof of Lemma A.23 shows that $f(r_Z)$ and $g(r_Z)$ are quantile functions for X and Y under c if r_Z is a quantile function for Z. An identical argument applied to the increasing function $h := f + g$ shows that $h(r_Z) = f(r_Z) + g(r_Z)$ is a quantile function for $X + Y$. The assertion now follows from the fact that all quantile functions of a random variable coincide almost everywhere, due to Lemma A.15. □

Remark 4.85. Applied to the special case of quantile function with respect to a probability measure, the preceding lemma yields that $V@R_\lambda$ and $AV@R_\lambda$ are comonotonic. ◇

Proof of Theorem 4.82. We already know from Proposition 4.78 that the Choquet integral of the loss is a monetary risk measure. Comonotonicity follows by combining Proposition 4.80 with Lemma 4.84.

Conversely, suppose now that ρ is comonotonic. Then ρ is coherent according to Lemma 4.77. Thus, we obtain a normalized monotone set function by letting $c(A) := \rho(-\mathrm{I}_A)$. Moreover, $\rho_c(X) := \int(-X)\,dc$ is a comonotonic monetary risk measure on $\mathcal{X}$ that coincides with ρ on indicator functions: $\rho(-\mathrm{I}_A) = c(A) = \rho_c(-\mathrm{I}_A)$. Let us now show that ρ and ρ_c coincide on simple random variables of the form

$$X = \sum_{i=1}^{n} x_i \mathrm{I}_{A_i}, \quad x_i \in \mathbb{R},\ A_i \in \mathcal{F}.$$

Since these random variables are dense in L^∞, Lemma 4.3 will then imply that $\rho = \rho_c$. In order to show that $\rho_c(X) = \rho(X)$ for X as above, we may assume without loss of generality that $x_1 \geq x_2 \geq \cdots \geq x_n$ and that the sets A_i are disjoint. By cash invariance, we may also assume $X \geq 0$, i.e., $x_n \geq 0$. Thus, we can write $X = \sum_{i=1}^{n} b_i \mathrm{I}_{B_i}$, where $b_i := x_i - x_{i+1} \geq 0$, $x_{n+1} := 0$, and $B_i := \bigcup_{k=1}^{i} A_k$. Note that $b_i \mathrm{I}_{B_i}$ and $b_k \mathrm{I}_{B_k}$ is a pair of comonotone functions. Hence, also $\sum_{i=1}^{k-1} b_i \mathrm{I}_{B_i}$ and $b_k \mathrm{I}_{B_k}$ are comonotone, and we get inductively

$$\rho(-X) = \sum_{i=1}^{n} b_i\, \rho(-\mathrm{I}_{B_i}) = \sum_{i=1}^{n} b_i\, \rho_c(-\mathrm{I}_{B_i}) = \rho_c(-X). \qquad \square$$

Remark 4.86. The argument at the end of the preceding proof shows that the Choquet integral of a simple random variable

$$X = \sum_{i=1}^{n} x_i \mathrm{I}_{A_i} \quad \text{with } x_1 \geq \cdots \geq x_n \geq x_{n+1} := 0$$

can be computed as

$$\int X\,dc = \sum_{i=1}^{n} (x_i - x_{i+1}) c(B_i) = \sum_{i=1}^{n} x_i \big(c(B_i) - c(B_{i-1})\big),$$

where $B_0 := \emptyset$ and $B_i := \bigcup_{k=1}^{i} A_k$ for $i = 1, \dots, n$. ◇

So far, we have shown that comonotonic monetary risk measures can be identified with Choquet integrals of normalized monotone set functions. Our next goal is to characterize those set functions that induce risk measures with the additional property of *convexity*. To this end, we will first consider law-invariant risk measures. The following result shows that the risk measures $AV@R_\lambda$ may be viewed as the extreme points in the convex class of all law-invariant convex risk measures on L^∞ that are comonotonic.

Theorem 4.87. *On an atomless probability space, the class of risk measures*

$$\rho_\mu(X) := \int AV@R_\lambda(X)\,\mu(d\lambda), \quad \mu \in \mathcal{M}_1([0,1]),$$

is precisely the class of all law-invariant convex risk measures on L^∞ *that are comonotonic. In particular, any convex measure of risk that is law-invariant and comonotonic is also coherent and continuous from above.*

Proof. Comonotonicity of ρ_μ follows from Corollary 4.71 and Theorem 4.82. Conversely, let us assume that ρ is a law-invariant convex measure of risk that is also comonotonic. By Theorem 4.82, $\rho(X) = \int(-X)\,dc$ for $c(A) := \rho(-\mathrm{I}_A)$. The law-invariance of ρ implies that $c(A)$ is a function of the probability $P[\,A\,]$, i.e., there exists an increasing function ψ on $[0,1]$ such that $\psi(0) = 0$, $\psi(1) = 1$, and $c(A) = \psi(P[\,A\,])$. Note that $\mathrm{I}_{A\cup B}$ and $\mathrm{I}_{A\cap B}$ is a pair of comonotone functions for all $A, B \in \mathcal{F}$. Hence, comonotonicity and subadditivity of ρ imply

$$\begin{aligned} c(A\cap B) + c(A\cup B) &= \rho(-\mathrm{I}_{A\cap B}) + \rho(-\mathrm{I}_{A\cup B}) = \rho\big(-\mathrm{I}_{A\cap B} - \mathrm{I}_{A\cup B}\big) \\ &= \rho\big(-\mathrm{I}_A - \mathrm{I}_B\big) \\ &\le c(A) + c(B). \end{aligned} \tag{4.48}$$

Proposition 4.69 thus implies that ψ is concave. Corollary 4.71 finally shows that the Choquet integral with respect to c can be identified with a risk measure ρ_μ, where μ is obtained from ψ via Lemma 4.63. □

Now we turn to the characterization of *all* comonotonic convex risk measures on $\mathcal{X}$. Recall that, for a positively homogeneous monetary risk measure, convexity is equivalent to subadditivity. Also recall that $\mathcal{M}_{1,f} := \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ denotes the set of all finitely additive normalized set functions $Q : \mathcal{F} \to [0,1]$, and that $E_Q[\,X\,]$ denotes the integral of $X \in \mathcal{X}$ with respect to $Q \in \mathcal{M}_{1,f}$, as constructed in Theorem A.50.

Theorem 4.88. *For the Choquet integral with respect to a normalized monotone set function c, the following conditions are equivalent.*

(a) $\rho(X) := \int(-X)\,dc$ *is a convex risk measure on* $\mathcal{X}$.

(b) $\rho(X) := \int(-X)\,dc$ *is a coherent risk measure on* $\mathcal{X}$.

(c) *For* $\mathcal{Q}_c := \{Q \in \mathcal{M}_{1,f} \mid Q[A] \le c(A)$ *for all* $A \in \mathcal{F}\}$,

$$\int X\,dc = \max_{Q \in \mathcal{Q}_c} E_Q[X] \quad \textit{for } X \in \mathcal{X}.$$

(d) *The set function c is submodular.*

In this case, $\mathcal{Q}_c$ *is equal to the maximal representing set* $\mathcal{Q}_{\max}$ *for* ρ.

Before giving the proof of this theorem, let us state the following corollary, which gives a complete characterization of all comonotonic convex risk measures, and a remark concerning the set $\mathcal{Q}_c$ in part (c), which is usually called the *core* of c.

Corollary 4.89. *A convex risk measure on* $\mathcal{X}$ *is comonotonic if and only if it arises as the Choquet integral of the loss with respect to a submodular, normalized, and monotone set function c. In this case, c is given by* $c(A) = \rho(-I_A)$, *and* ρ *has the representation*

$$\rho(X) = \max_{Q \in \mathcal{Q}_c} E_Q[-X],$$

where $\mathcal{Q}_c = \{Q \in \mathcal{M}_{1,f} \mid Q[A] \le c(A)$ *for all* $A \in \mathcal{F}\}$ *is equal to the maximal representing set* $\mathcal{Q}_{\max}$.

Proof. Theorems 4.82 and 4.88 state that $\rho(X) := \int (-X)\,dc$ is a comonotonic coherent risk measure, which can be represented as in the assertion, as soon as c is a submodular, normalized, and monotone set function. Conversely, any comonotonic convex risk measure ρ is coherent and arises as the Choquet integral of $c(A) := \rho(-I_A)$, due to Theorem 4.82. Theorem 4.88 then gives the submodularity of c. □

Remark 4.90. Let c be a normalized monotone submodular set function. Theorem 4.88 implies in particular that the core $\mathcal{Q}_c$ of c is non-empty. Moreover, c can be recovered from $\mathcal{Q}_c$:

$$c(A) = \max_{Q \in \mathcal{Q}_c} Q[A] \quad \text{for all } A \in \mathcal{F}.$$

If c has the additional continuity property that $c(A_n) \to 0$ for any decreasing sequence (A_n) of events such that $\bigcap_n A_n = \emptyset$, then this property is shared by any $Q \in \mathcal{Q}_c$, and it follows that Q is σ-additive. Thus, the corresponding coherent risk measure $\rho(X) = \int (-X)\,dc$ admits a representation in terms of σ-additive probability measures. It follows by Lemma 4.20 that ρ is continuous from above. ◇

The proof of Theorem 4.88 requires some preparations. The assertion of the following lemma is not entirely obvious, since Fubini's theorem may fail if $Q \in \mathcal{M}_{1,f}$ is not σ-additive.

Lemma 4.91. *For* $X \in \mathcal{X}$ *and* $Q \in \mathcal{M}_{1,f}$, *the integral* $E_Q[X]$ *is equal to the Choquet integral* $\int X\,dQ$.

Proof. It is enough to prove the result for $X \geq 0$. Suppose first that $X = \sum_{i=1}^{n} x_i \mathrm{I}_{A_i}$ is as in Remark 4.86. Then

$$\int X\,dQ = \sum_{i=1}^{n}(x_i - x_{i+1})Q\Big[\bigcup_{k=1}^{i} A_k\Big] = \sum_{i=1}^{n} x_i\,Q[\,A_i\,] = E_Q[\,X\,].$$

The result for general $X \in \mathcal{X}$ follows by approximating X uniformly with X_n which take only finitely many values, and by using the Lipschitz continuity of both $E_Q[\,\cdot\,]$ and $\int \cdot\,dQ$ with respect to the supremum norm. □

Lemma 4.92. *Let $A_1, \dots, A_n$ be a partition of Ω into disjoint measurable sets, and suppose that the normalized monotone set function c is submodular. Let Q be the probability measure on $\mathcal{F}_0 := \sigma(A_1, \dots, A_n)$ with weights*

$$Q[\,A_k\,] := c(B_k) - c(B_{k-1}) \quad \textit{for } B_0 := \emptyset \textit{ and } B_k := \bigcup_{j=1}^{k} A_j,\ k \geq 1. \tag{4.49}$$

Then $\int X\,dc \geq E_Q[\,X\,]$ for all bounded $\mathcal{F}_0$-measurable $X = \sum_{i=1}^{n} x_i \mathrm{I}_{A_i}$, and equality holds if the values of X are arranged in decreasing order: $x_1 \geq \cdots \geq x_n$.

Proof. Clearly, it suffices to consider only the case $X \geq 0$. Then Remark 4.86 implies $\int X\,dc = E_Q[\,X\,]$ as soon as the values of X are arranged in decreasing order.

Now we prove $\int X\,dc \geq E_Q[\,X\,]$ for arbitrary $\mathcal{F}_0$-measurable X. To this end, note that any permutation σ of $\{1, \dots, n\}$ induces a probability measure Q_σ on $\mathcal{F}_0$ by applying the definition of Q to the re-labeled partition $A_{\sigma(1)} \dots, A_{\sigma(n)}$. If σ is a permutation such that $x_{\sigma(1)} \geq \cdots \geq x_{\sigma(n)}$, then we have $\int X\,dc = E_{Q_\sigma}[\,X\,]$, and so the assertion will follow if we can prove that $E_{Q_\sigma}[\,X\,] \geq E_Q[\,X\,]$. To this end, it is enough to show that $E_{Q_\tau}[\,X\,] \geq E_Q[\,X\,]$ if τ is the transposition of two indices i and $i+1$ which are such that $x_i < x_{i+1}$, because σ can be represented as a finite product of such transpositions.

Note next that

$$E_{Q_\tau}[\,X\,] - E_Q[\,X\,] = x_i(Q_\tau[\,A_i\,] - Q[\,A_i\,]) + x_{i+1}(Q_\tau[\,A_{i+1}\,] - Q[\,A_{i+1}\,]). \tag{4.50}$$

To compute the probabilities $Q_\tau[\,A_k\,]$, let us introduce

$$B_0^\tau := \emptyset \quad \text{and} \quad B_k^\tau := \bigcup_{j=1}^{k} A_{\tau(j)}, \quad k = 1, \dots, n.$$

Then $B_k^\tau = B_k$ for $k \neq i$. Hence,

$$\begin{aligned} Q_\tau[\,A_i\,] + Q_\tau[\,A_{i+1}\,] &= Q_\tau[\,A_{\tau(i)}\,] + Q_\tau[\,A_{\tau(i+1)}\,] = c(B_{i+1}^\tau) - c(B_{i-1}^\tau) \\ &= c(B_{i+1}) - c(B_{i-1}) = Q[\,A_i\,] + Q[\,A_{i+1}\,]. \end{aligned} \tag{4.51}$$

Moreover, $B_i^\tau \cap B_i = B_{i-1}$, $B_i^\tau \cup B_i = B_{i+1}$, and hence $c(B_{i-1}) + c(B_{i+1}) \le c(B_i^\tau) + c(B_i)$, due to the submodularity of c. Thus,

$$Q[\, A_{i+1}\,] = c(B_{i+1}) - c(B_i) \le c(B_i^\tau) - c(B_{i-1}^\tau) = Q_\tau[\, A_{\tau(i)}] = Q_\tau[\, A_{i+1}].$$

Using (4.50), (4.51), and our assumption $x_i < x_{i+1}$ thus yields $E_{Q_\tau}[\, X\,] \ge E_Q[\, X\,]$. □

Proof of Theorem 4.88. (a)⇔(b): According to Proposition 4.78, the property of positive homogeneity is shared by all Choquet integrals, and the implication (b) ⇒ (a) is obvious.

(b) ⇒ (c): By Corollary 4.18, $\rho(-X) = \max_{Q \in \mathcal{Q}_{\max}} E_Q[\, X\,]$, where $Q \in \mathcal{M}_{1,f}$ belongs to $\mathcal{Q}_{\max}$ if and only if

$$E_Q[\, X\,] \le \rho(-X) = \int X\, dc \quad \text{for all } X \in \mathcal{X}. \tag{4.52}$$

We will now show that this set $\mathcal{Q}_{\max}$ coincides with the set $\mathcal{Q}_c$. If $Q \in \mathcal{Q}_{\max}$ then, in particular, $Q[\, A\,] \le \int I_A\, dc = c(A)$ for all $A \in \mathcal{F}$. Hence $Q \in \mathcal{Q}_c$. Conversely, suppose $Q \in \mathcal{Q}_c$. If $X \ge 0$ then

$$\int X\, dc = \int_0^\infty c(X > x)\, dx \ge \int_0^\infty Q[\, X > x\,]\, dx = E_Q[\, X\,],$$

where we have used Lemma 4.91. Cash invariance yields (4.52).

(c) ⇒ (b) is obvious.

(b) ⇒ (d): This follows precisely as in (4.48).

(d) ⇒ (a): We have to show that the Choquet integral is subadditive. By Lemma 4.3, it is again enough to prove this for random variables which only take finitely many values. Thus, let $A_1, \dots, A_n$ be a partition of Ω into finitely many disjoint measurable sets. Let us write $X = \sum_i x_i I_{A_i}$, $Y = \sum_i y_i I_{A_i}$, and let us assume that the indices $i = 1, \dots, n$ are arranged such that $x_1 + y_1 \ge \cdots \ge x_n + y_n$. Then the probability measure Q constructed in Lemma 4.92 is such that

$$\int (X + Y)\, dc = E_Q[\, X + Y\,] = E_Q[\, X\,] + E_Q[\, Y\,] \le \int X\, dc + \int Y\, dc.$$

But this is the required subadditivity of the Choquet integral. □

4.8 Measures of risk in a financial market

In this section, we will consider risk measures which arise in the financial market model of Section 1.1. In this model, $d + 1$ assets are priced at times $t = 0$ and $t = 1$. Prices at time 1 are modelled as non-negative random variables $S^0, S^1, \dots, S^d$ on some probability space $(\Omega, \mathcal{F}, P)$, with $S^0 \equiv 1 + r$. Prices at time 0 are given by a vector $\overline{\pi} = (1, \pi)$, with $\pi = (\pi^1, \dots, \pi^d)$. The discounted net gain of a trading strategy $\overline{\xi} = (\xi^0, \xi)$ is given by $\xi \cdot Y$, where the random vector $Y = (Y^1, \dots, Y^d)$ is defined by

$$Y^i = \frac{S^i}{1+r} - \pi^i \quad \text{for } i = 1, \dots, d.$$

As in the previous two sections, risk measures will be defined on the space $L^\infty = L^\infty(\Omega, \mathcal{F}, P)$. A financial position X can be viewed as riskless if $X \geq 0$ or, more generally, if X can be hedged without additional costs, i.e., if there exists a trading strategy $\overline{\xi} = (\xi^0, \xi)$ such that $\overline{\pi} \cdot \overline{\xi} = 0$ and

$$X + \frac{\overline{\xi} \cdot \overline{S}}{1+r} = X + \xi \cdot Y \geq 0 \quad P\text{-a.s.} \tag{4.53}$$

Thus, we define the following set of acceptable positions in L^∞:

$$\mathcal{A}_0 := \big\{ X \in L^\infty \mid \exists\, \xi \in \mathbb{R}^d \text{ with } X + \xi \cdot Y \geq 0 \ P\text{-a.s.} \big\}.$$

Proposition 4.93. *Suppose that* $\inf\{m \in \mathbb{R}^d \mid m \in \mathcal{A}_0\} > -\infty$. *Then* $\rho_0 := \rho_{\mathcal{A}_0}$ *is a coherent measure of risk. Moreover,* ρ_0 *is sensitive in the sense of Definition* 4.32 *if and only if the market model is arbitrage-free. In this case,* ρ_0 *is continuous from above and can be represented in terms of the set* $\mathcal{P}$ *of equivalent risk-neutral measures:*

$$\rho_0(X) = \sup_{P^* \in \mathcal{P}} E^*[\,-X\,]. \tag{4.54}$$

Proof. The fact that ρ_0 is a coherent measure of risk follows from Proposition 4.7. If the model is arbitrage-free, then Theorem 1.31 yields the representation (4.54), and it follows that ρ_0 is sensitive and continuous from above.

Conversely, suppose that ρ_0 is sensitive, but the market model admits an arbitrage opportunity. Then there are $\xi \in \mathbb{R}^d$ and $\varepsilon > 0$ such that $0 \leq \xi \cdot Y$ P-a.s. and $A := \{\xi \cdot Y \geq \varepsilon\}$ satisfies $P[\,A\,] > 0$. It follows that $\xi \cdot Y - \varepsilon I_A \geq 0$, i.e., $-\varepsilon I_A$ is acceptable. However, the sensitivity of ρ_0 implies that

$$\rho_0(-\varepsilon I_A) = \varepsilon \rho_0(-I_A) > \rho_0(0) = 0,$$

where we have used the coherence of ρ_0, which follows from fact that $\mathcal{A}_0$ is a cone. Thus, we arrive at a contradiction. □

There are several reasons why it may make sense to allow in (4.53) only strategies ξ that belong to a proper subset $\mathcal{S}$ of the class $\mathbb{R}^d$ of all strategies. For instance, if the resources available to an investor are limited, only those strategies should be considered for which the initial investment in risky assets is below a certain amount. Such a restriction corresponds to an upper bound on $\xi \cdot \pi$. There may be other constraints. For instance, *short sales constraints* are lower bounds on the number of shares in the portfolio. In view of *market illiquidity*, the investor may also wish to avoid holding too many shares of one single asset, since the market capacity may not suffice to resell the shares. Such constraints will be taken into account by assuming throughout the remainder of this section that $\mathcal{S}$ has the following properties:

- $0 \in \mathcal{S}$.
- $\mathcal{S}$ is convex.
- Each $\xi \in \mathcal{S}$ is *admissible* in the sense that $\xi \cdot Y$ is P-a.s. bounded from below.

Under these conditions, the set

$$\mathcal{A}^{\mathcal{S}} := \left\{ X \in L^\infty \mid \exists\, \xi \in \mathcal{S} \text{ with } X + \xi \cdot Y \geq 0 \ P\text{-a.s.} \right\} \tag{4.55}$$

is non-empty, convex, and contains all $X \in \mathcal{X}$ which dominate some $Z \in \mathcal{A}^{\mathcal{S}}$. Moreover, we will assume from now on that

$$\inf\{ m \in \mathbb{R} \mid m \in \mathcal{A}^{\mathcal{S}} \} > -\infty. \tag{4.56}$$

Proposition 4.7 then guarantees that the induced risk measure

$$\rho^{\mathcal{S}}(X) := \rho_{\mathcal{A}^{\mathcal{S}}}(X) = \inf\{m \in \mathbb{R} \mid m + X \in \mathcal{A}^{\mathcal{S}}\}$$

is a convex risk measure on L^∞. Note that (4.56) holds, in particular, if $\mathcal{S}$ does not contain arbitrage opportunities in the sense that $\xi \cdot Y \geq 0$ P-a.s. for $\xi \in \mathcal{S}$ implies $P[\, \xi \cdot Y = 0\,] = 1$.

Remark 4.94. Admissibility of portfolios is a serious restriction; in particular, it prevents unhedged short sales of any unbounded asset. Note, however, that it is consistent with our notion of acceptability for bounded claims in (4.55), since $X + \xi \cdot Y \geq 0$ implies $\xi \cdot Y \geq -\|X\|$. ◇

Two questions arise: When is $\rho^{\mathcal{S}}$ continuous from above, and thus admits a representation (4.29) in terms of probability measures? And, if such a representation exists, how can we identify the minimal penalty function $\alpha^{\mathcal{S}}_{\min}$ on $\mathcal{M}_1(P)$? In the case $\mathcal{S} = \mathbb{R}^d$, both questions were addressed in Proposition 4.93. For general $\mathcal{S}$, only the second question has a straightforward answer, which will be given in Proposition 4.96. As can be seen from the proof of Proposition 4.93, an analysis of the first question requires an extension of the arbitrage theory in Chapter 1 for the case of portfolio constraints. Such a theory will be developed in Chapter 9 in a more general dynamic setting, and we will address both questions for the corresponding risk measures in Corollary 9.30. This result implies the following theorem for the simple one-period model of the present section:

Theorem 4.95. *In addition to the above assumptions, suppose that the market model is non-redundant in the sense of Definition* 1.13 *and that $\mathcal{S}$ is a closed subset of $\mathbb{R}^d$.*

Then $\rho^{\mathcal{S}}$ is sensitive if and only if $\mathcal{S}$ contains no arbitrage opportunities. In this case, $\rho^{\mathcal{S}}$ is continuous from above and admits the representation

$$\rho^{\mathcal{S}}(X) = \sup_{Q \in \mathcal{M}_1(P)} \big(E_Q[-X] - \sup_{\xi \in \mathcal{S}} E_Q[\xi \cdot Y] \big). \tag{4.57}$$

In the following proposition, we will explain the specific form of the penalty function in (4.57). This result will not require the additional assumptions of Theorem 4.95.

Proposition 4.96. *For $Q \in \mathcal{M}_1(P)$, the minimal penalty function $\alpha^{\mathcal{S}}_{\min}$ of $\rho^{\mathcal{S}}$ is given by*

$$\alpha^{\mathcal{S}}_{\min}(Q) = \sup_{\xi \in \mathcal{S}} E_Q[\xi \cdot Y].$$

In particular, $\rho^{\mathcal{S}}$ can be represented as in (4.57) *if $\rho^{\mathcal{S}}$ is continuous from above and, especially, if $\mathcal{S}$ does not contain arbitrage opportunities.*

Proof. Fix $Q \in \mathcal{M}_1(P)$. Clearly, the expectation $E_Q[\xi \cdot Y]$ is well defined for each $\xi \in \mathcal{S}$ by admissibility. If $X \in \mathcal{A}^{\mathcal{S}}$, there exists $\eta \in \mathcal{S}$ such that $-X \leq \eta \cdot Y$ P-almost surely. Thus,

$$E_Q[-X] \leq E_Q[\eta \cdot Y] \leq \sup_{\xi \in \mathcal{S}} E_Q[\xi \cdot Y]$$

for any $Q \in \mathcal{M}_1(P)$. Hence, the definition of the minimal penalty function yields

$$\alpha^{\mathcal{S}}_{\min}(Q) \leq \sup_{\xi \in \mathcal{S}} E_Q[\xi \cdot Y].$$

To prove the converse inequality, take $\xi \in \mathcal{S}$. Note that $X_k := -((\xi \cdot Y) \wedge k)$ is bounded since ξ is admissible. Moreover,

$$X_k + \xi \cdot Y = (\xi \cdot Y - k)\, \mathrm{I}_{\{\xi \cdot Y \geq k\}} \geq 0,$$

so that $X_k \in \mathcal{A}^{\mathcal{S}}$. Hence,

$$\alpha^{\mathcal{S}}_{\min}(Q) \geq E_Q[-X_k] = E_Q[(\xi \cdot Y) \wedge k],$$

and so $\alpha^{\mathcal{S}}_{\min}(Q) \geq E_Q[\xi \cdot Y]$ by monotone convergence. □

Remark 4.97. Suppose that $\mathcal{S}$ is a cone. Then the acceptance set $\mathcal{A}^{\mathcal{S}}$ is also a cone, and $\rho^{\mathcal{S}}$ is a coherent measure of risk. If $\rho^{\mathcal{S}}$ is continuous from above, then Corollary 4.34 yields the representation

$$\rho^{\mathcal{S}}(X) = \sup_{Q \in \mathcal{Q}^{\mathcal{S}}_{\max}} E_Q[-X]$$

in terms of the non-empty set $\mathcal{Q}^{\mathcal{S}}_{\max} = \{ Q \in \mathcal{M}_1(P) \mid \alpha^{\mathcal{S}}_{\min}(Q) = 0 \}$. It follows from Proposition 4.96 that for $Q \in \mathcal{M}_1(P)$

$$Q \in \mathcal{Q}^{\mathcal{S}}_{\max} \text{ if and only if } E_Q[\xi \cdot Y] \leq 0 \text{ for all } \xi \in \mathcal{S}.$$

If $\rho^{\mathcal{S}}$ is sensitive then the set $\mathcal{S}$ cannot contain any arbitrage opportunities, and $\mathcal{Q}^{\mathcal{S}}_{\max}$ contains the set $\mathcal{P}$ of all equivalent martingale measures whenever such measures exist. More precisely, $\mathcal{Q}^{\mathcal{S}}_{\max}$ can be described as the set of absolutely continuous supermartingale measures with respect to $\mathcal{S}$; this will be discussed in more detail in the dynamical setting of Chapter 9. ◇

Let us now relax the condition of acceptability in (4.55). We no longer insist that the final outcome of an acceptable position, suitably hedged, should always be non-negative. Instead, we only require that the hedged position is acceptable in terms of a given convex risk measure $\rho_{\mathcal{A}}$ with acceptance set $\mathcal{A}$. Thus, we define

$$\bar{\mathcal{A}} := \big\{ X \in L^\infty \mid \exists\, \xi \in \mathcal{S},\ A \in \mathcal{A} \text{ with } X + \xi \cdot Y \geq A \ P\text{-a.s.} \big\}. \tag{4.58}$$

Clearly, $\mathcal{A} \subset \bar{\mathcal{A}}$ and hence

$$\rho_{\mathcal{A}} \geq \overline{\rho} := \rho_{\bar{\mathcal{A}}}.$$

From now on, we assume that

$$\overline{\rho} > -\infty,$$

which implies our assumption (4.56) for $\mathcal{A}^{\mathcal{S}}$.

Proposition 4.98. *The minimal penalty function $\overline{\alpha}_{\min}$ for $\overline{\rho}$ is given by*

$$\overline{\alpha}_{\min}(Q) = \alpha^{\mathcal{S}}_{\min}(Q) + \alpha_{\min}(Q),$$

where $\alpha^{\mathcal{S}}_{\min}$ is the minimal penalty function for $\rho^{\mathcal{S}}$, and $\alpha_{\min}$ is the minimal penalty function for $\rho_{\mathcal{A}}$.

Proof. We claim that

$$\bar{\mathcal{A}} = \{ X^{\mathcal{S}} + A \mid X^{\mathcal{S}} \in \mathcal{A}^{\mathcal{S}},\ A \in \mathcal{A} \}. \tag{4.59}$$

If $X \in \bar{\mathcal{A}}$, then there exists $A \in \mathcal{A}$ and $\xi \in \mathcal{S}$ such that $\overline{X} + \xi \cdot Y \geq A$. Therefore $X^{\mathcal{S}} := X - A \in \mathcal{A}^{\mathcal{S}}$. Conversely, if $X^{\mathcal{S}} \in \mathcal{A}^{\mathcal{S}}$ then $X^{\mathcal{S}} + \xi \cdot Y \geq 0$ for some $\xi \in \mathcal{S}$. Hence, for any $A \in \mathcal{A}$, we get $X^{\mathcal{S}} + A + \xi \cdot Y \geq X \in \mathcal{A}$, i.e., $X := X^{\mathcal{S}} + A \in \bar{\mathcal{A}}$.

In view of (4.59), we have

$$\begin{aligned}\overline{\alpha}_{\min}(Q) &= \sup_{X \in \bar{\mathcal{A}}} E_Q[\,-X\,] \\ &= \sup_{X^{\mathcal{S}} \in \mathcal{A}^{\mathcal{S}}} \sup_{A \in \mathcal{A}} E_Q[\,-X^{\mathcal{S}} - A\,] \\ &= \alpha^{\mathcal{S}}_{\min}(Q) + \alpha_{\min}(Q).\end{aligned}$$

□

For the rest of this section, we consider the following case study, which is based on [36]. Let us fix a finite class

$$\mathcal{Q}_0 = \{Q_1, \dots, Q_n\}$$

of equivalent probability measures $Q_i \approx P$ such that $|Y| \in L^1(Q_i)$; as in [36], we call the measures in $\mathcal{Q}_0$ *valuation measures*. Define the sets

$$\mathcal{B} := \{ X \in L^0 \mid E_{Q_i}[X] \text{ exists and is } \geq 0,\ i = 1, \dots, n \} \tag{4.60}$$

and

$$\mathcal{B}_0 := \{ X \in \mathcal{B} \mid E_{Q_i}[X] = 0 \text{ for } i = 1, \dots, n \}.$$

Note that

$$\mathcal{B}_0 \cap L^0_+ = \{0\}, \tag{4.61}$$

since $X = 0$ P-a.s. as soon as $X \geq 0$ P-a.s. and $E_{Q_i}[X] = 0$, due to the equivalence $Q_i \approx P$.

As the initial acceptance set, we take the convex cone

$$\mathcal{A} := \mathcal{B} \cap L^\infty \tag{4.62}$$

The corresponding set $\bar{\mathcal{A}}$ of positions which become acceptable if combined with a suitable hedge is defined as in (4.58):

$$\bar{\mathcal{A}} := \{ X \in L^\infty \mid \exists\, \xi \in \mathbb{R}^d \text{ with } X + \xi \cdot Y \in \mathcal{B} \}.$$

Let us now introduce the following stronger version of the no-arbitrage condition $\mathcal{K} \cap L^0_+ = \{0\}$, where $\mathcal{K} := \{ \xi \cdot Y \mid \xi \in \mathbb{R}^d \}$:

$$\mathcal{K} \cap \mathcal{B} = \mathcal{K} \cap \mathcal{B}_0. \tag{4.63}$$

In other words, there is no portfolio $\xi \in \mathbb{R}^d$ such that the result satisfies the valuation inequalities in (4.60) and is strictly favorable in the sense that at least one of the inequalities is strict.

Note that (4.63) implies the absence of arbitrage opportunities:

$$\mathcal{K} \cap L^0_+ = \mathcal{K} \cap \mathcal{B} \cap L^0_+ = \mathcal{K} \cap \mathcal{B}_0 \cap L^0_+ = \{0\},$$

where we have used (4.61) and $\mathcal{B} \cap L^0_+ = L^0_+$. Thus, (4.63) implies, in particular, the existence of an equivalent martingale measure, i.e., $\mathcal{P} \neq \emptyset$. The following proposition may be viewed as an extension of the "fundamental theorem of asset pricing". Let us denote by

$$\mathcal{R} := \Big\{ \sum_{i=1}^n \lambda_i Q_i \ \Big|\ \lambda_i > 0,\ \sum_{i=1}^n \lambda_i = 1 \Big\}$$

the class of all "representative" models for the class $\mathcal{Q}_0$, i.e., all mixtures such that each $Q \in \mathcal{Q}_0$ appears with a positive weight.

Proposition 4.99. *The following two properties are equivalent:*

(a) $\mathcal{K} \cap \mathcal{B} = \mathcal{K} \cap \mathcal{B}_0$.

(b) $\mathcal{P} \cap \mathcal{R} \neq \emptyset$.

Proof. (b)$\Rightarrow$(a): For $V \in \mathcal{K} \cap \mathcal{B}$ and $R \in \mathcal{R}$, we have $E_R[\,V\,] \geq 0$. If we can choose $R \in \mathcal{P} \cap \mathcal{R}$ then we get $E_R[\,V\,] = 0$, hence $V \in \mathcal{B}_0$.

(a)$\Rightarrow$(b): Consider the convex set

$$\mathcal{C} := \{\, E_R[\,Y\,] \mid R \in \mathcal{R} \,\} \subset \mathbb{R}^d\,;$$

we have to show that $\mathcal{C}$ contains the origin. If this is not the case then there exists $\xi \in \mathbb{R}^d$ such that

$$\xi \cdot x \geq 0 \quad \text{for } x \in \mathcal{C}, \tag{4.64}$$

and

$$\xi \cdot x^* > 0 \quad \text{for some } x^* \in \mathcal{C}\,;$$

see Proposition A.1. Define $V := \xi \cdot Y \in \mathcal{K}$. Condition (4.64) implies

$$E_R[\,V\,] \geq 0 \quad \text{for all } R \in \mathcal{R},$$

hence $V \in \mathcal{K} \cap \mathcal{B}$. Let $R^* \in \mathcal{R}$ be such that $x^* = E_{R^*}[\,Y\,]$. Then V satisfies $E_{R^*}[\,Y\,] > 0$, hence $V \notin \mathcal{K} \cap \mathcal{B}_0$, in contradiction to our assumption (a). □

We can now state a representation theorem for the coherent risk measure $\overline{\rho}$ corresponding to the convex cone $\bar{\mathcal{A}}$. It is a special case of Theorem 4.102 which will be proved below.

Theorem 4.100. *Under assumption* (4.63)*, the coherent risk measure* $\overline{\rho} := \rho_{\bar{\mathcal{A}}}$ *corresponding to the acceptance set* $\bar{\mathcal{A}}$ *is given by*

$$\overline{\rho}(X) = \sup_{P^* \in \mathcal{P} \cap \mathcal{R}} E^*[\,-X\,].$$

Let us now introduce a second finite set $\mathcal{Q}_1 \subset \mathcal{M}_1(P)$ of probability measures $Q \ll P$ with $|Y| \in \mathcal{L}^1(Q)$; as in [36], we call them *stress test measures*. In addition to the valuation inequalities in (4.60), we require that an admissible position passes a stress test specified by a "floor"

$$\gamma(Q) < 0 \quad \text{for each } Q \in \mathcal{Q}_1.$$

Thus, the convex cone $\mathcal{A}$ in (4.62) is reduced to the convex set

$$\mathcal{A}_1 := \mathcal{A} \cap \mathcal{B}_1 = L^\infty \cap \big(\mathcal{B} \cap \mathcal{B}_1\big),$$

where

$$\mathcal{B}_1 := \big\{\, X \in L^0 \mid E_Q[\,X\,] \geq \gamma(Q) \text{ for } Q \in \mathcal{Q}_1 \,\big\}.$$

Let

$$\bar{\mathcal{A}}_1 := \big\{\, X \in L^\infty \mid \exists\, \xi \in \mathbb{R}^d \text{ with } X + \xi \cdot Y \in \mathcal{B} \cap \mathcal{B}_1 \,\big\}$$

denote the resulting acceptance set for positions combined with a suitable hedge.

Remark 4.101. The analogue

$$\mathcal{K} \cap (\mathcal{B} \cap \mathcal{B}_1) = \mathcal{K} \cap \mathcal{B}_0 \tag{4.65}$$

of our condition (4.63) looks weaker, but it is in fact equivalent to (4.63). Indeed, for $X \in \mathcal{K} \cap \mathcal{B}$ we can find $\varepsilon > 0$ such that $X_1 := \varepsilon X$ satisfies the additional constraints

$$E_Q[\, X_1\,] \geq \gamma(Q) \quad \text{for } Q \in \mathcal{Q}_1.$$

Since $X_1 \in \mathcal{K} \cap \mathcal{B} \cap \mathcal{B}_1$, condition (4.65) implies $X_1 \in \mathcal{K} \cap \mathcal{B}_0$, hence $X = \frac{1}{\varepsilon} X_1 \in \mathcal{K} \cap \mathcal{B}_0$, since $\mathcal{K} \cap \mathcal{B}_0$ is a cone. ◇

Let us now identify the convex measure of risk $\overline{\rho}_1$ induced by the convex acceptance set $\bar{\mathcal{A}}_1$. Define

$$\mathcal{R}_1 := \Big\{ \sum_{Q \in \mathcal{Q}} \lambda(Q) \cdot Q \;\Big|\; \lambda(Q) \geq 0,\ \sum_{Q \in \mathcal{Q}} \lambda(Q) = 1 \Big\} \supset \mathcal{R}$$

as the convex hull of $\mathcal{Q} := \mathcal{Q}_0 \cup \mathcal{Q}_1$, and define

$$\gamma(R) := \sum_{Q \in \mathcal{Q}} \lambda(Q) \gamma(Q)$$

for $R = \sum_Q \lambda(Q) Q \in \mathcal{R}$ with $\gamma(Q) := 0$ for $Q \in \mathcal{Q}_0$.

Theorem 4.102. *Under assumption* (4.63)*, the convex risk measure* $\overline{\rho}_1$ *induced by the acceptance set* $\bar{\mathcal{A}}_1$ *is given by*

$$\overline{\rho}_1(X) = \sup_{P^* \in \mathcal{P} \cap \mathcal{R}_1} \big(E^*[\,-X\,] + \gamma(P^*) \big), \tag{4.66}$$

i.e., $\overline{\rho}_1$ *is determined by the penalty function*

$$\overline{\alpha}_1(Q) := \begin{cases} +\infty & \textit{for } Q \notin \mathcal{P} \cap \mathcal{R}_1, \\ -\gamma(Q) & \textit{for } Q \in \mathcal{P} \cap \mathcal{R}_1. \end{cases}$$

Proof. Let ρ^* denote the convex risk measure defined by the right-hand side of (4.66), and let $\mathcal{A}^*$ denote the corresponding acceptance set:

$$\mathcal{A}^* := \big\{ X \in L^\infty \mid E^*[\, X\,] \geq \gamma(P^*) \text{ for all } P^* \in \mathcal{P} \cap \mathcal{R}_1 \big\}.$$

It is enough to show $\mathcal{A}^* = \bar{\mathcal{A}}_1$.

(a): In order to show $\bar{\mathcal{A}}_1 \subset \mathcal{A}^*$, take $X \in \bar{\mathcal{A}}_1$ and $P^* \in \mathcal{P} \cap \mathcal{R}_1$. There exists $\xi \in \mathbb{R}^d$ and $A_1 \in \mathcal{A}_1$ such that $X + \xi \cdot Y \geq A_1$. Thus,

$$E^*[\, X + \xi \cdot Y\,] \geq E^*[\, A_1\,] \geq \gamma(P^*),$$

due to $P^* \in \mathcal{R}_1$. Since $E^*[\,\xi \cdot Y\,] = 0$ due to $P^* \in \mathcal{P}$, we obtain $E^*[\,X\,] \geq \gamma(P^*)$, hence $X \in \mathcal{A}^*$.

(b): In order to show $\mathcal{A}^* \subset \bar{\mathcal{A}}_1$, we take $X \in \mathcal{A}^*$ and assume that $X \notin \bar{\mathcal{A}}_1$. This means that the vector $x^* = (x_1^*, \dots, x_N^*)$ with components

$$x_i^* := E_{Q_i}[\,X\,] - \gamma(Q_i)$$

does *not* belong to the convex cone

$$\mathcal{C} := \left\{ (E_{Q_i}[\,\xi \cdot Y\,])_{i=1,\dots,N} + y \mid \xi \in \mathbb{R}^d,\ y \in \mathbb{R}_+^N \right\} \subset \mathbb{R}^N,$$

where $\mathcal{Q} = \mathcal{Q}_0 \cup \mathcal{Q}_1 = \{Q_1, \dots, Q_N\}$ with $N \geq n$. In part (c) of this proof we will show that $\mathcal{C}$ is closed. Thus, there exists $\lambda \in \mathbb{R}^N$ such that

$$\lambda \cdot x^* < \inf_{x \in \mathcal{C}} \lambda \cdot x\,; \tag{4.67}$$

see Proposition A.1. Since $\mathcal{C} \supset \mathbb{R}_+^N$, we obtain $\lambda_i \geq 0$ for $i = 1, \dots, N$, and we may assume $\sum_i \lambda_i = 1$ since $\lambda \neq 0$. Define

$$R := \sum_{i=1}^N \lambda_i Q_i \in \mathcal{R}_1.$$

Since $\mathcal{C}$ contains the linear space of vectors $\big(E_{Q_i}[\,V\,]\big)_{i=1,\dots,N}$ with $V \in \mathcal{K}$, (4.67) implies

$$E_R[\,V\,] = 0 \quad \text{for } V \in \mathcal{K},$$

hence $R \in \mathcal{P}$. Moreover, the right-hand side of (4.67) must be zero, and the condition $\lambda \cdot x^* < 0$ translates into

$$E_R[\,X\,] < \gamma(R),$$

contradicting our assumption $X \in \mathcal{A}^*$.

(c): It remains to show that $\mathcal{C}$ is *closed*. For $\xi \in \mathbb{R}^d$ we define $y(\xi)$ as the vector in $\mathbb{R}^N$ with coordinates $y_i(\xi) = E_{Q_i}[\,\xi \cdot Y\,]$. Any $x \in \mathcal{C}$ admits a representation

$$x = y(\xi) + z$$

with $z \in \mathbb{R}_+^N$ and $\xi \in N^\perp$, where

$$N := \left\{ \eta \in \mathbb{R}^d \mid E_{Q_i}[\,\eta \cdot Y\,] = 0 \text{ for } i = 1, \dots, N \right\},$$

and

$$N^\perp := \left\{ \xi \in \mathbb{R}^d \mid \xi \cdot \eta = 0 \text{ for all } \eta \in N \right\}.$$

Take a sequence

$$x_n = y(\xi_n) + z_n, \quad n = 1, 2, \dots,$$

with $\xi_n \in N^{\perp}$ and $z_n \in \mathbb{R}^N_+$, such that x_n converges to $x \in \mathbb{R}^N$. If $\liminf_n |\xi_n| < \infty$, then we may assume, passing to a subsequence if necessary, that ξ_n converges to $\xi \in \mathbb{R}^d$. In this case, z_n must converge to some $z \in \mathbb{R}^N_+$, and we have $x = y(\xi)+z \in \mathcal{C}$. Let us now show that the case $\lim_n |\xi_n| = \infty$ is in fact excluded. In that case, $\alpha_n := (1 + |\xi_n|)^{-1}$ converges to 0, and the vectors $\zeta_n := \alpha_n \xi_n$ stay bounded. Thus, we may assume that ζ_n converges to $\zeta \in N^{\perp}$. This implies

$$y(\zeta) = \lim_{n\uparrow\infty} y(\zeta_n) = -\lim_{n\uparrow\infty} \alpha_n z_n \in -\mathbb{R}^N_+.$$

Since $\zeta \in N^{\perp}$ and $|\zeta| = \lim_n |\zeta_n| = 1$, we obtain $y(\zeta) \neq 0$. Thus, the inequality

$$E_{Q_i}[\,(-\zeta)\cdot Y\,] = -y_i(\zeta) \geq 0$$

holds for all i and is strict for some i, in contradiction to our assumption (4.63). □

4.9 Shortfall risk

In this section, we will establish a connection between convex measures of risk and the expected utility theory of Chapter 2.

Suppose that a risk-averse investor assesses the downside risk of a financial position $X \in \mathcal{X}$ by taking the expected utility $E[\,u(-X^-)\,]$ derived from the *shortfall* X^-, or by considering the expected utility $E[\,u(X)\,]$ of the position itself. If the focus is on the downside risk, then it is natural to change the sign and to replace u by the function $\ell(x) := -u(-x)$. Then ℓ is a strictly convex and increasing function, and the maximization of expected utility is equivalent to minimizing the expected loss $E[\,\ell(-X)\,]$ or the *shortfall risk* $E[\,\ell(X^-)\,]$. In order to unify the discussion of both cases, we do not insist on *strict* convexity. In particular, ℓ may vanish on $(-\infty, 0]$, and in this case the shortfall risk takes the form

$$E[\,\ell(X^-)\,] = E[\,\ell(-X)\,].$$

Definition 4.103. A function $\ell : \mathbb{R} \to \mathbb{R}$ is called a *loss function* if it is increasing and not identically constant.

In this section, we will only consider *convex* loss functions. Let us return to the setting where we consider risk measures defined on the class $\mathcal{X}$ of all bounded measurable functions on some given measurable space $(\Omega, \mathcal{F})$. First, we fix a probability measure P on $(\Omega, \mathcal{F})$. For a given convex loss function ℓ and an interior point x_0 in the range of ℓ, we define the following acceptance set:

$$\mathcal{A} := \{\, X \in \mathcal{X} \mid E[\,\ell(-X)\,] \leq x_0 \,\}. \tag{4.68}$$

Proposition 4.104. *The acceptance set $\mathcal{A}$ defines a convex measure of risk $\rho := \rho_{\mathcal{A}}$ which is continuous from below. Moreover, the minimal penalty function $\alpha_{\min}$ for ρ is concentrated on $\mathcal{M}_1(P)$, and ρ can be represented in the form*

$$\rho(X) = \max_{Q\in\mathcal{M}_1(P)} \big(E_Q[-X] - \alpha_{\min}(Q)\big). \tag{4.69}$$

Proof. Clearly, the convex set $\mathcal{A}$ satisfies the first two properties of Proposition 4.6 (a), and so ρ is a convex measure of risk. We have to show that ρ is continuous from below. Note first that

$$\rho(X) = \inf\big\{ m \in \mathbb{R} \mid E[\,\ell(-m-X)] \le x_0 \big\}$$

is the unique solution to the equation

$$E[\,\ell(-z-X)\,] = x_0. \tag{4.70}$$

Indeed, that $\rho(X)$ solves (4.70) follows by dominated convergence, since the finite convex function ℓ is continuous. The solution is unique, since ℓ is strictly increasing on $(\ell^{-1}(x_0)-\varepsilon, \infty)$ for some $\varepsilon > 0$.

Suppose now that (X_n) is a sequence in $\mathcal{X}$ which increases pointwise to some $X \in \mathcal{X}$. Then $\rho(X_n)$ decreases to some finite limit R. Using the continuity of ℓ and dominated convergence, it follows that

$$E[\,\ell(-\rho(X_n)-X_n)] \longrightarrow E[\,\ell(-R-X)\,].$$

But each of the approximating expectations equals x_0, and so R is a solution to (4.70). Hence $R = \rho(X)$, and this proves continuity from below. Since ρ satisfies (4.28), the representation (4.69) follows from Proposition 4.21 and Lemma 4.30. □

Let us now compute the minimal penalty function $\alpha_{\min}$.

Example 4.105. For an exponential loss function $\ell(x) = e^{\beta x}$, the minimal penalty function can be described in terms of relative entropy, and the resulting risk measure coincides, up to an additive constant, with the entropic risk measure introduced in Example 4.33. In fact,

$$\rho(X) = \inf\big\{ m \in \mathbb{R} \mid E[\,e^{-\beta(m+X)}\,] \le x_0 \big\} = \frac{1}{\beta}\big(\log E[\,e^{-\beta X}\,] - \log x_0\big).$$

In this special case, the general formula (4.17) for $\alpha_{\min}$ reduces to the variational formula for the relative entropy $H(Q|P)$ of Q with respect to P:

$$\begin{aligned}\alpha_{\min}(Q) &= \sup_{X\in\mathcal{X}} \Big(E_Q[-X\,] - \frac{1}{\beta}\log E[\,e^{-\beta X}\,]\Big) - \frac{\log x_0}{\beta}\\ &= \frac{1}{\beta}\big(H(Q|P) - \log x_0\big)\,;\end{aligned}$$

see Lemma 3.29. Thus, the representation (4.69) of ρ is equivalent to the following dual variational identity:

$$\log E[\, e^X \,] = \max_{Q\in\mathcal{M}_1(P)} \big(E_Q[\, X \,] - H(Q|P)\big). \qquad \diamond$$

In general, the minimal penalty function $\alpha_{\min}$ on $\mathcal{M}_1(P)$ can be expressed in terms of the *Fenchel–Legendre transform* or *conjugate function* ℓ^* of the convex function ℓ defined by

$$\ell^*(z) := \sup_{x\in\mathbb{R}} \big(zx - \ell(x) \big).$$

Theorem 4.106. *For any convex loss function* ℓ*, the minimal penalty function in the representation* (4.69) *is given by*

$$\alpha_{\min}(Q) = \inf_{\lambda>0} \frac{1}{\lambda}\left(x_0 + E\left[\ell^*\left(\lambda \frac{dQ}{dP} \right) \right] \right), \quad Q \in \mathcal{M}_1(P). \tag{4.71}$$

To prepare the proof of Theorem 4.106, we summarize some properties of the functions ℓ and ℓ^* as stated in Appendix A.1. First note that ℓ^* is a proper convex function, i.e., it is convex and takes some finite value. We denote by $J := (\ell^*)'_+$ its right-continuous derivative. Then, for $x, z \in \mathbb{R}$,

$$xz \le \ell(x) + \ell^*(z) \quad \text{with equality if } x = J(z). \tag{4.72}$$

Lemma 4.107. *Let* (ℓ_n) *be a sequence of convex loss functions which decreases pointwise to the convex loss function* ℓ*. Then the corresponding conjugate functions* ℓ_n^* *increase pointwise to* ℓ^**.*

Proof. It follows immediately from the definition of the Fenchel–Legendre transform that each ℓ_n^* is dominated by ℓ^*, and that $\ell_n^*(z)$ increases to some limit $\ell_\infty^*(z)$. We have to prove that $\ell_\infty^* = \ell^*$.

The function $z \mapsto \ell_\infty^*(z)$ is a lower semicontinuous convex function as the increasing limit of such functions. Moreover, ℓ_∞^* is a proper convex function, since it is dominated by the proper convex function ℓ^*. Consider the conjugate function ℓ_∞^{**} of ℓ_∞^*. Clearly, $\ell_\infty^{**} \ge \ell$, since $\ell_\infty^* \le \ell^*$ and since $\ell^{**} = \ell$ by Proposition A.6. On the other hand, we have by a similar argument that $\ell_\infty^{**} \le \ell_n$ for each n. By taking $n \uparrow \infty$, this shows $\ell_\infty^{**} = \ell$, which in turn gives $\ell_\infty^* = \ell^*$. □

Lemma 4.108. *The functions* ℓ *and* ℓ^* *have the following properties.*

(a) $\ell^*(0) = -\inf_{x\in\mathbb{R}} \ell(x)$ *and* $\ell^*(z) \ge -\ell(0)$ *for all* z*.*

(b) *There exists some* $z_1 \in [0, \infty)$ *such that*

$$\ell^*(z) = \sup_{x\ge 0} \big(xz - \ell(x)\big) \quad \textit{for } z \ge z_1.$$

In particular, ℓ^* *is increasing on* $[z_1, \infty)$*.*

(c) $\dfrac{\ell^*(z)}{z} \longrightarrow \infty$ *as* $z \uparrow \infty$.

Proof. Part (a) is obvious.

(b): Let $N := \{ z \in \mathbb{R} \mid \ell^*(z) = -\ell(0) \}$. We show in a first step that $N \neq \emptyset$. Note that convexity of ℓ implies that the set S of all z with $zx \leq \ell(x) - \ell(0)$ for all $x \in \mathbb{R}$ is non-empty. For $z \in S$ we clearly have $\ell^*(z) \leq -\ell(0)$. On the other hand, $\ell^*(z) \geq -\ell(0)$ by (a).

Now we take $z_1 := \sup N$. It is clear that $z_1 \geq 0$. If $z > z_1$ and $x < 0$, then

$$xz - \ell(x) \leq xz_1 - \ell(x) \leq \ell^*(z_1) \leq -\ell(0),$$

where the last inequality follows from the lower semicontinuity of ℓ^*. But $\ell^*(z) > -\ell(0)$, hence

$$\sup_{x<0} \big(xz - \ell(x) \big) < \ell^*(z).$$

(c): For $z \geq z_1$,

$$\ell^*(z)/z = \sup_{x \geq 0} \big(x - \ell(x)/z \big)$$

by (b). Hence

$$\frac{\ell^*(z)}{z} \geq x_z - 1,$$

where $x_z := \sup\{ x \mid \ell(x) \leq z \}$. Since ℓ is convex, increasing, and takes only finite values, we have $x_z \to \infty$ as $z \uparrow \infty$. □

Proof of Theorem 4.106. Fix $Q \in \mathcal{M}_1(P)$, and denote by $\varphi := dQ/dP$ its density. First, we show that it suffices to prove the claim for $x_0 > \ell(0)$. Otherwise we can find some $a \in \mathbb{R}$ such that $\ell(-a) < x_0$, since x_0 was assumed to be an interior point of $\ell(\mathbb{R})$. Let $\widetilde{\ell}(x) := \ell(x - a)$, and

$$\widetilde{\mathcal{A}} := \big\{ \widetilde{X} \in \mathcal{X} \mid E[\, \widetilde{\ell}(-\widetilde{X}) \,] \leq x_0 \big\}.$$

Then $\widetilde{\mathcal{A}} = \{ X - a \mid X \in \mathcal{A} \}$, and hence

$$\sup_{\widetilde{X} \in \widetilde{\mathcal{A}}} E_Q[\, -\widetilde{X} \,] = \sup_{X \in \mathcal{A}} E_Q[\, -X \,] + a. \tag{4.73}$$

The convex loss function $\widetilde{\ell}$ satisfies the requirement $\widetilde{\ell}(0) < x_0$. So if the assertion is established in this case, we find that

$$\sup_{\widetilde{X} \in \widetilde{\mathcal{A}}} E_Q[\, -\widetilde{X} \,] = \inf_{\lambda > 0} \frac{1}{\lambda} (x_0 + E[\, \widetilde{\ell}^*(\lambda\varphi) \,]) = \inf_{\lambda > 0} \frac{1}{\lambda} (x_0 + E[\, \ell^*(\lambda\varphi) \,]) + a\,;$$

here we have used the fact that the Fenchel–Legendre transform $\widetilde{\ell}^*$ of $\widetilde{\ell}$ satisfies $\widetilde{\ell}^*(z) = \ell^*(z) + az$. Together with (4.73), this proves that the reduction to the case $\ell(0) < x_0$ is indeed justified.

For any $\lambda > 0$ and $X \in \mathcal{A}$, (4.72) implies

$$-X\varphi = \frac{1}{\lambda}(-X)(\lambda\varphi) \leq \frac{1}{\lambda}\big(\ell(-X) + \ell^*(\lambda\varphi)\big).$$

Hence, for any $\lambda > 0$

$$\alpha_{\min}(Q) \leq \sup_{X \in \mathcal{A}} \frac{1}{\lambda}\big(E[\,\ell(-X)\,] + E[\,\ell^*(\lambda\varphi)]\big) \leq \frac{1}{\lambda}\big(x_0 + E[\,\ell^*(\lambda\varphi)]\big)$$

Thus, it remains to prove that

$$\alpha_{\min}(Q) \geq \inf_{\lambda > 0} \frac{1}{\lambda}\big(x_0 + E[\,\ell^*(\lambda\varphi)]\big) \tag{4.74}$$

in case where $\alpha_{\min}(Q) < \infty$. This will be done first under the following extra conditions:

$$\text{There exists } \kappa \in \mathbb{R} \text{ such that } \ell(x) = \inf \ell \text{ for all } x \leq \kappa. \tag{4.75}$$

$$\ell^* \text{ is finite on } (0, \infty). \tag{4.76}$$

$$J \text{ is continuous on } (0, \infty). \tag{4.77}$$

Note that these assumptions imply that $\ell^*(0) < \infty$ and that $J(0+) \geq \kappa$. Moreover, $J(z)$ increases to $+\infty$ as $z \uparrow \infty$, and hence so does $\ell(J(z))$. Since

$$\ell^*(z) \geq -\ell(0) > -x_0 \quad \text{for all } z, \tag{4.78}$$

it follows from (4.72) that

$$\lim_{z \downarrow 0} \ell(J(z)) - x_0 < \lim_{z \downarrow 0} \big(\ell(J(z)) + \ell^*(z)\big) = \lim_{z \downarrow 0} zJ(z) = 0.$$

These facts and the continuity of J imply that for large enough n there exists some $\lambda_n > 0$ such that

$$E\big[\,\ell\big(J(\lambda_n\varphi)\mathrm{I}_{\{\varphi \leq n\}}\big)\,\big] = x_0.$$

Let us define

$$X^n := -J(\lambda_n\varphi)\mathrm{I}_{\{\varphi \leq n\}}.$$

Then X^n is bounded and belongs to $\mathcal{A}$. Hence, it follows from (4.72) and (4.78) that

$$\begin{aligned}
\alpha_{\min}(Q) &\geq E_Q[\,-X^n\,] \\
&= \frac{1}{\lambda_n} E\big[\,\mathrm{I}_{\{\varphi \leq n\}} J(\lambda_n\varphi)(\lambda_n\varphi)\,\big] \\
&= \frac{1}{\lambda_n} E\big[\,\big(\ell(-X^n) + \ell^*(\lambda_n\varphi)\big) \cdot \mathrm{I}_{\{\varphi \leq n\}}\,\big] \\
&= \frac{1}{\lambda_n}\big(x_0 - \ell(0) \cdot P[\,\varphi > n\,] + E\big[\,\ell^*(\lambda_n\varphi)\mathrm{I}_{\{\varphi \leq n\}}\,\big]\big) \\
&\geq \frac{x_0 - \ell(0)}{\lambda_n}.
\end{aligned}$$

Since we assumed that $\alpha_{\min}(Q) < \infty$, the decreasing limit λ_∞ of λ_n must be strictly positive. The fact that ℓ^* is bounded from below allows us to apply Fatou's lemma:

$$\begin{aligned}\alpha_{\min}(Q) &\geq \liminf_{n\uparrow\infty} \frac{1}{\lambda_n}\big(x_0 - \ell(0)\cdot P[\,\varphi > n\,] + E\big[\,\ell^*(\lambda_n\varphi)\mathrm{I}_{\{\varphi\leq n\}}\,\big]\big)\\ &\geq \frac{1}{\lambda_\infty}\big(x_0 + E[\,\ell^*(\lambda_\infty\varphi)\,]\big).\end{aligned}$$

This proves (4.74) under the assumptions (4.75), (4.76), and (4.77).

If (4.75) and (4.76) hold, but J is not continuous, then we can approximate the upper semicontinuous function J from above with an increasing continuous function $\widetilde{J}$ on $[0,\infty)$ such that

$$\widetilde{\ell}^*(z) := \ell^*(0) + \int_0^z \widetilde{J}(y)\,dy$$

satisfies

$$\ell^*(z) \leq \widetilde{\ell}^*(z) \leq \ell^*\big((1+\varepsilon)z\big) \quad \text{for } z \geq 0.$$

Let $\widetilde{\ell} := \widetilde{\ell}^{**}$ denote the Fenchel–Legendre transform of $\widetilde{\ell}^*$. Since $\ell^{**} = \ell$ by Proposition A.6, it follows that

$$\ell\Big(\frac{x}{1+\varepsilon}\Big) \leq \widetilde{\ell}(x) \leq \ell(x).$$

Therefore,

$$\widetilde{\mathcal{A}} := \Big\{\, X \in \mathcal{X} \mid E[\,\widetilde{\ell}(-X)\,] \leq x_0 \Big\} \subseteq \Big\{\,(1+\varepsilon)X \mid X \in \mathcal{A}\Big\} =: \mathcal{A}_\varepsilon.$$

Since we already know that the assertion holds for $\widetilde{\ell}$, we get that

$$\begin{aligned}\inf_{\lambda>0}\frac{1}{\lambda}\bigg(x_0 + E\bigg[\,\ell^*\Big(\lambda\frac{dQ}{dP}\Big)\bigg]\bigg) &\leq \inf_{\lambda>0}\frac{1}{\lambda}\bigg(x_0 + E\bigg[\,\widetilde{\ell}^*\Big(\lambda\frac{dQ}{dP}\Big)\bigg]\bigg)\\ &= \sup_{X\in\widetilde{\mathcal{A}}} E_Q[\,-X\,]\\ &\leq \sup_{X\in\mathcal{A}_\varepsilon} E_Q[\,-X\,]\\ &= (1+\varepsilon)\alpha_{\min}(Q).\end{aligned}$$

By letting $\varepsilon \downarrow 0$, we obtain (4.74).

Finally, we remove conditions (4.75) and (4.76). If $\ell^*(z) = +\infty$ for some z, then z must be an upper bound for the slope of ℓ. So we will approximate ℓ by a sequence (ℓ_n) of convex loss functions whose slope is unbounded. Simultaneously, we can handle the case where ℓ does not take on its infimum. To this end, we choose a sequence $\kappa_n \downarrow \inf \ell$ such that $\kappa_n \leq \ell(0) < x_0$. We can define, for instance,

$$\ell_n(x) := \ell(x) \vee \kappa_n + \frac{1}{n}(e^x - 1)^+.$$

Then ℓ_n decreases pointwise to ℓ. Each loss function ℓ_n satisfies (4.75) and (4.76). Hence, for any $\varepsilon > 0$ there are λ_n^ε such that

$$\infty > \alpha_{\min}(Q) \geq \alpha_{\min}^n(Q) \geq \frac{1}{\lambda_n^\varepsilon}\big(x_0 + E[\,\ell_n^*(\lambda_n^\varepsilon\varphi)\,]\big) - \varepsilon \quad \text{for each } n,$$

where $\alpha_{\min}^n(Q)$ is the penalty function arising from ℓ_n. Note that $\ell_n^* \nearrow \ell^*$ by Lemma 4.107. Our assumption $\alpha_{\min}(Q) < \infty$, the fact that

$$\inf_{z\in\mathbb{R}} \ell_n^*(z) \geq -\ell_n(0) = -\ell(0) > -x_0,$$

and part (c) of Lemma 4.108 show that the sequence $(\lambda_n^\varepsilon)_{n\in\mathbb{N}}$ must be bounded away from zero and from infinity. Therefore, we may assume that λ_n^ε converges to some $\lambda^\varepsilon \in (0,\infty)$. Using again the fact that $\ell_n^*(z) \geq -\ell(0)$ uniformly in n and z, Fatou's lemma yields

$$\alpha_{\min}(Q) + \varepsilon \geq \liminf_{n\uparrow\infty} \frac{1}{\lambda_n^\varepsilon}\big(x_0 + E[\,\ell_n^*(\lambda_n^\varepsilon\varphi)\,]\big) \geq \frac{1}{\lambda^\varepsilon}\big(x_0 + E[\,\ell^*(\lambda^\varepsilon\varphi)\,]\big).$$

This completes the proof of the theorem. □

Example 4.109. Take

$$\ell(x) := \begin{cases} \frac{1}{p}x^p & \text{if } x \geq 0,\\ 0 & \text{otherwise,}\end{cases}$$

where $p > 1$. Then

$$\ell^*(z) := \begin{cases} \frac{1}{q}z^q & \text{if } z \geq 0,\\ +\infty & \text{otherwise,}\end{cases}$$

where $q = p/(p-1)$ is the usual dual coefficient. We may apply Theorem 4.106 for any $x_0 > 0$. Let $Q \in \mathcal{M}_1(P)$ with density $\varphi := dQ/dP$. Clearly, $\alpha_{\min}(Q) = +\infty$ if $\varphi \notin L^q(\Omega, \mathcal{F}, P)$. Otherwise, the infimum in (4.71) is attained for

$$\lambda_Q = \left(\frac{px_0}{E[\,\varphi^q\,]}\right)^{1/q}.$$

Hence, we can identify $\alpha_{\min}(Q)$ for any $Q \ll P$ as

$$\alpha_{\min}^p(Q) = (px_0)^{1/p} \cdot E\bigg[\Big(\frac{dQ}{dP}\Big)^q\bigg]^{1/q}.$$

Taking the limit $p \downarrow 1$, we obtain the case $\ell(x) = x^+$ where we measure the risk in terms of the *expected shortfall*. Here we have

$$\alpha_{\min}^1(Q) = x_0 \cdot \left\|\frac{dQ}{dP}\right\|_\infty. \qquad \diamond$$

Together with Proposition 4.19, Theorem 4.106 yields the following result for risk measures which are defined in terms of a robust notion of bounded shortfall risk. Here it is convenient to define $\ell^*(\infty) := \infty$.

Corollary 4.110. *Suppose that $\mathcal{Q}$ is a family of probability measures on $(\Omega, \mathcal{F})$, and that ℓ, ℓ^*, and x_0 are as in Theorem 4.106. We define a set of acceptable positions by*

$$\mathcal{A} := \big\{ X \in \mathcal{X} \mid E_P[\,\ell(-X)\,] \le x_0 \text{ for all } P \in \mathcal{Q} \big\}.$$

Then the corresponding convex risk measure can be represented in terms of the penalty function

$$\alpha(Q) = \inf_{\lambda>0} \frac{1}{\lambda}\left(x_0 + \inf_{P\in\mathcal{Q}} E_P\left[\ell^*\left(\lambda \frac{dQ}{dP} \right) \right] \right), \quad Q \in \mathcal{M}_1(\Omega, \mathcal{F}),$$

where dQ/dP is the density appearing in the Lebesgue decomposition of Q with respect to P as in Theorem A.13.

Example 4.111. In the case of Example 4.105, the corresponding robust problem in Corollary 4.110 leads to the following entropy minimization problem: For a given Q and a set $\mathcal{Q}$ of probability measures, find

$$\inf_{P\in\mathcal{Q}} H(Q|P).$$

Note that this problem is different from the standard problem of minimizing $H(Q|P)$ with respect to the first variable Q as it appears in Section 3.2. ◇

Example 4.112. Take $x_0 = 0$ in (4.68) and $\ell(x) := x$. Then

$$\ell^*(z) := \begin{cases} 0 & \text{if } z = 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore, $\alpha(Q) = \infty$ if $Q \neq P$, and $\rho(X) = E[\,-X\,]$. If $\mathcal{Q}$ is a set of probability measures, the "robust" risk measure ρ of Corollary 4.110 is coherent, and it is given by

$$\rho(X) = \sup_{P\in\mathcal{Q}} E_P[\,-X\,]. \qquad ◇$$

Part II

Dynamic hedging

Chapter 5
Dynamic arbitrage theory

In this chapter we develop a dynamic version of the arbitrage theory of Chapter 1. Here we will work in a multiperiod setting, where the stochastic price fluctuation of a financial asset is described as a stochastic process in discrete time. Portfolios will be successively readjusted, taking into account the information available at each time. Market efficiency requires that such dynamic trading strategies do not create arbitrage opportunities. In Section 5.2 we show that an arbitrage-free model is characterized by the existence of an equivalent martingale measure. Under such a measure, the discounted price processes of the traded assets are martingales, that is, they have the mathematical structure of a fair game. In Section 5.3 we introduce European contingent claims. These are financial instruments whose payoff at the expiration date depends on the behavior of the underlying primary assets, and possibly on other factors. We discuss the problem of pricing such contingent claims in a manner which does not create new arbitrage opportunities. The pricing problem is closely related to the problem of hedging a given claim by using a dynamic trading strategy based on the primary assets. An ideal situation occurs if any contingent claim can be perfectly replicated by the final outcome of such a strategy. In such a *complete* model, the equivalent martingale measure P^* is unique, and derivatives are priced in a canonical manner by taking the expectation of the discounted payoff with respect to the measure P^*. Section 5.5 contains a simple case study for completeness, the binomial model introduced by Cox, Ross, and Rubinstein. In this context, it is possible to obtain explicit pricing formulas for a number of exotic options, as explained in Section 5.6. In Section 5.7 we pass to the limiting diffusion model of geometric Brownian motion. Using a suitable version of the central limit theorem, we are led to the general Black–Scholes formula for European contingent claims and to explicit pricing formulas for some exotic options such as the up-and-in call.

The general structure of complete models is described in Section 5.4. There it will become clear that completeness is the exception rather than the rule: Typical market models in discrete time are *incomplete*.

5.1 The multi-period market model

Throughout this chapter, we consider a market model in which $d+1$ assets are priced at times $t = 0, 1, \dots, T$. The price of the i^{th} asset at time t is modelled as a non-negative random variable S_t^i on a given probability space $(\Omega, \mathcal{F}, P)$. The random vector $\overline{S}_t = (S_t^0, S_t) = (S_t^0, S_t^1, \dots, S_t^d)$ is assumed to be measurable with respect to a σ-algebra $\mathcal{F}_t \subset \mathcal{F}$. One should think of $\mathcal{F}_t$ as the class of all events which are observable up to time t. Thus, it is natural to assume that

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_T. \tag{5.1}$$

Definition 5.1. A family $(\mathcal{F}_t)_{t=0,\dots,T}$ of σ-algebras satisfying (5.1) is called a *filtration*. In this case, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t=0,\dots,T}, P)$ is also called a *filtered probability space*.

To simplify the presentation, we will assume that

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \quad \text{and} \quad \mathcal{F} = \mathcal{F}_T. \tag{5.2}$$

Let $(E, \mathcal{E})$ be a measurable space. A *stochastic process* with state space $(E, \mathcal{E})$ is given by a family of E-valued random variables on $(\Omega, \mathcal{F}, P)$ indexed by time. In our context, the typical parameter sets will be $\{0, \dots, T\}$ or $\{1, \dots, T\}$, and the state space will be some Euclidean space.

Definition 5.2. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{F}_t)_{t=0,\dots,T}$ be a filtration.

(a) A stochastic process $Y = (Y_t)_{t=0,\dots,T}$ is called *adapted* with respect to the filtration $(\mathcal{F}_t)_{t=0,\dots,T}$ if each Y_t is $\mathcal{F}_t$-measurable.

(b) A stochastic process $Z = (Z_t)_{t=1,\dots,T}$ is called *predictable* with respect to $(\mathcal{F}_t)_{t=0,\dots,T}$ if each Z_t is $\mathcal{F}_{t-1}$-measurable.

Note that in our definition predictable processes start at $t = 1$ while adapted processes are also defined at $t = 0$. In particular, the asset prices $\overline{S} = (\overline{S}_t)_{t=0,\dots,T}$ form an adapted stochastic process with values in $\mathbb{R}^{d+1}$.

Definition 5.3. A *trading strategy* is a predictable $\mathbb{R}^{d+1}$-valued process $\overline{\xi} = (\xi^0, \xi) = (\xi_t^0, \xi_t^1, \dots, \xi_t^d)_{t=1,\dots,T}$.

The value ξ_t^i of a trading strategy $\overline{\xi}$ corresponds to the quantity of shares of the i^{th} asset held during the t^{th} trading period between $t-1$ and t. Thus, $\xi_t^i S_{t-1}^i$ is the amount invested into the i^{th} asset at time $t-1$, while $\xi_t^i S_t^i$ is the resulting value at time t. The total value of the portfolio $\overline{\xi}_t$ at time $t-1$ is

$$\overline{\xi}_t \cdot \overline{S}_{t-1} = \sum_{i=0}^d \xi_t^i S_{t-1}^i.$$

By time t, the value of the portfolio $\overline{\xi}_t$ has changed to

$$\overline{\xi}_t \cdot \overline{S}_t = \sum_{i=0}^d \xi_t^i S_t^i.$$

The predictability of $\overline{\xi}$ expresses the fact that investments must be allocated at the beginning of each trading period, without anticipating future price increments.

Definition 5.4. A trading strategy $\overline{\xi}$ is called *self-financing* if

$$\overline{\xi}_t \cdot \overline{S}_t = \overline{\xi}_{t+1} \cdot \overline{S}_t \quad \text{for } t = 1, \dots, T-1. \tag{5.3}$$

Intuitively, (5.3) means that the portfolio is always rearranged in such a way that its present value is preserved. It follows that the accumulated gains and losses resulting from the asset price fluctuations are the only source of variations of the portfolio value:

$$\overline{\xi}_{t+1} \cdot \overline{S}_{t+1} - \overline{\xi}_t \cdot \overline{S}_t = \overline{\xi}_{t+1} \cdot (\overline{S}_{t+1} - \overline{S}_t). \tag{5.4}$$

In fact, a trading strategy is self-financing if and only if (5.4) holds for $t = 1, \dots, T-1$. It follows through summation over (5.4) that

$$\overline{\xi}_t \cdot \overline{S}_t = \overline{\xi}_1 \cdot \overline{S}_0 + \sum_{k=1}^{t} \overline{\xi}_k \cdot (\overline{S}_k - \overline{S}_{k-1}) \quad \text{for } t = 1, \dots, T.$$

Here, the constant $\overline{\xi}_1 \cdot \overline{S}_0$ can be interpreted as the initial investment for the purchase of the portfolio $\overline{\xi}_1$.

Example 5.5. Often it is assumed that the 0^{th} asset plays the role of a *locally riskless bond*. In this case, one takes $S_0^0 \equiv 1$ and one lets S_t^0 evolve according to a *spot rate* $r_t \geq 0$: At time t, an investment x made at time $t-1$ yields the payoff $x(1+r_t)$. Thus, a unit investment at time 0 produces the value

$$S_t^0 = \prod_{k=1}^{t} (1 + r_k)$$

at time t. An investment in S^0 is "locally riskless" if the spot rate r_t is known beforehand at time $t-1$. This idea can be made precise by assuming that the process r is *predictable*. ◇

Without assuming predictability as in the preceding example, we assume from now on that

$$S_t^0 > 0 \quad P\text{-a.s. for all } t.$$

This assumption allows us to use the 0^{th} asset as a *numéraire* and to form the *discounted* price processes

$$X_t^i := \frac{S_t^i}{S_t^0}, \quad t = 0, \dots, T, \ i = 0, \dots, d.$$

Then $X_t^0 \equiv 1$, and $X_t = (X_t^1, \dots, X_t^d)$ expresses the value of the remaining assets in units of the numéraire. As explained in Remark 1.9, discounting allows comparison of asset prices which are quoted at different times.

Definition 5.6. The (discounted) *value process* $V = (V_t)_{t=0,\dots,T}$ associated with a trading strategy $\overline{\xi}$ is given by

$$V_0 := \overline{\xi}_1 \cdot \overline{X}_0 \quad \text{and} \quad V_t := \overline{\xi}_t \cdot \overline{X}_t \quad \text{for } t = 1, \dots, T.$$

The *gains process* associated with $\overline{\xi}$ is defined as

$$G_0 := 0 \quad \text{and} \quad G_t := \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) \quad \text{for } t = 1, \dots, T.$$

Clearly,

$$V_t = \overline{\xi}_t \cdot \overline{X}_t = \frac{\overline{\xi}_t \cdot \overline{S}_t}{S_t^0},$$

so V_t can be interpreted as the portfolio value at the end of the t^{th} trading period expressed in units of the numéraire asset. The gains process

$$G_t = \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1})$$

reflects, in terms of the numéraire, the net gains which have accumulated through the trading strategy $\overline{\xi}$ up to time t. For a self-financing trading strategy $\overline{\xi}$, the identity

$$\overline{\xi}_t \cdot \overline{S}_t = \overline{\xi}_1 \cdot \overline{S}_0 + \sum_{k=1}^{t} \overline{\xi}_k \cdot (\overline{S}_k - \overline{S}_{k-1}) \tag{5.5}$$

remains true if all relevant quantities are computed in units of the numéraire. This is the content of the following simple proposition.

Proposition 5.7. *For a trading strategy $\overline{\xi}$ the following conditions are equivalent:*

(a) $\overline{\xi}$ *is self-financing.*

(b) $\overline{\xi}_t \cdot \overline{X}_t = \overline{\xi}_{t+1} \cdot \overline{X}_t$ *for* $t = 1, \dots, T-1$.

(c) $V_t = V_0 + G_t = \overline{\xi}_1 \cdot \overline{X}_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1})$ *for all* t.

Proof. By dividing both sides of (5.3) by S_t^0 it is seen that condition (b) is a reformulation of Definition 5.4. Moreover, (b) holds if and only if

$$\overline{\xi}_{t+1} \cdot \overline{X}_{t+1} - \overline{\xi}_t \cdot \overline{X}_t = \overline{\xi}_{t+1} \cdot (\overline{X}_{t+1} - \overline{X}_t) = \xi_{t+1} \cdot (X_{t+1} - X_t)$$

for $t = 1, \dots, T-1$, and this identity is equivalent to (c). □

Remark 5.8. The numéraire component of a self-financing trading strategy $\overline{\xi}$ satisfies

$$\xi^0_{t+1} - \xi^0_t = -(\xi_{t+1} - \xi_t) \cdot X_t \quad \text{for } t = 1, \dots, T-1. \tag{5.6}$$

Since

$$\xi^0_1 = V_0 - \xi_1 \cdot X_0, \tag{5.7}$$

the entire process ξ^0 is determined by the initial investment V_0 and the d-dimensional process ξ. Consequently, if a constant V_0 and an arbitrary d-dimensional predictable process ξ are given, then we can use (5.7) and (5.6) as the definition of a predictable process ξ^0, and this construction yields a self-financing trading strategy $\overline{\xi} := (\xi^0, \xi)$. In dealing with self-financing strategies $\overline{\xi}$, it is thus sufficient to focus on the initial investment V_0 and the d-dimensional processes X and ξ. ◇

Remark 5.9. Different economic agents investing into the same market may choose different numéraires. For example, consider the following simple market model in which prices are quoted in euros (€) as the domestic currency. Let S^0 be a locally riskless €-bond with the predictable spot rate process r^0, i.e.,

$$S^0_t = \prod_{k=1}^{t} (1 + r^0_k),$$

and let S^1 describe the price of a locally riskless investment into US dollars ($). Since the price of this $-bond is quoted in €, the asset S^1 is modeled as

$$S^1_t = U_t \cdot \prod_{k=1}^{t} (1 + r^1_k),$$

where r^1 is the spot rate for a $-investment, and U_t denotes the price of 1$ in terms of €, i.e., U_t is the exchange rate of the $ versus the €. While it may be natural for European investors to take S^0 as their numéraire, it may be reasonable for an American investor to choose S^1. This simple example explains why it may be relevant to check which concepts and results of our theory are invariant under a *change of numéraire*; see, e.g., the discussion at the end of Section 5.2. ◇

5.2 Arbitrage opportunities and martingale measures

Intuitively, an arbitrage opportunity is an investment strategy that yields a positive profit with positive probability but without any downside risk.

Definition 5.10. A self-financing trading strategy is called an *arbitrage opportunity* if its value process V satisfies

$$V_0 \le 0, \quad V_T \ge 0 \ \ P\text{-a.s.}, \quad \text{and} \quad P[\, V_T > 0\,] > 0.$$

The existence of such an arbitrage opportunity may be regarded as a market inefficiency in the sense that certain assets are not priced in a reasonable way. In this section, we will characterize those market models which do not allow for arbitrage opportunities. Such models will be called *arbitrage-free*. The following proposition shows that the market model is arbitrage-free if and only if there are no arbitrage opportunities for each single trading period. Later on, this fact will allow us to apply the results of Section 1.6 to our multi-period model.

Proposition 5.11. *The market model admits an arbitrage opportunity if and only if there exist $t \in \{1, \dots, T\}$ and $\eta \in \mathcal{L}^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$ such that*

$$\eta \cdot (X_t - X_{t-1}) \geq 0 \;\; P\text{-a.s., and} \;\; P[\, \eta \cdot (X_t - X_{t-1}) > 0\,] > 0. \tag{5.8}$$

Proof. To prove necessity, take an arbitrage opportunity $\overline{\xi} = (\xi^0, \xi)$ with value process V, and let

$$t := \min\big\{\, k \mid V_k \geq 0 \; P\text{-a.s., and } P[\, V_k > 0\,] > 0 \,\big\}.$$

Then $t \leq T$ by assumption, and either $V_{t-1} = 0$ P-a.s. or $P[\, V_{t-1} < 0\,] > 0$. In the first case, it follows that

$$\xi_t \cdot (X_t - X_{t-1}) = V_t - V_{t-1} = V_t \quad P\text{-a.s.}$$

Thus, $\eta := \xi_t$ satisfies (5.8). In the second case, we let $\eta := \xi_t \, \mathrm{I}_{\{V_{t-1}<0\}}$. Then η is $\mathcal{F}_{t-1}$-measurable, and

$$\eta \cdot (X_t - X_{t-1}) = (V_t - V_{t-1})\, \mathrm{I}_{\{V_{t-1}<0\}} \geq -V_{t-1}\, \mathrm{I}_{\{V_{t-1}<0\}}.$$

The expression on the right-hand side is non-negative and strictly positive with a positive probability, so (5.8) holds.

Now we prove sufficiency. For t and η as in (b), define a d-dimensional predictable process ξ by

$$\xi_s := \begin{cases} \eta & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

Via (5.7) and (5.6), ξ uniquely defines a self-financing trading strategy $\overline{\xi} = (\xi^0, \xi)$ with initial investment $V_0 = 0$. Since the corresponding value process satisfies $V_T = \eta \cdot (X_t - X_{t-1})$, the strategy $\overline{\xi}$ is an arbitrage opportunity. □

Definition 5.12. A stochastic process $M = (M_t)_{t=0,\dots,T}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), Q)$ is called a *martingale* if M is adapted, satisfies $E_Q[\,|M_t|\,] < \infty$ for all t, and if

$$M_s = E_Q[\, M_t \mid \mathcal{F}_s\,] \quad \text{for } 0 \leq s \leq t \leq T. \tag{5.9}$$

A martingale can be regarded as the mathematical formalization of a "fair game": For each time s and for each horizon $t > s$, the conditional expectation of the future gain $M_t - M_s$ is zero, given the information available at s.

Remark 5.13. Clearly, (5.9) is equivalent to the following condition:

$$M_t = E_Q[\, M_{t+1} \mid \mathcal{F}_t \,] \quad \text{for } 0 \le t \le T-1.$$

Taking $t = T$ in (5.9), we see that, in our context of a finite time horizon, a process M is a martingale if and only if it arises as a sequence of successive conditional expectations $E_Q[\, F \mid \mathcal{F}_t \,]$ for some $F \in \mathcal{L}^1(\Omega, \mathcal{F}_T, Q)$. Consider, for example, the *density process*

$$Z_t := \frac{d\widetilde{Q}}{dQ}\bigg|_{\mathcal{F}_t}, \quad t = 0, \dots, T,$$

of a probability measure $\widetilde{Q}$ that is absolutely continuous with respect to Q. It follows from Proposition A.11 that $Z_t = E_Q[\, Z_T \mid \mathcal{F}_t \,]$, so that Z is a martingale. ◇

Whether or not a given process M is a martingale depends on the underlying probability measure Q. If we wish to emphasize the dependence of the martingale property of M on a particular measure Q, we will say that M is a *Q-martingale* or that M is a *martingale under the measure Q*.

Definition 5.14. A probability measure Q on $(\Omega, \mathcal{F}_T)$ is called a *martingale measure* if the discounted price process X is a (d-dimensional) Q-martingale, i.e.,

$$E_Q[\, X_t^i \,] < \infty \quad \text{and} \quad X_s^i = E_Q[\, X_t^i \mid \mathcal{F}_s \,], \quad 0 \le s \le t \le T,\ i = 1, \dots, d.$$

A martingale measure P^* is called an *equivalent martingale measure* if it is equivalent to the original measure P on $\mathcal{F}_T$. The set of all equivalent martingale measures is denoted by $\mathcal{P}$.

The following result is a version of Doob's fundamental "systems theorem" for martingales. It states that a fair game admits no realistic gambling system which produces a positive expected gain. Here, Y^- denotes the negative part $-Y \wedge 0$ of a random variable Y.

Theorem 5.15. *For a probability measure Q, the following conditions are equivalent.*

(a) *Q is a martingale measure.*

(b) *If $\overline{\xi} = (\xi^0, \xi)$ is self-financing and ξ is bounded, then the value process V of $\overline{\xi}$ is a Q-martingale.*

(c) *If $\overline{\xi} = (\xi^0, \xi)$ is self-financing and its value process V satisfies $E_Q[\, V_T^- \,] < \infty$, then V is a Q-martingale.*

(d) *If $\overline{\xi} = (\xi^0, \xi)$ is self-financing and its value process V satisfies $V_T \ge 0$ Q-a.s., then $E_Q[\, V_T \,] = V_0$.*

Proof. (a) ⇒ (b): Let V be the value process of a self-financing trading strategy $\overline{\xi} = (\xi^0, \xi)$ such that $|\xi|$ is bounded by a constant c. Then

$$|V_t| \leq |V_0| + \sum_{k=1}^{t} c\big(|X_k| + |X_{k-1}|\big).$$

Since each $|X_k|$ belongs to $\mathcal{L}^1(Q)$, we have $E_Q[\,|V_t|\,] < \infty$. Moreover, for $0 \leq t \leq T-1$,

$$\begin{aligned} E_Q[\, V_{t+1} \mid \mathcal{F}_t \,] &= E_Q\big[\, V_t + \xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t \,\big] \\ &= V_t + \xi_{t+1} \cdot E_Q[\, X_{t+1} - X_t \mid \mathcal{F}_t \,] \\ &= V_t, \end{aligned}$$

where we have used that ξ_{t+1} is $\mathcal{F}_t$-measurable and bounded.

(b) ⇒ (c): We will show the following implication:

$$\text{If } E_Q[\, V_t^- \,] < \infty \text{ then } E_Q[\, V_t \mid \mathcal{F}_{t-1} \,] = V_{t-1}. \tag{5.10}$$

Since $E_Q[\, V_T^- \,] < \infty$ by assumption, we will then get

$$E_Q[\, V_{T-1}^- \,] = E_Q\big[\, E_Q[\, V_T \mid \mathcal{F}_{T-1} \,]^- \,\big] \leq E_Q[\, V_T^- \,] < \infty,$$

due to Jensen's inequality for conditional expectations. Repeating this argument will yield $E_Q[\, V_t^- \,] < \infty$ and $E_Q[\, V_t \mid \mathcal{F}_{t-1} \,] = V_{t-1}$ for all t. Since V_0 is a finite constant, we will also get $E_Q[\, V_t \,] = V_0$, which together with the fact that $E_Q[\, V_t^- \,] < \infty$ implies $V_t \in \mathcal{L}^1(Q)$ for all t. Thus, the martingale property of V will follow.

To prove (5.10), note first that $E_Q[\, V_t \mid \mathcal{F}_{t-1} \,]$ is well defined due to our assumption $E_Q[\, V_t^- \,] < \infty$. Next, let $\xi_t^{(a)} := \xi_t \, \mathrm{I}_{\{|\xi_t| \leq a\}}$ for $a > 0$. Then $\xi_t^{(a)} \cdot (X_t - X_{t-1})$ is a martingale increment by condition (b). In particular, $\xi_t^{(a)} \cdot (X_t - X_{t-1}) \in \mathcal{L}^1(Q)$ and $E_Q[\, \xi_t^{(a)} \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] = 0$. Hence,

$$\begin{aligned} E_Q[\, V_t \mid \mathcal{F}_{t-1} \,] \mathrm{I}_{\{|\xi_t| \leq a\}} &= E_Q[\, V_t \, \mathrm{I}_{\{|\xi_t| \leq a\}} \mid \mathcal{F}_{t-1} \,] - E_Q[\, \xi_t^{(a)} \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] \\ &= E_Q\big[\, V_t \, \mathrm{I}_{\{|\xi_t| \leq a\}} - \xi_t^{(a)} \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,\big] \\ &= E_Q\big[\, V_{t-1} \, \mathrm{I}_{\{|\xi_t| \leq a\}} \mid \mathcal{F}_{t-1} \,\big] \\ &= V_{t-1} \, \mathrm{I}_{\{|\xi_t| \leq a\}}. \end{aligned}$$

By sending $a \uparrow \infty$, we obtain (5.10).

(c) ⇒ (d): By (5.2), every Q-martingale M satisfies

$$M_0 = E_Q[\, M_T \mid \mathcal{F}_0 \,] = E_Q[\, M_T \,].$$

(d) ⇒ (a): To prove that $X_t^i \in \mathcal{L}^1(Q)$ for given i and t, consider the deterministic process ξ defined by $\xi_s^i := \mathrm{I}_{\{s \leq t\}}$ and $\xi_s^j := 0$ for $j \neq i$. By Remark 5.8, ξ can be complemented with a predictable process ξ^0 such that $\overline{\xi} = (\xi^0, \xi)$ is a self-financing strategy with initial investment $V_0 = X_0^i$. The corresponding value process satisfies

$$V_T = V_0 + \sum_{s=1}^{T} \xi_s \cdot (X_s - X_{s-1}) = X_t^i \geq 0.$$

From (d) we get

$$E_Q[\, X_t^i \,] = E_Q[\, V_T \,] = V_0 = X_0^i, \tag{5.11}$$

which yields $X_t^i \in \mathcal{L}^1(Q)$.

Condition (a) will follow if we can show that $E_Q[\, X_t^i;\ A \,] = E_Q[\, X_{t-1}^i;\ A \,]$ for given t, i, and $A \in \mathcal{F}_{t-1}$. To this end, we define a d-dimensional predictable process η by $\eta_s^i := \mathrm{I}_{\{s<t\}} + \mathrm{I}_{A^c}\mathrm{I}_{\{s=t\}}$ and $\eta_s^j := 0$ for $j \neq i$. As above, we take a predictable process η^0 such that $\overline{\eta} = (\eta^0, \eta)$ is a self-financing strategy with initial investment $\widetilde{V}_0 = X_0^i$. Its terminal value is given by

$$\widetilde{V}_T = \widetilde{V}_0 + \sum_{s=1}^{T} \eta_s \cdot (X_s - X_{s-1}) = X_t^i \mathrm{I}_{A^c} + X_{t-1}^i \mathrm{I}_A \geq 0.$$

Using (d) yields

$$X_0^i = \widetilde{V}_0 = E_Q[\, \widetilde{V}_T \,] = E_Q[\, X_t^i;\ A^c \,] + E_Q[\, X_{t-1}^i;\ A \,].$$

By comparing this identity with (5.11), we conclude that $E_Q[X_t^i;\ A] = E_Q[X_{t-1}^i;\ A]$. □

Remark 5.16. (a) Suppose that the "objective" measure P is itself a martingale measure, so that the fluctuation of prices may be viewed as a fair game. In this case, the preceding proposition shows that there are no realistic self-financing strategies which would generate a positive expected gain. Thus, the assumption $P \in \mathcal{P}$ is a strong version of the so-called *efficient market hypothesis*. For a market model containing a locally risk-less bond, this strong hypothesis would imply that risk-averse investors would not be attracted towards investing into the risky assets if their expectations are consistent with P; see Example 2.43.

(b) The strong assumption $P \in \mathcal{P}$ implies, in particular, that there is no arbitrage opportunity, i.e., no self-financing strategy with positive expected gain and without any downside risk. Indeed, Theorem 5.15 implies that the value process of any self-financing strategy with $V_0 \leq 0$ and $V_T \geq 0$ satisfies $E[\, V_T \,] = V_0$, hence $V_T = 0$ P-almost surely. The assumption that the market model is arbitrage-free may be viewed as a much milder and hence more flexible form of the *efficient market hypothesis*. ◇

We can now state the following dynamic version of the "fundamental theorem of asset pricing", which relates the absence of arbitrage opportunities to the existence of equivalent martingale measures.

Theorem 5.17. *The market model is arbitrage-free if and only if the set $\mathcal{P}$ of all equivalent martingale measures is non-empty. In this case, there exists a $P^* \in \mathcal{P}$ with bounded density dP^*/dP.*

Proof. Suppose first that there exists an equivalent martingale measure P^*. Then it follows as in Remark 5.16 (b) that the market model in which the probability measure P is replaced by P^* is arbitrage-free. Since the notion of an arbitrage opportunity depends on the underlying measure only through its null sets and since these are common for the two equivalent measures P and P^*, it follows that also the original market model is arbitrage-free.

Let us turn to the proof of the converse assertion. For $t \in \{1, \dots, T\}$, we define

$$\mathcal{K}_t := \left\{ \eta \cdot (X_t - X_{t-1}) \mid \eta \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d) \right\}. \tag{5.12}$$

By Proposition 5.11, the market model is arbitrage-free if and only if

$$\mathcal{K}_t \cap L^0_+(\Omega, \mathcal{F}_t, P) = \{0\} \tag{5.13}$$

holds for all t. Note that (5.13) depends on the measure P only through its null sets.

Condition (5.13) allows us to apply Theorem 1.54 to the t^{th} trading period. For $t = T$ we obtain a probability measure $\widetilde{P}_T \approx P$ which has a bounded density $d\widetilde{P}_T/dP$ and which satisfies

$$\widetilde{E}_T[\, X_T - X_{T-1} \mid \mathcal{F}_{T-1}\,] = 0.$$

Now suppose that we already have a probability measure $\widetilde{P}_{t+1} \approx P$ with a bounded density $d\widetilde{P}_{t+1}/dP$ such that

$$\widetilde{E}_{t+1}[\, X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = 0 \quad \text{for } t+1 \le k \le T. \tag{5.14}$$

The equivalence of $\widetilde{P}_{t+1}$ and P implies that (5.13) also holds with P replaced by $\widetilde{P}_{t+1}$. Applying Theorem 1.54 to the t^{th} trading period yields a probability measure $\widetilde{P}_t$ with a bounded $\mathcal{F}_t$-measurable density $Z_t := d\widetilde{P}_t/d\widetilde{P}_{t+1} > 0$ such that

$$\widetilde{E}_t[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,] = 0.$$

Clearly, $\widetilde{P}_t$ is equivalent to P and has a bounded density, since

$$\frac{d\widetilde{P}_t}{dP} = \frac{d\widetilde{P}_t}{d\widetilde{P}_{t+1}} \cdot \frac{d\widetilde{P}_{t+1}}{dP}$$

is the product of two bounded densities. Moreover, if $t+1 \le k \le T$, Proposition A.12 and the $\mathcal{F}_t$-measurability of $Z_t = d\widetilde{P}_t/d\widetilde{P}_{t+1}$ imply

$$\begin{aligned}
\widetilde{E}_t[\, X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] &= \frac{\widetilde{E}_{t+1}[\,(X_k - X_{k-1})Z_t \mid \mathcal{F}_{k-1}\,]}{\widetilde{E}_{t+1}[\, Z_t \mid \mathcal{F}_{k-1}\,]} \\
&= \widetilde{E}_{t+1}[\, X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] \\
&= 0.
\end{aligned}$$

Hence, (5.14) carries over from $\widetilde{P}_{t+1}$ to $\widetilde{P}_t$. We can repeat this recursion until finally $P^* := \widetilde{P}_1$ yields the desired equivalent martingale measure. □

Clearly, the absence of arbitrage in the market is independent of the choice of the numéraire, while the set $\mathcal{P}$ of equivalent martingale measures generally does depend on the numéraire. In order to investigate the structure of this dependence, suppose that the first asset S^1 is P-a.s. strictly positive, so that it can serve as an alternative numéraire. The price process discounted by S^1 is denoted by

$$\overline{Y}_t = (Y_t^0, Y_t^1, \dots, Y_t^d) := \left(\frac{S_t^0}{S_t^1}, 1, \frac{S_t^2}{S_t^1}, \dots, \frac{S_t^d}{S_t^1} \right) = \frac{S_t^0}{S_t^1} \overline{X}_t, \quad t = 0, \dots, T.$$

Let $\widetilde{\mathcal{P}}$ be the set of equivalent martingale measures for $\overline{Y}$. Then $\widetilde{\mathcal{P}} \neq \emptyset$ if and only if $\mathcal{P} \neq \emptyset$, according to Theorem 5.17 and the fact that the existence of arbitrage opportunities is independent of the choice of the numéraire.

Proposition 5.18. *The two sets $\mathcal{P}$ and $\widetilde{\mathcal{P}}$ are related via the identity*

$$\widetilde{\mathcal{P}} = \left\{ \widetilde{P}^* \;\Big|\; \frac{d\widetilde{P}^*}{dP^*} = \frac{X_T^1}{X_0^1} \textit{ for some } P^* \in \mathcal{P} \right\}.$$

Proof. The process X_t^1/X_0^1 is a P^*-martingale for any $P^* \in \mathcal{P}$. In particular, $E^*[\, X_T^1/X_0^1 \,] = 1$, and the formula

$$\frac{d\widetilde{P}^*}{dP^*} = \frac{X_T^1}{X_0^1}$$

defines a probability measure $\widetilde{P}^*$ which is equivalent to P. Moreover, by Proposition A.12,

$$\begin{aligned} \widetilde{E}^*[\, \overline{Y}_t \mid \mathcal{F}_s \,] &= \frac{1}{X_s^1} \cdot E^*[\, \overline{Y}_t \cdot X_t^1 \mid \mathcal{F}_s \,] \\ &= \frac{1}{X_s^1} \cdot E^*[\, \overline{X}_t \mid \mathcal{F}_s \,] \\ &= \overline{Y}_s. \end{aligned}$$

Hence, $\widetilde{P}^*$ is an equivalent martingale measure for $\overline{Y}$, and it follows that

$$\widetilde{\mathcal{P}} \supseteq \left\{ \widetilde{P}^* \;\Big|\; \frac{d\widetilde{P}^*}{dP^*} = \frac{X_T^1}{X_0^1} \text{ for some } P^* \in \mathcal{P} \right\}.$$

Reversing the roles of $\overline{X}$ and $\overline{Y}$ yields the identity of the two sets. □

Remark 5.19. Unless X_T^1 is P-a.s. constant, the two sets $\mathcal{P}$ and $\widetilde{\mathcal{P}}$ satisfy

$$\mathcal{P} \cap \widetilde{\mathcal{P}} = \emptyset.$$

This can be proved as in Remark 1.10. ◇

5.3 European contingent claims

A key topic of mathematical finance is the analysis of *derivative securities* or *contingent claims*, i.e., of certain assets whose payoff depends on the behavior of the primary assets $S^0, S^1, \dots, S^d$ and, in some cases, also on other factors.

Definition 5.20. A non-negative random variable C on $(\Omega, \mathcal{F}_T, P)$ is called a *European contingent claim*. A European contingent claim C is called a *derivative* of the underlying assets $S^0, S^1, \dots, S^d$ if C is measurable with respect to the σ-algebra generated by the price process $(\overline{S}_t)_{t=0,\dots,T}$.

A European contingent claim has the interpretation of an asset which yields at time T the amount $C(\omega)$, depending on the scenario ω of the market evolution. T is called the *expiration date* or the *maturity* of C. Of course, maturities prior to the final trading period T of our model are also possible, but unless it is otherwise mentioned, we will assume that our European contingent claims expire at T. In Chapter 6, we will meet another class of derivative securities, the so-called *American contingent claims*. As long as there is no risk of confusion between European and American contingent claims, we will use the term “contingent claim” to refer to a European contingent claim.

Example 5.21. The owner of a *European call option* has the right, but not the obligation, to buy an asset at time T for a fixed price K, called the *strike price*. This corresponds to a contingent claim of the form

$$C^{\text{call}} = (S_T^i - K)^+.$$

Conversely, a *European put option* gives the right, but not the obligation, to sell the asset at time T for a strike price K. This corresponds to the contingent claim

$$C^{\text{put}} = (K - S_T^i)^+. \qquad \diamond$$

Example 5.22. The payoff of an *Asian option* depends on the average price

$$S_{\text{av}}^i := \frac{1}{|\mathbb{T}|} \sum_{t \in \mathbb{T}} S_t^i$$

of the underlying asset during a predetermined set of periods $\mathbb{T} \subset \{0, \dots, T\}$. For instance, an *average price call* with strike K corresponds to the contingent claim

$$C_{\text{av}}^{\text{call}} := (S_{\text{av}}^i - K)^+,$$

and an *average price put* has the payoff

$$C_{\text{av}}^{\text{put}} := (K - S_{\text{av}}^i)^+.$$

Average price options can be used, for instance, to secure regular cash streams against exchange rate fluctuations. For example, assume that an economic agent receives at each time $t \in \mathbb{T}$ a fixed amount of a foreign currency with exchange rates S_t^i. In this case, an average price put option may be an efficient instrument for securing the incoming cash stream against the risk of unfavorable exchange rates.

An *average strike call* corresponds to the contingent claim

$$(S_T^i - S_{\mathrm{av}}^i)^+,$$

while an *average strike put* pays off the amount

$$(S_{\mathrm{av}}^i - S_T^i)^+.$$

An average strike put can be used, for example, to secure the risk from selling at time T a quantity of an asset which was bought at successive times over the period $\mathbb{T}$. ◇

Example 5.23. The payoff of a *barrier option* depends on whether the price of the underlying asset reaches a certain level before maturity. Most barrier options are either *knock-out* or *knock-in options*. A knock-in option pays off only if the barrier B is reached. The simplest example is a *digital option*

$$C^{\mathrm{dig}} := \begin{cases} 1 & \text{if } \max\limits_{0 \le t \le T} S_t^i \ge B, \\ 0 & \text{otherwise,} \end{cases}$$

which has a unit payoff if the price processes reaches a given upper barrier $B > S_0^i$. Another example is the *down-and-in put* with strike price K and lower barrier $\widetilde{B} < S_0^i$ which pays off

$$C_{\mathrm{d\&i}}^{\mathrm{put}} := \begin{cases} (K - S_T^i)^+ & \text{if } \min\limits_{0 \le t \le T} S_t^i \le \widetilde{B}, \\ 0 & \text{otherwise.} \end{cases}$$

A knock-out barrier option has a zero payoff once the price of the underlying asset reaches the predetermined barrier. For instance, an *up-and-out call* corresponds to the contingent claim

$$C_{\mathrm{u\&o}}^{\mathrm{call}} := \begin{cases} (S_T^i - K)^+ & \text{if } \max\limits_{0 \le t \le T} S_t^i < B, \\ 0 & \text{otherwise;} \end{cases}$$

see Figure 5.1. *Down-and-out* and *up-and-in options* are defined analogously. ◇

Example 5.24. Using a *lookback option*, one can trade the underlying asset at the maximal or minimal price that occurred during the life of the option. A *lookback call* has the payoff

$$S_T^i - \min_{0 \le t \le T} S_t^i,$$

while a *lookback put* corresponds to the contingent claim

$$\max_{0 \le t \le T} S_t^i - S_T^i. \qquad \diamond$$

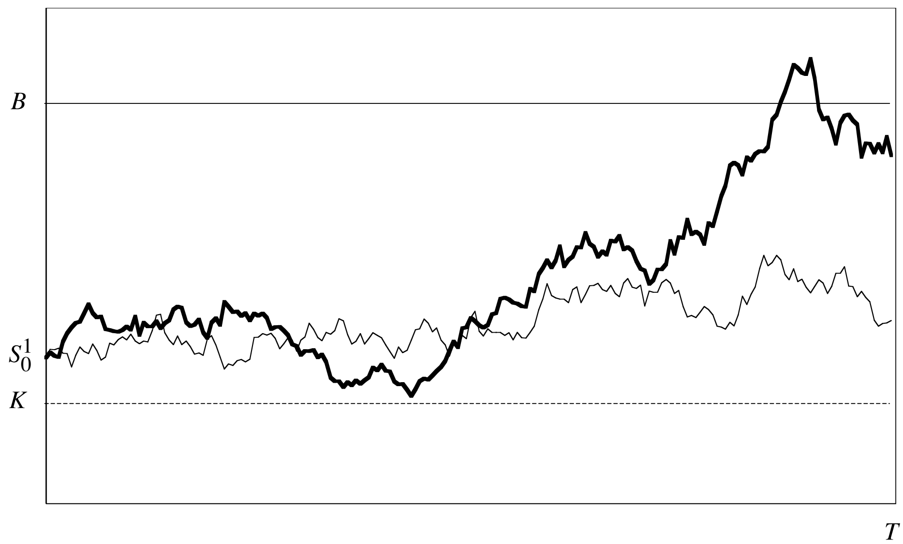


Figure 5.1. In one scenario, the payoff of the up-and-out call becomes zero because the stock price hits the barrier B before time T. In the other scenario, the payoff is given by $(S_T - K)^+$.

The discounted value of a contingent claim C when using the numéraire S^0 is given by

$$H := \frac{C}{S_T^0}.$$

We will call H the *discounted European claim* or just the *discounted claim* associated with C. In the remainder of this text, "H" will be the generic notation for the discounted payoff of any type of contingent claim.

The reader may wonder why we work simultaneously with the notions of a contingent claim and a discounted claim. From a purely mathematical point of view, there would be no loss of generality in assuming that the numéraire asset is identically equal to one. In fact, the entire theory to be developed in Part II can be seen as a discrete-time "stochastic analysis" for the d-dimensional process $X = (X^1, \dots, X^d)$ and its "stochastic integrals"

$$\sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1})$$

of predictable d-dimensional processes ξ. However, some of the economic intuition would be lost if we would limit the discussion to this level. For instance, we have already seen the economic relevance of the particular choice of the numéraire, even though this choice may be irrelevant from the mathematician's point of view. As a compromise between the mathematician's preference for conciseness and the economist's concern for keeping track explicitly of economically relevant quantities, we develop the mathematics on the level of discounted prices, but we will continue to discuss definitions and results in terms of undiscounted prices whenever it seems appropriate.

From now on, we will assume that our market model is arbitrage-free or, equivalently, that

$$\mathcal{P} \neq \emptyset.$$

Definition 5.25. A contingent claim C is called *attainable* (*replicable, redundant*) if there exists a self-financing trading strategy $\overline{\xi}$ whose terminal portfolio value coincides with C, i.e.,

$$C = \overline{\xi}_T \cdot \overline{S}_T \quad P\text{-a.s.}$$

Such a trading strategy $\overline{\xi}$ is called a *replicating strategy* for C.

Clearly, a contingent claim C is attainable if and only if the corresponding discounted claim $H = C/S_T^0$ is of the form

$$H = \overline{\xi}_T \cdot \overline{X}_T = V_T = V_0 + \sum_{t=1}^{T} \xi_t \cdot (X_t - X_{t-1}),$$

for a self-financing trading strategy $\overline{\xi} = (\xi^0, \xi)$ with value process V. In this case, we will say that the discounted claim H is *attainable*, and we will call $\overline{\xi}$ a replicating strategy for H. The following theorem yields the surprising result that an attainable discounted claim is automatically integrable with respect to every equivalent martingale measure. Note, however, that integrability may not hold for an attainable contingent claim *prior* to discounting.

Theorem 5.26. *Any attainable discounted claim H is integrable with respect to each equivalent martingale measure, i.e.,*

$$E^*[\, H \,] < \infty \quad \textit{for all } P^* \in \mathcal{P}.$$

Moreover, for each $P^* \in \mathcal{P}$ the value process of any replicating strategy satisfies

$$V_t = E^*[\, H \mid \mathcal{F}_t \,] \quad P\textit{-a.s. for } t = 0, \dots, T.$$

In particular, V is a non-negative P^*-martingale.

Proof. This follows from $V_T = H \geq 0$ and the systems theorem in the form of Theorem 5.15. □

Remark 5.27. The identity

$$V_t = E^*[\, H \mid \mathcal{F}_t \,], \quad t = 0, \dots, T,$$

appearing in Theorem 5.26 has two remarkable implications. Since its right-hand side is independent of the particular replicating strategy, all such strategies must have the same value process. Moreover, the left-hand side does not depend on the choice of $P^* \in \mathcal{P}$. Hence, V_t is a version of the conditional expectation $E^*[\,H \mid \mathcal{F}_t\,]$ for every $P^* \in \mathcal{P}$. In particular, $E^*[\,H\,]$ is the same for all $P^* \in \mathcal{P}$. ◇

Remark 5.28. When applied to an attainable contingent claim C prior to discounting, Theorem 5.26 states that

$$\overline{\xi}_t \cdot \overline{S}_t = S_t^0 E^*\left[\frac{C}{S_T^0} \;\middle|\; \mathcal{F}_t\right], \quad t = 0, \dots, T,$$

P-a.s. for all $P^* \in \mathcal{P}$ and for every replicating strategy $\overline{\xi}$. In particular, the initial investment which is needed for a replication of C is given by

$$\overline{\xi}_1 \cdot \overline{S}_0 = S_0^0 E^*\left[\frac{C}{S_T^0}\right].$$ ◇

Let us now turn to the problem of *pricing* a contingent claim. Consider first a discounted claim H which is attainable. Then the (discounted) initial investment

$$\overline{\xi}_1 \cdot \overline{X}_0 = V_0 = E^*[\,H\,] \tag{5.15}$$

needed for the replication of H can be interpreted as the unique (discounted) "*fair price*" of H. In fact, a different price for H would create an arbitrage opportunity. For instance, if H could be sold at time 0 for a price $\widetilde{\pi}$ which is higher than (5.15), then selling H and buying the replicating portfolio $\overline{\xi}$ yields the profit

$$\widetilde{\pi} - \overline{\xi}_1 \cdot \overline{X}_0 > 0$$

at time 0, although the terminal portfolio value $V_T = \overline{\xi}_T \cdot \overline{X}_T$ suffices for settling the claim H at maturity T. In order to make this idea precise, let us formalize the idea of an "arbitrage-free price" of a general discounted claim H.

Definition 5.29. A real number $\pi^H \geq 0$ is called an *arbitrage-free price* of a discounted claim H, if there exists an adapted stochastic process X^{d+1} such that

$$\begin{aligned} X_0^{d+1} &= \pi^H, \\ X_t^{d+1} &\geq 0 \ \text{ for } t = 1, \dots, T-1, \text{ and} \\ X_T^{d+1} &= H, \end{aligned} \tag{5.16}$$

and such that the enlarged market model with price process $(X^0, X^1, \dots, X^d, X^{d+1})$ is arbitrage-free. The set of all arbitrage-free prices of H is denoted by $\Pi(H)$. The lower and upper bounds of $\Pi(H)$ are denoted by

$$\pi_{\inf}(H) := \inf \Pi(H) \quad \text{and} \quad \pi_{\sup}(H) := \sup \Pi(H).$$

Thus, an arbitrage-free price π^H of a discounted claim H is by definition a price at which H can be traded at time 0 without introducing arbitrage opportunities into the market model: If H is sold for π^H, then neither buyer nor seller can find an investment strategy which both eliminates all the risk and yields an opportunity to make a positive profit. Our aim in this section is to characterize the set of all arbitrage-free prices of a discounted claim H.

Note that an arbitrage-free price π^H is quoted in units of the numéraire asset. The amount that corresponds to π^H in terms of currency units prior to discounting is equal to

$$\pi^C := S_0^0 \pi^H,$$

and π^C is an (undiscounted) arbitrage-free price of the contingent claim $C := S_T^0 H$.

Theorem 5.30. *The set of arbitrage-free prices of a discounted claim H is non-empty and given by*

$$\Pi(H) = \left\{ E^*[\, H \,] \mid P^* \in \mathcal{P} \text{ and } E^*[\, H \,] < \infty \right\}. \tag{5.17}$$

Moreover, the lower and upper bounds of $\Pi(H)$ are given by

$$\pi_{\text{inf}}(H) = \inf_{P^* \in \mathcal{P}} E^*[\, H \,] \quad \text{and} \quad \pi_{\text{sup}}(H) = \sup_{P^* \in \mathcal{P}} E^*[\, H \,].$$

Proof. By Theorem 5.17, π^H is an arbitrage-free price for H if and only if we can find an equivalent martingale measure $\hat{P}$ for the market model extended via (5.16). $\hat{P}$ must satisfy

$$X_t^i = \hat{E}[\, X_T^i \mid \mathcal{F}_t] \quad \text{for } t = 0, \dots, T \text{ and } i = 1, \dots, d+1.$$

In particular, $\hat{P}$ belongs to $\mathcal{P}$ and satisfies $\pi^H = \hat{E}[\, H \,]$. Thus, we obtain the inclusion $\subseteq$ in (5.17).

Conversely, if $\pi^H = E^*[\, H \,]$ for some $P^* \in \mathcal{P}$, then we can define the stochastic process

$$X_t^{d+1} := E^*[\, H \mid \mathcal{F}_t], \quad t = 0, \dots, T,$$

which satisfies all the requirements of (5.16). Moreover, the same measure P^* is clearly an equivalent martingale measure for the extended market model, which hence is arbitrage-free. Thus, we obtain the identity of the two sets in (5.17).

To show that $\Pi(H)$ is non-empty, we first fix some measure $\widetilde{P} \approx P$ such that $\widetilde{E}[\, H \,] < \infty$. For instance, we can take $d\widetilde{P} = c(1+H)^{-1}\, dP$, where c is the normalizing constant. Under $\widetilde{P}$, the market model is arbitrage-free. Hence, Theorem 5.17 yields $P^* \in \mathcal{P}$ such that $dP^*/d\widetilde{P}$ is bounded. In particular, $E^*[\, H \,] < \infty$ and hence $E^*[\, H \,] \in \Pi(H)$.

The formula for $\pi_{\text{inf}}(H)$ follows immediately from (5.17) and the fact that $\Pi(H) \neq \emptyset$. The one for $\pi_{\text{sup}}(H)$ needs an additional argument. Suppose that $P^\infty \in \mathcal{P}$ is such that $E^\infty[\, H \,] = \infty$. We must show that for any $c > 0$ there exists some

$\pi \in \Pi(H)$ with $\pi > c$. To this end, let n be such that $\widetilde{\pi} := E^\infty[\, H \wedge n \,] > c$, and define

$$X_t^{d+1} := E^\infty[\, H \wedge n \mid \mathcal{F}_t \,], \quad t = 0, \dots, T.$$

Then P^∞ is an equivalent martingale measure for the extended market model $(X^0, \dots, X^d, X^{d+1})$, which hence is arbitrage-free. Applying the already established fact that the set of arbitrage-free prices of any contingent claim is nonempty to the extended market model yields an equivalent martingale measure P^* for $(X^0, \dots, X^d, X^{d+1})$ such that $E^*[\, H \,] < \infty$. Since P^* is also a martingale measure for the original market model, the first part of this proof implies that $\pi := E^*[\, H \,] \in \Pi(H)$. Finally, note that

$$\pi = E^*[\, H \,] \geq E^*[\, H \wedge n \,] = E^*[\, X_T^{d+1} \,] = X_0^{d+1} = \widetilde{\pi} > c.$$

Hence, the formula for $\pi_{\text{sup}}(H)$ is proved. □

Example 5.31. In an arbitrage-free market model, we consider a European call option $C^{\text{call}} = (S_T^1 - K)^+$ with strike $K > 0$ and with maturity T. We assume that the numéraire S^0 is the predictable price process of a locally riskless bond as in Example 5.5. Then S_t^0 is increasing in t and satisfies $S_0^0 \equiv 1$. For any $P^* \in \mathcal{P}$, Theorem 5.30 yields an arbitrage-free price π^{call} of C^{call} which is given by

$$\pi^{\text{call}} = E^*\left[\, \frac{C^{\text{call}}}{S_T^0} \,\right] = E^*\left[\left(X_T^1 - \frac{K}{S_T^0} \right)^+\right].$$

Due to the convexity of the function $x \mapsto x^+ = x \vee 0$ and our assumptions on S^0, π^{call} can be bounded below as follows:

$$\pi^{\text{call}} \geq \left(E^*\left[\, X_T^1 - \frac{K}{S_T^0} \,\right]\right)^+ = \left(S_0^1 - E^*\left[\, \frac{K}{S_T^0} \,\right]\right)^+ \geq (S_0^1 - K)^+.$$

In financial language, this fact is usually expressed by saying that the value of the option is higher than its "intrinsic value" $(S_0^1 - K)^+$, i.e., the payoff if the option were exercised immediately. The difference of the price π^{call} of an option and its intrinsic value is often called the "time-value" of the European call option; see Figure 5.2. ◇

Example 5.32. For a European put option $C^{\text{put}} = (K - S_T^1)^+$, the situation is more complicated. If we consider the same situation as in Example 5.31, then the analogue of (5.31) fails unless the numéraire S^0 is constant. In fact, as a consequence of the put-call parity, the "time value" of a put option whose intrinsic value is large (i.e., the option is "in the money") usually becomes negative; see Figure 5.3. ◇

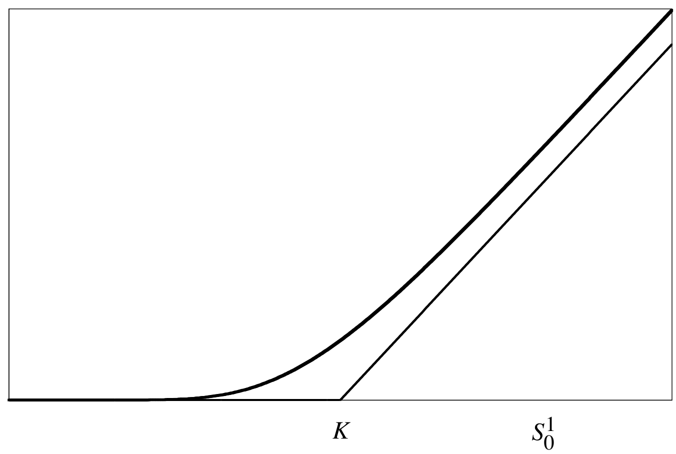


Figure 5.2. The typical price of a call option as a function of S_0^1 is always above the option's intrinsic value $(S_0^1 - K)^+$.

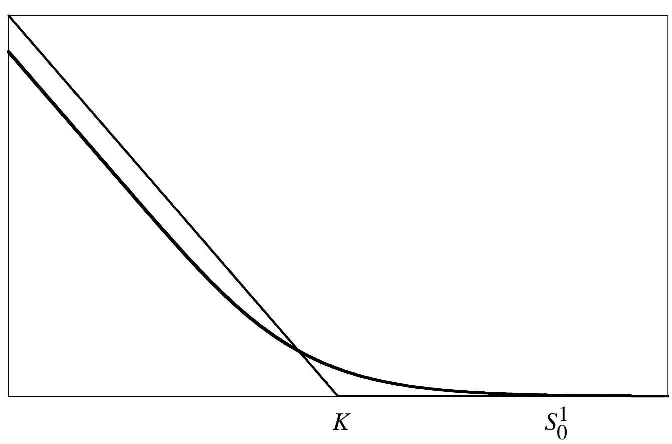


Figure 5.3. The typical price of a European put option as a function of S_0^1 compared to the option's intrinsic value $(K - S_0^1)^+$.

Our next aim is to characterize the structure of the set of arbitrage-free prices of a discounted claim H. It follows from Theorem 5.30 that every arbitrage-free price π^H of H must lie between the two numbers

$$\pi_{\inf}(H) = \inf_{P^* \in \mathcal{P}} E^*[\, H \,] \quad \text{and} \quad \pi_{\sup}(H) = \sup_{P^* \in \mathcal{P}} E^*[\, H \,].$$

We also know that $\pi_{\inf}(H)$ and $\pi_{\sup}(H)$ are equal if H is attainable. The following theorem shows that also the converse implication holds, i.e., H is attainable if and only if $\pi_{\inf}(H) = \pi_{\sup}(H)$.

Theorem 5.33. *Let H be a discounted claim.*

(a) *If H is attainable, then the set $\Pi(H)$ of arbitrage-free prices for H consists of the single element V_0, where V is the value process of any replicating strategy for H.*

(b) *If H is not attainable, then $\pi_{\inf}(H) < \pi_{\sup}(H)$ and*

$$\Pi(H) = \big(\pi_{\inf}(H), \pi_{\sup}(H)\big).$$

Proof. The first assertion follows from Remark 5.27 and Theorem 5.30.

To prove (b), note first that

$$\Pi(H) = \{\, E^*[\, H\,] \mid P^* \in \mathcal{P},\ E^*[\, H\,] < \infty \,\}$$

is an interval because $\mathcal{P}$ is a convex set. We will show that $\Pi(H)$ is open by constructing for any $\pi \in \Pi(H)$ two arbitrage-free prices $\check{\pi}$ and $\hat{\pi}$ for H such that $\check{\pi} < \pi < \hat{\pi}$. To this end, take $P^* \in \mathcal{P}$ such that $\pi = E^*[\, H\,]$. We will first construct an equivalent martingale measure $\hat{P} \in \mathcal{P}$ such that $\hat{E}[\, H\,] > E^*[\, H\,]$. Let

$$U_t := E^*[\, H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T,$$

so that

$$H = U_0 + \sum_{t=1}^{T} (U_t - U_{t-1}).$$

Since H is not attainable, there must be some $t \in \{1, \dots, T\}$ such that $U_t - U_{t-1} \notin \mathcal{K}_t \cap L^1(P^*)$, where

$$\mathcal{K}_t := \big\{\, \eta \cdot (X_t - X_{t-1}) \mid \eta \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d) \,\big\}.$$

By Lemma 1.68, $\mathcal{K}_t \cap L^1(P^*)$ is a closed linear subspace of $L^1(\Omega, \mathcal{F}_t, P^*)$. Therefore, Theorem A.56 applied with $\mathcal{B} := \{U_t - U_{t-1}\}$ and $\mathcal{C} := \mathcal{K}_t \cap L^1(P^*)$ yields some $Z \in L^\infty(\Omega, \mathcal{F}_t, P^*)$ such that

$$\sup \big\{ E^*[\, W Z\,] \mid W \in \mathcal{K}_t \cap L^1(P^*) \big\} < E^*[\, (U_t - U_{t-1})\, Z\,] < \infty.$$

From the linearity of $\mathcal{K}_t \cap L^1(P^*)$ we deduce that

$$E^*[\, W Z\,] = 0 \quad \text{for all } W \in \mathcal{K}_t \cap L^1(P^*), \tag{5.18}$$

and hence that

$$E^*[\, (U_t - U_{t-1})\, Z\,] > 0. \tag{5.19}$$

There is no loss of generality in assuming that $|Z| \le 1/3$, so that

$$\hat{Z} := 1 + Z - E^*[\, Z \mid \mathcal{F}_{t-1}\,]$$

can be taken as the density $d\hat{P}/dP^* = \hat{Z}$ of a new probability measure $\hat{P} \approx P$. Since Z is $\mathcal{F}_t$-measurable, the expectation of H under $\hat{P}$ satisfies

$$\begin{aligned}\hat{E}[\,H\,] &= E^*[\,H\,\hat{Z}\,]\\ &= E^*[\,H\,] + E^*\big[\,E^*[\,H \mid \mathcal{F}_t]\,Z\,\big] - E^*\big[\,E^*[\,H \mid \mathcal{F}_{t-1}]\,E^*[\,Z \mid \mathcal{F}_{t-1}]\,\big]\\ &= E^*[\,H\,] + E^*[\,U_t\,Z\,] - E^*[\,U_{t-1}\,Z\,]\\ &> E^*[\,H\,],\end{aligned}$$

where we have used (5.19) in the last step. On the other hand, $\hat{E}[\,H\,] \le \frac{5}{3}E^*[\,H\,] < \infty$. Thus, $\hat{\pi} := \hat{E}[\,H\,]$ will yield the desired arbitrage-free price larger than π if we have $\hat{P} \in \mathcal{P}$.

Let us prove that $\hat{P} \in \mathcal{P}$. For $k > t$, the $\mathcal{F}_t$-measurability of $\hat{Z}$ and Proposition A.12 yield that

$$\hat{E}[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = E^*[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = 0.$$

For $k = t$, (5.18) yields $E^*[\,(X_t - X_{t-1})\,Z \mid \mathcal{F}_{t-1}\,] = 0$. Thus, it follows from $E^*[\,\hat{Z} \mid \mathcal{F}_{t-1}\,] = 1$ that

$$\begin{aligned}&\hat{E}[\,X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]\\ &\quad= E^*\big[\,(X_t - X_{t-1})\big(1 - E^*[\,Z \mid \mathcal{F}_{t-1}]\big) \mid \mathcal{F}_{t-1}\,\big] + E^*[\,(X_t - X_{t-1})Z \mid \mathcal{F}_{t-1}\,]\\ &\quad= 0.\end{aligned}$$

Finally, if $k < t$ then P^* and $\hat{P}$ coincide on $\mathcal{F}_k$. Hence

$$\hat{E}[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = E^*[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = 0,$$

and we may conclude that $\hat{P} \in \mathcal{P}$.

It remains to construct another equivalent martingale measure $\check{P}$ such that

$$\check{\pi} := \check{E}[\,H\,] < E^*[\,H\,] = \pi. \tag{5.20}$$

But this is simply achieved by letting

$$\frac{d\check{P}}{dP^*} := 2 - \frac{d\hat{P}}{dP^*},$$

which defines a probability measure $\check{P} \approx P$, because the density $d\hat{P}/dP^*$ is bounded above by 5/3 and below by 1/3. $\check{P} \in \mathcal{P}$ is then obvious as is (5.20). □

Remark 5.34. So far, we have assumed that a contingent claim is settled at the terminal time T. A natural way of dealing with an $\mathcal{F}_{T_0}$-measurable payoff $C_0 \ge 0$ maturing at some time $T_0 < T$ is to apply our results to the corresponding discounted claim

$$H_0 := \frac{C_0}{S^0_{T_0}}$$

in the market model with the restricted time horizon T_0. Clearly, this restricted model is arbitrage-free. An alternative approach is to invest the payoff C_0 at time T_0 into the numéraire asset S^0. At time T, this yields the contingent claim

$$C := C_0 \cdot \frac{S_T^0}{S_{T_0}^0},$$

whose discounted claim

$$H = \frac{C}{S_T^0} = \frac{C_0}{S_{T_0}^0}$$

is formally identical to H_0. Moreover, our results can be directly applied to H. It is intuitively clear that these two approaches for determining the arbitrage-free prices of C_0 should be equivalent. A formal proof must show that the set $\Pi(H)$ is equal to the set

$$\Pi(H_0) := \left\{ E_0^*[\, H_0 \,] \mid P_0^* \in \mathcal{P}_0 \text{ and } E_0^*[\, H_0 \,] < \infty \right\}$$

of arbitrage-free prices of H_0 in the market model whose time horizon is T_0. Here, $\mathcal{P}_0$ denotes the set of measures P_0^* on $(\Omega, \mathcal{F}_{T_0})$ which are equivalent to P on $\mathcal{F}_{T_0}$ and which are martingale measures for the restricted price process $(X_t)_{t=0,\dots,T_0}$. Clearly, each $P^* \in \mathcal{P}$ defines an element of $\mathcal{P}_0$ by restricting P^* to the σ-algebra $\mathcal{F}_{T_0}$. In fact, Proposition 5.35 below shows that every element in $\mathcal{P}_0$ arises in this way. Thus, the two sets of arbitrage-free prices for H and H_0 coincide, i.e.,

$$\Pi(H) = \Pi(H_0)$$

It follows, in particular, that H_0 is attainable if and only if H is attainable. ◇

Proposition 5.35. *Consider the situation described in Remark* 5.34 *and let* $P_0^* \in \mathcal{P}_0$ *be given. Then there exists some* $P^* \in \mathcal{P}$ *whose restriction to* $\mathcal{F}_{T_0}$ *is equal to* P_0^*.

Proof. Let $\hat{P} \in \mathcal{P}$ be arbitrary, and denote by Z_{T_0} the density of P_0^* with respect to the restriction of $\hat{P}$ to the σ-algebra $\mathcal{F}_{T_0}$. Then Z_{T_0} is $\mathcal{F}_{T_0}$-measurable, and

$$dP^* := Z_{T_0}\, d\hat{P}$$

defines a probability measure on $\mathcal{F}$. Clearly, P^* is equivalent to $\hat{P}$ and to P, and it coincides with P_0^* on $\mathcal{F}_{T_0}$. To check that $P^* \in \mathcal{P}$, it suffices to show that $X_t - X_{t-1}$ is a martingale increment under P^* for $t > T_0$. For these t, the density Z_{T_0} is $\mathcal{F}_{t-1}$-measurable, so Proposition A.12 implies that

$$E^*[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,] = \hat{E}[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,] = 0.$$ □

Example 5.36. Let us consider the situation of Example 5.31, where the numéraire S^0 is a locally riskless bond. Remark 5.34 allows us to compare the arbitrage-free prices of two European call options $C_0 = (S_{T_0}^1 - K)^+$ and $C = (S_T^1 - K)^+$ with the same strikes and underlyings but with different maturities $T_0 < T$. As in Example 5.31, we get that for $P^* \in \mathcal{P}$

$$E^*\Big[\frac{C}{S_T^0}\Bigm|\mathcal{F}_{T_0}\Big] \geq \frac{1}{S_{T_0}^0}\bigg(S_{T_0}^1 - K\,E^*\Big[\frac{S_{T_0}^0}{S_T^0}\Bigm|\mathcal{F}_{T_0}\Big]\bigg)^+ \geq \frac{C_0}{S_{T_0}^0}. \tag{5.21}$$

Hence, if P^* is used to calculate arbitrage-free prices for C_0 and C, the resulting price of C_0 is lower than the price of C:

$$E^*\Big[\frac{C}{S_T^0}\Big] \geq E^*\Big[\frac{C_0}{S_{T_0}^0}\Big].$$

This argument suggests that the price of a European call option should be an increasing function of the maturity. $\diamond$

5.4 Complete markets

We have seen in Theorem 5.33 that any attainable claim in an arbitrage-free market model has a unique arbitrage-free price. Thus, the situation becomes particularly transparent if all contingent claims are attainable.

Definition 5.37. An arbitrage-free market model is called *complete* if every contingent claim is attainable.

Complete market models are precisely those models in which every contingent claim has a unique and unambiguous arbitrage-free price. However, in discrete time, only a very limited class of models enjoys this property. The following characterization of market completeness is sometimes called the "second fundamental theorem of asset pricing".

Theorem 5.38. *An arbitrage-free market model is complete if and only if there exists exactly one equivalent martingale measure. In this case, the number of atoms in* $(\Omega, \mathcal{F}_T, P)$ *is bounded above by* $(d+1)^T$.

Proof. If the model is complete, then $H := I_A$ for $A \in \mathcal{F}_T$ is an attainable discounted claim. It follows from the results of Section 5.3 that the mapping $P^* \mapsto E^*[\,H\,] = P^*[\,A\,]$ is constant over the set $\mathcal{P}$. Hence, there can be only one equivalent martingale measure.

Conversely, if $|\mathcal{P}| = 1$, then the set $\Pi(H)$ of arbitrage-free prices of every discounted claim H has exactly one element. Hence, Theorem 5.33 implies that H is attainable.

To prove the second assertion, note first that the asserted bound on the number of atoms in $\mathcal{F}_T$ holds for $T = 1$ by Theorem 1.40 and Proposition 1.39. We proceed by induction on T. Suppose that the assertion holds for $T - 1$. By assumption, any bounded $\mathcal{F}_T$-measurable random variable $H \geq 0$ can be written as

$$H = V_{T-1} + \xi_T \cdot (X_T - X_{T-1}),$$

where both V_{T-1} and ξ_T are $\mathcal{F}_{T-1}$-measurable and hence constant on each atom A of $(\Omega, \mathcal{F}_{T-1}, P)$. It follows that the dimension of the linear space $L^\infty(\Omega, \mathcal{F}_T, P[\,\cdot\,|A])$ is less than or equal to $d + 1$. Thus, Proposition 1.39 implies that $(\Omega, \mathcal{F}_T, P[\,\cdot\,|A])$ has at most $d + 1$ atoms. Applying the induction hypothesis concludes the proof. □

Below we state additional characterizations of market completeness. Denote by $\mathcal{Q}$ the set of all martingale measures in the sense of Definition 5.14. Then both $\mathcal{P}$ and $\mathcal{Q}$ are convex sets. Recall that an element of a convex set is called an *extreme point* of this set if it cannot be written as a non-trivial convex combination of members of this set.

Property (d) in the following theorem is usually called the *predictable representation property*, or the *martingale representation property*, of the P^*-martingale X.

Theorem 5.39. *For $P^* \in \mathcal{P}$ the following conditions are equivalent:*

(a) $\mathcal{P} = \{P^*\}$.

(b) *P^* is an extreme point of $\mathcal{P}$*

(c) *P^* is an extreme point of $\mathcal{Q}$.*

(d) *Every P^*-martingale M can be represented as a "stochastic integral" of a d-dimensional predictable process ξ:*

$$M_t = M_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) \quad \textit{for } t = 0, \dots, T.$$

Proof. (a) ⇒ (c): If P^* can be written as $P^* = \alpha Q_1 + (1 - \alpha) Q_2$ for $\alpha \in (0, 1)$ and $Q_1, Q_2 \in \mathcal{Q}$, then Q_1 and Q_2 are both absolutely continuous with respect to P^*. By defining

$$P_i := \frac{1}{2}(Q_i + P^*), \quad i = 1, 2,$$

we thus obtain two martingale measures P_1 and P_2 which are equivalent to P^*. Hence, $P_1 = P_2 = P^*$ and, in turn, $Q_1 = Q_2 = P^*$.

(c) ⇒ (b): This is obvious since $\mathcal{P} \subset \mathcal{Q}$.

(b) ⇒ (a): Suppose that there exists a $\hat{P} \in \mathcal{P}$ which is different from P^*. We will show below that in this case $\hat{P}$ can be chosen such that the density $d\hat{P}/dP^*$ is bounded by some constant $c > 0$. Then, if $\varepsilon > 0$ is less than $1/c$,

$$\frac{dP'}{dP^*} := 1 + \varepsilon - \varepsilon \frac{d\hat{P}}{dP^*}$$

defines another measure $P' \in \mathcal{P}$ different from P^*. Moreover, P^* can be represented as the convex combination

$$P^* = \frac{\varepsilon}{1+\varepsilon}\hat{P} + \frac{1}{1+\varepsilon}P',$$

which contradicts condition (b). Hence, P^* must be the unique equivalent martingale measure.

It remains to prove the existence of $\hat{P} \in \mathcal{P}$ with a bounded density $d\hat{P}/dP^*$ if there exists some $\widetilde{P} \in \mathcal{P}$ which is different from P^*. Then there exists a set $A \in \mathcal{F}_T$ such that $P^*[\,A\,] \neq \widetilde{P}[\,A\,]$. We enlarge our market model by introducing the additional asset

$$X_t^{d+1} := \widetilde{P}[\,A \mid \mathcal{F}_t\,], \quad t = 0, \dots, T,$$

and we take P^* instead of P as our reference measure. By definition, $\widetilde{P}$ is an equivalent martingale measure for $(X^0, X^1, \dots, X^d, X^{d+1})$. Hence, the extended market model is arbitrage-free, and Theorem 5.17 guarantees the existence of an equivalent martingale measure $\hat{P}$ such that the density $d\hat{P}/dP^*$ is bounded. Moreover, $\hat{P}$ must be different from P^*, since P^* is not a martingale measure for X^{d+1}:

$$X_0^{d+1} = \widetilde{P}[\,A\,] \neq P^*[\,A\,] = E^*[\,X_T^{d+1}\,].$$

(a) ⇒(d): The terminal value M_T of a P^*-martingale M can be decomposed into the difference of its positive and negative parts:

$$M_T = M_T^+ - M_T^-.$$

M_T^+ and M_T^- can be regarded as two discounted claims, which are attainable by Theorem 5.38. Hence, there exist two d-dimensional predictable process ξ^+ and ξ^- such that

$$M_T^{\pm} = V_0^{\pm} + \sum_{k=1}^{T} \xi_k^{\pm} \cdot (X_k - X_{k-1}) \quad P^*\text{-a.s.}$$

for two non-negative constants V_0^+ and V_0^-. Since the value processes

$$V_t^{\pm} := V_0^{\pm} + \sum_{k=1}^{t} \xi_k^{\pm} \cdot (X_k - X_{k-1})$$

are P^*-martingales by Theorem 5.26, we get that

$$M_t = E^*[\,M_T^+ - M_T^- \mid \mathcal{F}_t\,] = V_t^+ - V_t^-.$$

This proves that the desired representation of M holds in terms of the d-dimensional predictable process $\xi := \xi^+ - \xi^-$.

(d) $\Rightarrow$(a): Applying our assumption to the martingale $M_t := P^*[\, A \mid \mathcal{F}_t\,]$ shows that $H = I_A$ is an attainable contingent claim. Hence, it follows from the results of Section 5.3 that the mapping $P^* \mapsto P^*[\, A\,]$ is constant over the set $\mathcal{P}$. Thus, there can be only one equivalent martingale measure. □

5.5 The binomial model

A complete financial market model with only one risky asset must have a binary tree structure, as we have seen in Theorem 5.38. Under an additional homogeneity assumption, this reduces to the following particularly simple model, which was introduced by Cox, Ross, and Rubinstein in [42]. It involves the riskless bond

$$S_t^0 := (1+r)^t, \quad t = 0, \dots, T,$$

with $r > -1$ and one risky asset $S^1 = S$, whose return

$$R_t := \frac{S_t - S_{t-1}}{S_{t-1}}$$

in the t^{th} trading period can only take two possible values $a, b \in \mathbb{R}$ such that

$$-1 < a < b.$$

Thus, the stock price jumps from S_{t-1} either to the higher value $S_t = S_{t-1}(1+b)$ or to the lower value $S_t = S_{t-1}(1+a)$. In this context, we are going to derive explicit formulas for the arbitrage-free prices and replicating strategies of various contingent claims.

Let us construct the model on the sample space

$$\Omega := \{-1, +1\}^T = \big\{\, \omega = (y_1, \dots, y_T) \mid y_i \in \{-1, +1\}\,\big\}.$$

Denote by

$$Y_t(\omega) := y_t \quad \text{for } \omega = (y_1, \dots, y_T) \tag{5.22}$$

the projection on the t^{th} coordinate, and let

$$R_t(\omega) := a\,\frac{1 - Y_t(\omega)}{2} + b\,\frac{1 + Y_t(\omega)}{2} = \begin{cases} a & \text{if } Y_t(\omega) = -1, \\ b & \text{if } Y_t(\omega) = +1. \end{cases}$$

The price process of the risky asset is modeled as

$$S_t := S_0 \prod_{k=1}^{t} (1 + R_k),$$

where the initial value $S_0 > 0$ is a given constant. The discounted price process takes the form

$$X_t = \frac{S_t}{S_t^0} = S_0 \prod_{k=1}^{t} \frac{1+R_k}{1+r}.$$

As filtration we take

$$\mathcal{F}_t := \sigma(S_0, \dots, S_t) = \sigma(X_0, \dots, X_t), \quad t = 0, \dots, T.$$

Note that $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and

$$\mathcal{F}_t = \sigma(Y_1, \dots, Y_t) = \sigma(R_1, \dots, R_t) \quad \text{for } t = 1, \dots, T;$$

$\mathcal{F} := \mathcal{F}_T$ coincides with the power set of Ω. Let us fix any probability measure P on $(\Omega, \mathcal{F})$ such that

$$P[\{\omega\}] > 0 \quad \text{for all } \omega \in \Omega. \tag{5.23}$$

Such a model will be called a *binomial model* or a *CRR model*. The following theorem characterizes those parameter values a, b, r for which the model is arbitrage-free.

Theorem 5.40. *The CRR model is arbitrage-free if and only if $a < r < b$. In this case, the CRR model is complete, and there is a unique martingale measure P^*. The martingale measure is characterized by the fact that the random variables $R_1, \dots, R_T$ are independent under P^* with common distribution*

$$P^*[R_t = b] = p^* := \frac{r-a}{b-a}, \quad t = 1, \dots, T.$$

Proof. A measure Q on $(\Omega, \mathcal{F})$ is a martingale measure if and only if the discounted price process is a martingale under Q, i.e.,

$$X_t = E_Q[X_{t+1} \mid \mathcal{F}_t] = X_t\, E_Q\left[\frac{1+R_{t+1}}{1+r} \,\middle|\, \mathcal{F}_t\right] \quad Q\text{-a.s.}$$

for all $t \leq T-1$. This identity is equivalent to the equation

$$r = E_Q[R_{t+1} \mid \mathcal{F}_t] = b \cdot Q[R_{t+1} = b \mid \mathcal{F}_t] + a \cdot \big(1 - Q[R_{t+1} = b \mid \mathcal{F}_t]\big),$$

i.e., to the condition

$$Q[R_{t+1} = b \mid \mathcal{F}_t](\omega) = p^* = \frac{r-a}{b-a} \quad \text{for } Q\text{-a.e. } \omega \in \Omega.$$

But this holds if and only if the random variables $R_1, \dots, R_T$ are independent under Q with common distribution $Q[R_t = b] = p^*$. In particular, there can be at most one martingale measure for X.

If the market model is arbitrage-free, then there exists an *equivalent* martingale measure P^*. The condition $P^* \approx P$ implies

$$p^* = P^*[\, R_1 = b\,] \in (0, 1),$$

which holds if and only if $a < r < b$.

Conversely, if $a < r < b$ then we can define a measure $P^* \approx P$ on $(\Omega, \mathcal{F})$ by

$$P^*[\,\{\omega\}\,] := (p^*)^k \cdot (1 - p^*)^{T-k} > 0$$

where k denotes the number of occurrences of $+1$ in $\omega = (y_1, \dots, y_T)$. Under P^*, $Y_1, \dots, Y_T$, and hence $R_1, \dots, R_T$, are independent random variables with common distribution $P^*[\, Y_t = 1\,] = P^*[\, R_t = b\,] = p^*$, and so P^* is an equivalent martingale measure. □

From now on, we consider only CRR models which are arbitrage-free, and we denote by P^* the unique equivalent martingale measure.

Remark 5.41. Note that the unique martingale measure P^*, and hence the valuation of any contingent claim, is completely independent of the initial choice of the "objective" measure P within the class of measures satisfying (5.23). ◇

Let us now turn to the problem of *pricing* and *hedging* a given contingent claim C. The discounted claim $H = C/S_T^0$ can be written as

$$H = h(S_0, \dots, S_T)$$

for a suitable function h.

Proposition 5.42. *The value process*

$$V_t = E^*[\, H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T,$$

of a replicating strategy for H is of the form

$$V_t(\omega) = v_t\big(S_0, S_1(\omega), \dots, S_t(\omega)\big),$$

where the function v_t is given by

$$v_t(x_0, \dots, x_t) = E^*\bigg[\, h\bigg(x_0, \dots, x_t, x_t\,\frac{S_1}{S_0}, \dots, x_t\,\frac{S_{T-t}}{S_0}\bigg)\bigg]. \tag{5.24}$$

Proof. Clearly,

$$V_t = E^*\bigg[\, h\bigg(S_0, S_1, \dots, S_t, S_t\,\frac{S_{t+1}}{S_t}, \dots, S_t\,\frac{S_T}{S_t}\bigg) \,\bigg|\, \mathcal{F}_t\bigg].$$

Each quotient S_{t+s}/S_t is independent of $\mathcal{F}_t$ and has under P^* the same distribution as

$$\frac{S_s}{S_0} = \prod_{k=1}^{s}(1+R_k).$$

Hence (5.24) follows from the standard properties of conditional expectations. □

Since V is characterized by the recursion

$$V_T := H \quad \text{and} \quad V_t = E^*[\, V_{t+1} \mid \mathcal{F}_t \,], \quad t = T-1, \dots, 0,$$

we obtain a recursive formula for the functions v_t defined in (5.24):

$$\begin{aligned} v_T(x_0, \dots, x_T) &= h(x_0, \dots, x_T), \\ v_t(x_0, \dots, x_t) &= p^* \cdot v_{t+1}(x_0, \dots, x_t, x_t\hat{b}) + (1-p^*) \cdot v_{t+1}(x_0, \dots, x_t, x_t\hat{a}), \end{aligned} \tag{5.25}$$

where

$$\hat{a} := 1 + a \quad \text{and} \quad \hat{b} := 1 + b.$$

Example 5.43. If $H = h(S_T)$ depends only on the terminal value S_T of the stock price, then V_t depends only on the value S_t of the current stock price:

$$V_t(\omega) = v_t\big(S_t(\omega)\big).$$

Moreover, the formula (5.24) for v_t reduces to an expectation with respect to the binomial distribution with parameter p^*:

$$v_t(x_t) = \sum_{k=0}^{T-t} h\big(x_t\, \hat{a}^{T-t-k}\, \hat{b}^k\big) \binom{T-t}{k} (p^*)^k (1-p^*)^{T-t-k}.$$

In particular, the unique arbitrage-free price of H is given by

$$\pi(H) = v_0(S_0) = \sum_{k=0}^{T} h\big(S_0\, \hat{a}^{T-k}\, \hat{b}^k\big) \binom{T}{k} (p^*)^k (1-p^*)^{T-k}.$$

For $h(x) = (x-K)^+/(1+r)^T$ or $h(x) = (K-x)^+/(1+r)^T$, we obtain explicit formulas for the arbitrage-free prices of European call or put options with strike price K. For instance, the price of $H^{\text{call}} := (S_T - K)^+/(1+r)^T$ is given by

$$\pi(H^{\text{call}}) = \frac{1}{(1+r)^T} \sum_{k=0}^{T} \big(S_0\, \hat{a}^{T-k}\, \hat{b}^k - K\big)^+ \binom{T}{k} (p^*)^k (1-p^*)^{T-k}. \qquad \diamond$$

Example 5.44. Denote by

$$M_t := \max_{0 \le s \le t} S_s, \quad 0 \le t \le T,$$

the *running maximum* of S, and consider a discounted claim with payoff $H = h(S_T, M_T)$. For instance, H can be an up-and-in or up-and-out barrier option or a lookback put. Then the value process of H is of the form

$$V_t = v_t(S_t, M_t),$$

where

$$v_t(x_t, m_t) = E^*\left[h\left(x_t \frac{S_{T-t}}{S_0}, m_t \vee \left(x_t \frac{M_{T-t}}{S_0} \right) \right) \right].$$

This follows from (5.24) or directly from the fact that

$$M_T = M_t \vee \left(S_t \max_{t \le u \le T} \frac{S_u}{S_t} \right),$$

where $\max_{t \le u \le T} S_u/S_t$ is independent of $\mathcal{F}_t$ and has the same law as M_{T-t}/S_0 under P^*. The same argument works for options that depend on the minimum of the stock price such as lookback calls or down-and-in barrier options.

For an *Asian option* depending on the average price

$$S_{\mathrm{av}} := \frac{1}{|\mathbb{T}|} \sum_{t \in \mathbb{T}} S_t$$

during a predetermined set of periods $\mathbb{T} \subset \{0, \dots, T\}$, one introduces the process

$$A_t := \sum_{s \in \mathbb{T},\, s \le t} S_s.$$

Using a similar argument as above, one shows that the value process V_t of the Asian option is a function of S_t, A_t, and t. ◇

Let us now derive a formula for the hedging strategy $\overline{\xi} = (\xi^0, \xi)$ of our discounted claim $H = h(S_0, \dots, S_T)$.

Proposition 5.45. *The hedging strategy is given by*

$$\xi_t(\omega) = \Delta_t\big(S_0, S_1(\omega), \dots, S_{t-1}(\omega)\big),$$

where

$$\Delta_t(x_0, \dots, x_{t-1}) := (1+r)^t \frac{v_t(x_0, \dots, x_{t-1}, x_{t-1}\,\hat{b}) - v_t(x_0, \dots, x_{t-1}, x_{t-1}\,\hat{a})}{x_{t-1}\hat{b} - x_{t-1}\hat{a}}.$$

Proof. For each $\omega = (y_1, \dots, y_T)$, ξ_t must satisfy

$$\xi_t(\omega)\big(X_t(\omega) - X_{t-1}(\omega)\big) = V_t(\omega) - V_{t-1}(\omega). \tag{5.26}$$

In this equation, the random variables ξ_t, X_{t-1}, and V_{t-1} depend only on the first $t-1$ components of ω. For a fixed t, let us define ω^+ and ω^- by

$$\omega^\pm := (y_1, \dots, y_{t-1}, \pm 1, y_{t+1}, \dots, y_T).$$

Plugging ω^+ and ω^- into (5.26) shows

$$\xi_t(\omega) \cdot \big(X_{t-1}(\omega)\,\hat{b}(1+r)^{-1} - X_{t-1}(\omega)\big) = V_t(\omega^+) - V_{t-1}(\omega)$$
$$\xi_t(\omega) \cdot \big(X_{t-1}(\omega)\,\hat{a}(1+r)^{-1} - X_{t-1}(\omega)\big) = V_t(\omega^-) - V_{t-1}(\omega).$$

Solving for $\xi_t(\omega)$ and using our formula (5.25) for V_t, we obtain

$$\xi_t(\omega) = (1+r)\frac{V_t(\omega^+) - V_t(\omega^-)}{X_{t-1}(\omega)(\hat{b} - \hat{a})} = \Delta_t\big(S_0, S_1(\omega), \dots, S_{t-1}(\omega)\big). \qquad \square$$

Remark 5.46. The term Δ_t may be viewed as a discrete "derivative" of the value function v_t with respect to the possible stock price changes. In financial language, a hedging strategy based on a derivative of the value process is often called a *Delta hedge*. ◇

Remark 5.47. Let $H = h(S_T)$ be a discounted claim which depends on the terminal value of S by way of an *increasing* function h. For instance, h can be the discounted payoff function $h(x) = (x-K)^+/(1+r)^T$ of a European call option. Then

$$v_t(x) = E^*[\, h(x\, S_{T-t}/S_0)\,]$$

is also increasing in x, and so the hedging strategy satisfies

$$\xi_t(\omega) = (1+r)^t \frac{v_t\big(S_{t-1}(\omega)\,\hat{b}\big) - v_t\big(S_{t-1}(\omega)\,\hat{a}\big)}{S_{t-1}(\omega)\,\hat{b} - S_{t-1}(\omega)\,\hat{a}} \ge 0.$$

In other words, the hedging strategy for H does not involve *short sales* of the risky asset. ◇

5.6 Exotic derivatives

The recursion formula (5.25) can be used for the numeric computation of the value process of any contingent claim. For the value processes of certain *exotic derivatives* which depend on the *maximum* of the stock price, it is even possible to obtain simple closed-form solutions if we make the additional assumption that

$$\hat{a} = \frac{1}{\hat{b}},$$

where $\hat{a} = 1 + a$ and $\hat{b} = 1 + b$. In this case, the price process of the risky asset is of the form

$$S_t(\omega) = S_0 \hat{b}^{Z_t(\omega)}$$

where, for Y_k as in (5.22),

$$Z_0 := 0 \quad \text{and} \quad Z_t := Y_1 + \cdots + Y_t, \quad t = 1, \dots, T.$$

Let $\mathbb{P}$ denote the uniform distribution

$$\mathbb{P}[\,\{\omega\}\,] := \frac{1}{|\Omega|} = 2^{-T}, \quad \omega \in \Omega.$$

Under the measure $\mathbb{P}$, the random variables Y_t are independent with common distribution $\mathbb{P}[\,Y_t = +1\,] = \frac{1}{2}$. Thus, the stochastic process Z becomes a *standard random walk* under $\mathbb{P}$. Therefore,

$$\mathbb{P}[\,Z_t = k\,] = \begin{cases} 2^{-t}\binom{t}{\frac{t+k}{2}} & \text{if } t + k \text{ is even,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.27}$$

The following lemma is the key to numerous explicit results on the distribution of Z under the measure $\mathbb{P}$; see, e.g., Chapter III of [83]. For its statement, it will be convenient to assume that the random walk Z is defined up to time $T + 1$; this can always be achieved by enlarging our probability space $(\Omega, \mathcal{F})$. We denote by

$$M_t := \max_{0 \le s \le t} Z_s$$

the *running maximum* of Z.

Lemma 5.48 (Reflection principle). *For all $k \in \mathbb{N}$ and $l \in \mathbb{N}_0$,*

$$\mathbb{P}\big[\,M_T \ge k \text{ and } Z_T = k - l\,\big] = \mathbb{P}[\,Z_T = k + l\,],$$

and

$$\mathbb{P}\big[\,M_T = k \text{ and } Z_T = k - l\,\big] = 2\frac{k + l + 1}{T + 1}\,\mathbb{P}[\,Z_{T+1} = 1 + k + l\,].$$

Proof. Let

$$\tau(\omega) := \inf\{\,t \ge 0 \mid Z_t(\omega) = k\,\} \wedge T.$$

For $\omega = (y_1, \dots, y_T) \in \Omega$ we define $\phi(\omega)$ by $\phi(\omega) = \omega$ if $\tau(\omega) = T$ and by

$$\phi(\omega) = \big(y_1, \dots, y_{\tau(\omega)}, -y_{\tau(\omega)+1}, \dots, -y_T\big)$$

otherwise, i.e., if the level k is reached before the deadline T. Intuitively, the two trajectories $\big(Z_t(\omega)\big)_{t=0,\dots,T}$ and $\big(Z_t(\phi(\omega)\big)_{t=0,\dots,T}$ coincide up to $\tau(\omega)$, but from then

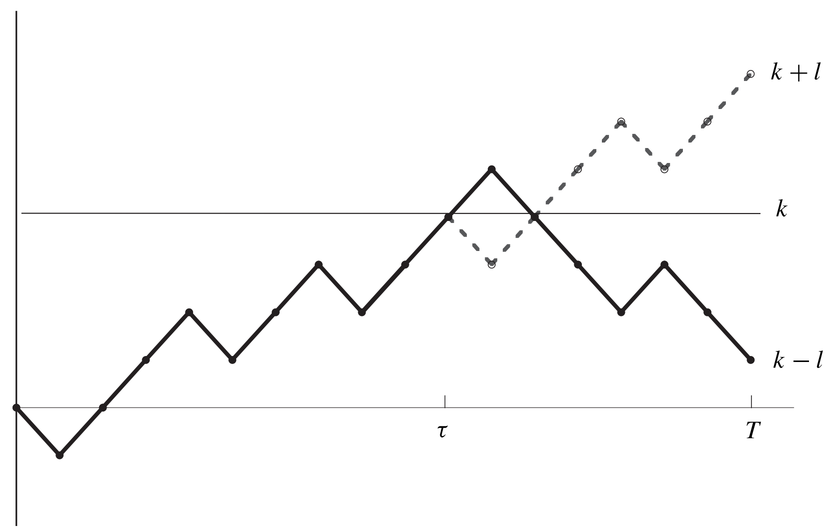


Figure 5.4. The reflection principle.

on the latter path is obtained by reflecting the original one on the horizontal axis at level k; see Figure 5.4.

Let $A_{k,l}$ denote the set of all $\omega \in \Omega$ such that $Z_T(\omega) = k - l$ and $M_T \ge k$. Then ϕ is a bijection from $A_{k,l}$ to the set

$$\{ M_T \ge k \text{ and } Z_T = k + l \},$$

which coincides with $\{Z_T = k + l\}$, due to our assumption $l \ge 0$. Hence, the uniform distribution $\mathbb{P}$ must assign the same probability to $A_{k,l}$ and $\{Z_T = k + l\}$, and we obtain our first formula.

The second formula is trivial in case $T + k + l$ is not even. Otherwise, we let $j := (T + k + l)/2$ and apply (5.27) together with part one of this lemma:

$$\begin{aligned}
&\mathbb{P}\big[M_T = k;\ Z_T = k - l \big] \\
&\quad = \mathbb{P}\big[M_T \ge k;\ Z_T = k - l \big] - \mathbb{P}\big[M_T \ge k + 1;\ Z_T = k - l \big] \\
&\quad = \mathbb{P}[Z_T = k + l] - \mathbb{P}[Z_T = k + l + 2] \\
&\quad = 2^{-T}\binom{T}{j} - 2^{-T}\binom{T}{j+1} \\
&\quad = 2^{-T}\binom{T+1}{j+1}\frac{2j + 1 - T}{T + 1},
\end{aligned}$$

and this expression is equal to the right-hand side of our second formula. □

Formula (5.27) changes as follows if we replace the uniform distribution $\mathbb{P}$ by our martingale measure P^*, described in Theorem 5.40:

$$P^*[\, Z_t = k\,] = \begin{cases} (p^*)^{\frac{t+k}{2}}(1-p^*)^{\frac{t-k}{2}}\binom{t}{\frac{t+k}{2}} & \text{if } t+k \text{ is even,} \\ 0 & \text{otherwise.} \end{cases}$$

Let us now show how the reflection principle carries over to P^*.

Lemma 5.49 (Reflection principle for P^*). *For all $k \in \mathbb{N}$ and $l \in \mathbb{N}_0$,*

$$\begin{aligned} P^*\big[\, M_T \geq k,\ Z_T = k - l\,\big] &= \Big(\frac{1-p^*}{p^*}\Big)^l P^*[\, Z_T = k + l\,] \\ &= \Big(\frac{p^*}{1-p^*}\Big)^k P^*[\, Z_T = -k - l\,], \end{aligned}$$

and

$$\begin{aligned} &P^*\big[\, M_T = k,\ Z_T = k - l\,\big] \\ &\quad = \frac{1}{p^*} \cdot \Big(\frac{1-p^*}{p^*}\Big)^l \cdot \frac{k+l+1}{T+1}\, P^*[\, Z_{T+1} = 1 + k + l\,] \\ &\quad = \frac{1}{1-p^*} \cdot \Big(\frac{p^*}{1-p^*}\Big)^k \cdot \frac{k+l+1}{T+1}\, P^*[\, Z_{T+1} = -1 - k - l\,]. \end{aligned}$$

Proof. We show first that the density of P^* with respect to $\mathbb{P}$ is given by

$$\frac{dP^*}{d\mathbb{P}} = 2^T \cdot (p^*)^{\frac{T+Z_T}{2}}(1-p^*)^{\frac{T-Z_T}{2}}.$$

Indeed, P^* puts the weight

$$P^*[\,\{\omega\}\,] = (p^*)^k(1-p^*)^{T-k}$$

to each $\omega = (y_1, \dots, y_T) \in \Omega$ which contains exactly k components with $y_i = +1$. But for such an ω we have $Z_T(\omega) = k - (T-k) = 2k - T$, and our formula follows.

From the density formula, we get

$$\begin{aligned} &P^*\big[\, M_T \geq k \text{ and } Z_T = k - l\,\big] \\ &\qquad = 2^T (p^*)^{\frac{T+k-l}{2}}(1-p^*)^{\frac{T+l-k}{2}}\, \mathbb{P}\big[\, M_T \geq k \text{ and } Z_T = k - l\,\big]. \end{aligned}$$

Applying the reflection principle and using again the density formula, we see that the probability term on the right is equal to

$$\mathbb{P}[\, Z_T = k + l\,] = 2^{-T}(p^*)^{-\frac{T+k+l}{2}}(1-p^*)^{-\frac{T-k-l}{2}}\, P^*[\, Z_T = k + l\,],$$

which gives the first identity. The proof of the remaining ones is analogous. □

Example 5.50 (Up-and-in call option). Consider an up-and-in call option of the form

$$C^{\text{call}}_{\text{u\&i}} = \begin{cases} (S_T - K)^+ & \text{if } \max\limits_{0 \le t \le T} S_t \ge B, \\ 0 & \text{otherwise,} \end{cases}$$

where $B > S_0 \vee K$ denotes a given barrier, and where $K > 0$ is the strike price. Our aim is to compute the arbitrage-free price

$$\pi(C^{\text{call}}_{\text{u\&i}}) = \frac{1}{(1+r)^T} E^*[\, C^{\text{call}}_{\text{u\&i}}\,].$$

Clearly,

$$\begin{aligned} E^*\big[C^{\text{call}}_{\text{u\&i}}\big] &= E^*\big[(S_T - K)^+;\ \max_{0 \le t \le T} S_t \ge B\big] \\ &= E^*\big[(S_T - K)^+;\ S_T \ge B\big] + E^*\big[(S_T - K)^+;\ \max_{0 \le t \le T} S_t \ge B,\ S_T < B\big]. \end{aligned}$$

The first expectation on the right can be computed explicitly in terms of the binomial distribution. Thus, it remains to compute the second expectation, which we denote by I. To this end, we may assume without loss of generality that B lies within the range of possible asset prices, i.e., there exists some $k \in \mathbb{N}$ such that $B = S_0 \hat{b}^k$. Then, by Lemma 5.49,

$$\begin{aligned} I &= \sum_{l \ge 1} E^*\big[\,(S_T - K)^+;\ M_T \ge k,\ Z_T = k - l\,\big] \\ &= \sum_{l \ge 1} (S_0 \hat{b}^{k-l} - K)^+ P^*[\, M_T \ge k,\ Z_T = k - l\,] \\ &= \sum_{l \ge 1} (S_0 \hat{b}^{k-l} - K)^+ \Big(\frac{p^*}{1 - p^*}\Big)^k P^*[\, Z_T = -k - l\,] \\ &= \Big(\frac{p^*}{1 - p^*}\Big)^k \hat{b}^{2k} \sum_{l \ge 1} (S_0 \hat{b}^{-k-l} - \widetilde{K})^+ P^*[\, Z_T = -k - l\,] \\ &= \Big(\frac{p^*}{1 - p^*}\Big)^k \Big(\frac{B}{S_0}\Big)^2 E^*\big[\,(S_T - \widetilde{K})^+;\ S_T < B\,\big], \end{aligned}$$

where

$$\widetilde{K} = K \hat{b}^{-2k} = K \Big(\frac{S_0}{B}\Big)^2.$$

Hence, we obtain the formula

$$\begin{aligned} \pi(C^{\text{call}}_{\text{u\&i}}) = \frac{1}{(1+r)^T} \bigg(& E^*\big[\,(S_T - K)^+;\ S_T \ge B\,\big] \\ & + \Big(\frac{p^*}{1 - p^*}\Big)^k \Big(\frac{B}{S_0}\Big)^2 E^*\big[\,(S_T - \widetilde{K})^+;\ S_T < B\,\big] \bigg). \end{aligned}$$

Both expectations on the right now only involve the binomial distribution with parameters p^* and T. They can be computed as in Example 5.43, and so we get the explicit formula

$$\pi(C_{\text{u\&i}}^{\text{call}}) = \frac{1}{(1+r)^T}\Bigg[\sum_{n=0}^{n_k}(S_0\hat{b}^{T-2n}-K)^+(p^*)^{T-n}(1-p^*)^n\binom{T}{T-n} + \Big(\frac{p^*}{1-p^*}\Big)^k\Big(\frac{B}{S_0}\Big)^2\sum_{n=n_k+1}^{T}(S_0\hat{b}^{T-2n}-\widetilde{K})^+(p^*)^{T-n}(1-p^*)^n\binom{T}{T-n}\Bigg],$$

where n_k is the largest integer n such that $T-2n \geq k$. ◇

Example 5.51 (Up-and-out call option). Consider an up-and-out call option of the form

$$C_{\text{u\&o}}^{\text{call}} = \begin{cases} 0 & \text{if } \max\limits_{0\leq t\leq T} S_t \geq B, \\ (S_T-K)^+ & \text{otherwise,} \end{cases}$$

where $K > 0$ is the strike price and $B > S_0 \vee K$ is an upper barrier for the stock price. As in the preceding example, we assume that $B = S_0(1+b)^k$ for some $k \in \mathbb{N}$. Let

$$C^{\text{call}} := (S_T-K)^+$$

denote the corresponding "plain vanilla call", whose arbitrage-free price is given by

$$\pi\big(C^{\text{call}}\big) = \frac{1}{(1+r)^T}E^*[\,(S_T-K)^+\,].$$

Since $C^{\text{call}} = C_{\text{u\&o}}^{\text{call}} + C_{\text{u\&i}}^{\text{call}}$, we get from Example 5.50 that

$$\begin{aligned}\pi\big(C_{\text{u\&o}}^{\text{call}}\big) &= \pi\big(C^{\text{call}}\big) - \pi\big(C_{\text{u\&i}}^{\text{call}}\big)\\ &= \frac{1}{(1+r)^T}\Bigg[E^*\big[\,(S_T-K)^+;\ S_T<B\,\big]\\ &\qquad - \Big(\frac{p^*}{1-p^*}\Big)^k\Big(\frac{B}{S_0}\Big)^2E^*\big[\,(S_T-\widetilde{K})^+;\ S_T<B\,\big]\Bigg].\end{aligned}$$

where $\widetilde{K} = KS_0^2/B^2$. These expectations can be computed as in Example 5.50. ◇

Similarly, one obtains pricing formulas for barrier options with a lower stock price barrier such as down-and-out put options or down-and-in calls. In the following example, we compute the price of a lookback put option. Lookback call options are handled in the same manner.

Example 5.52 (Lookback put option). A lookback put option corresponds to the contingent claim

$$C_{\max}^{\text{put}} := \max_{0\leq t\leq T} S_t - S_T\,;$$

see Example 5.24. In the CRR model, the discounted arbitrage-free price of $C_{\max}^{\text{put}}$ is given by

$$\pi\big(C_{\max}^{\text{put}}\big) = \frac{1}{(1+r)^T}\,E^*\big[\max_{0\le t\le T} S_t\big] - S_0.$$

The expectation of the maximum can be computed as

$$E^*\big[\max_{0\le t\le T} S_t\big] = S_0 \sum_{k=0}^{T} \hat{b}^k\, P^*\big[M_T = k\big]$$

Lemma 5.49 yields

$$\begin{aligned} P^*[\,M_T = k\,] &= \sum_{l\ge 0} P^*[\,M_T = k,\ Z_T = k-l\,] \\ &= \sum_{l\ge 0} \frac{1}{1-p^*}\Big(\frac{p^*}{1-p^*}\Big)^k \frac{k+l+1}{T+1} P^*[\,Z_{T+1} = -1-k-l\,] \\ &= \frac{1}{1-p^*}\Big(\frac{p^*}{1-p^*}\Big)^k \frac{1}{T+1} E^*[\,-Z_{T+1};\ Z_{T+1} \le -1-k\,]. \end{aligned}$$

Thus, we arrive at the formula

$$\begin{aligned} \pi\big(C_{\max}^{\text{put}}\big) + S_0 = {} & \frac{S_0}{(1+r)^T(1-p^*)(T+1)}\,\cdot \\ & \cdot \sum_{k=0}^{T} \hat{b}^k \Big(\frac{p^*}{1-p^*}\Big)^k E^*[\,-Z_{T+1};\ Z_{T+1} \le -1-k\,]. \end{aligned}$$

As before, one can give explicit formulas for the expectations occurring on the right. ◇

5.7 Convergence to the Black–Scholes price

In practice, a huge number of trading periods may occur between the current time $t = 0$ and the maturity T of a European contingent claim. Thus, the computation of option prices in terms of some martingale measure may become rather elaborate. On the other hand, one can hope that the pricing formulas in discrete time converge to a transparent limit as the number of intermediate trading periods grows larger and larger. In this section, we will formulate conditions under which such a convergence occurs.

Throughout this section, T will *not* denote the *number* of trading periods in a fixed discrete-time market model but rather a physical *date*. The time interval $[0, T]$ will be divided into N equidistant time steps $\frac{T}{N}, \frac{2T}{N}, \dots, \frac{NT}{N}$, and the date $\frac{kT}{N}$ will correspond to the k^{th} trading period of an arbitrage-free market model. For simplicity, we will assume that each market model contains a riskless bond and just one risky asset. In the N^{th} approximation, the risky asset will be denoted by $S^{(N)}$, and the riskless bond will be defined by a constant interest rate $r_N > -1$.

The question is whether the prices of contingent claims in the approximating market models converge as N tends to infinity. Since the terminal values of the riskless bonds should converge, we assume that

$$\lim_{N\uparrow\infty}(1+r_N)^N = e^{rT},$$

where r is a finite constant. This condition is in fact equivalent to the following one:

$$\lim_{N\uparrow\infty} N\, r_N = r\,T.$$

Let us now consider the risky assets. We assume that the initial prices $S_0^{(N)}$ do not depend on N, i.e., $S_0^{(N)} = S_0$ for some constant $S_0 > 0$. The prices $S_k^{(N)}$ are random variables on some probability space $(\Omega_N, \mathcal{F}^{(N)}, P_N^*)$, where P_N^* is a risk-neutral measure for each approximating market model, i.e., the discounted price process

$$X_k^{(N)} := \frac{S_k^{(N)}}{(1+r_N)^k}, \quad k = 0, \dots, N,$$

is a P_N^*-martingale with respect to the filtration $\mathcal{F}_k^{(N)} := \sigma(S_1^{(N)}, \dots, S_k^{(N)})$. Our remaining conditions will be stated in terms of the returns

$$R_k^{(N)} := \frac{S_k^{(N)} - S_{k-1}^{(N)}}{S_{k-1}^{(N)}}, \quad k = 1, \dots, N.$$

First, we assume that, for each N, the random variables $R_1^{(N)}, \dots, R_N^{(N)}$ are independent under P_N^* and satisfy

$$-1 < \alpha_N \le R_k^{(N)} \le \beta_N, \quad k = 1, \dots, N,$$

for constants α_N and β_N such that

$$\lim_{N\uparrow\infty}\alpha_N = \lim_{N\uparrow\infty}\beta_N = 0.$$

Second, we assume that the variances $\operatorname{var}_N\left(R_k^{(N)}\right)$ under P_N^* are such that

$$\sigma_N^2 := \frac{1}{T}\sum_{k=1}^N \operatorname{var}_N\left(R_k^{(N)}\right) \longrightarrow \sigma^2 \in (0,\infty).$$

The following result can be regarded as a multiplicative version of the *central limit theorem*.

Theorem 5.53. *Under the above assumptions, the distributions of $S_N^{(N)}$ under P_N^* converge weakly to the log-normal distribution with parameters* $\log S_0 + rT - \frac{1}{2}\sigma^2 T$ *and $\sigma\sqrt{T}$, i.e., to the distribution of*

$$S_T := S_0 \exp\left(\sigma\, W_T + \left(r - \frac{1}{2}\sigma^2\right)T\right), \tag{5.28}$$

where W_T has a centered normal law $N(0, T)$ with variance T.

Proof. We may assume without loss of generality that $S_0 = 1$. Consider the Taylor expansion

$$\log(1 + x) = x - \frac{1}{2}x^2 + \rho(x)\,x^2 \tag{5.29}$$

where the remainder term ρ is such that

$$|\rho(x)| \le \delta(\alpha, \beta) \quad \text{for } -1 < \alpha \le x \le \beta,$$

and where $\delta(\alpha, \beta) \to 0$ for $\alpha, \beta \to 0$. Applied to

$$S_N^{(N)} = \prod_{k=1}^{N} \big(1 + R_k^{(N)}\big),$$

this yields

$$\log S_N^{(N)} = \sum_{k=1}^{N} \Big(R_k^{(N)} - \frac{1}{2}\big(R_k^{(N)}\big)^2\Big) + \Delta_N,$$

where

$$|\Delta_N| \le \delta(\alpha_N, \beta_N) \sum_{k=1}^{N} \big(R_k^{(N)}\big)^2.$$

Since P_N^* is a martingale measure, we have $E_N^*[\,R_k^{(N)}\,] = r_N$, and it follows that

$$E_N^*[\,|\Delta_N|\,] \le \delta(\alpha_N, \beta_N) \sum_{k=1}^{N} \Big(\operatorname{var}_N\big(R_k^{(N)}\big) + r_N^2\Big) \longrightarrow 0.$$

In particular, $\Delta_N \to 0$ in probability, and the corresponding laws converge weakly to the Dirac measure δ_0. Slutsky's theorem, as stated in Appendix A.6, asserts that it suffices to show that the distributions of

$$Z_N := \sum_{k=1}^{N} \Big(R_k^{(N)} - \frac{1}{2}\big(R_k^{(N)}\big)^2\Big) =: \sum_{k=1}^{N} Y_k^{(N)}$$

converge weakly to the normal law $N(rT - \frac{1}{2}\sigma^2 T, \sigma^2 T)$. To this end, we will check that the conditions of the central limit theorem in the form of Theorem A.36 are satisfied.

Note that

$$\max_{1\le k\le N}\left|Y_k^{(N)}\right| \le \gamma_N + \frac{1}{2}\gamma_N^2 \longrightarrow 0$$

for $\gamma_N := |\alpha_N| \vee |\beta_N|$, and that

$$E_N^*[\,Z_N\,] = Nr_N - \frac{1}{2}\big(\sigma_N^2\,T + Nr_N^2\big) \longrightarrow r\,T - \frac{1}{2}\,\sigma^2\,T.$$

Finally,

$$\operatorname{var}_N(Z_N) \longrightarrow \sigma^2\,T,$$

since for $p > 2$

$$\sum_{k=1}^N E_N^*\big[\big|R_k^{(N)}\big|^p\big] \le \gamma_N^{p-2}\sum_{k=1}^N E_N^*\big[\big(R_k^{(N)}\big)^2\big] \longrightarrow 0.$$

Thus, the conditions of Theorem A.36 are satisfied. □

Remark 5.54. The assumption of independent returns in Theorem 5.53 can be relaxed. Instead of Theorem A.36, we can apply a *central limit theorem for martingales* under suitable assumptions on the behavior of the *conditional* variances

$$\operatorname{var}_N(\,R_k^{(n)}\mid \mathcal{F}_{k-1}\,)\,;$$

for details see, e.g., Section 9.3 of [39]. ◇

Example 5.55. Suppose the approximating model in the N^{th} stage is a CRR model with interest rate

$$r_N = \frac{r\,T}{N},$$

and with returns $R_k^{(N)}$, which can take the two possible values a_N and b_N; see Section 5.5. We assume that

$$\hat{a}_N = 1 + a_N = e^{-\sigma\sqrt{T/N}} \quad \text{and} \quad \hat{b}_N = 1 + b_N = e^{\sigma\sqrt{T/N}}$$

for some given $\sigma > 0$. Since

$$\sqrt{N}r_N \longrightarrow 0, \quad \sqrt{N}a_N \longrightarrow -\sigma\sqrt{T}, \quad \sqrt{N}b_N \longrightarrow \sigma\sqrt{T} \quad \text{as } N\uparrow\infty, \tag{5.30}$$

we have $a_N < r_N < b_N$ for large enough N. Theorem 5.40 yields that the N^{th} model is arbitrage-free and admits a unique equivalent martingale measure P_N^*. The measure P_N^* is characterized by

$$P_N^*[\,R_k^{(N)} = b_N\,] =: p_N^* = \frac{r_N - a_N}{b_N - a_N},$$

and we obtain from (5.30) that

$$\lim_{N\uparrow\infty} p_N^* = \frac{1}{2}.$$

Moreover, $E_N^*[\,R_k^{(N)}\,] = r_N$, and we get

$$\sum_{k=1}^{N} \operatorname{var}_N\left(R_k^{(N)}\right) = N\big(p_N^* b_N^2 + (1-p_N^*)a_N^2 - r_N^2\big) \longrightarrow \sigma^2 T$$

as $N \uparrow \infty$. Hence, the assumptions of Theorem 5.53 are satisfied. ◇

Let us consider a derivative which is defined in terms of a function $f \geq 0$ of the risky asset's terminal value. In each approximating model, this corresponds to a contingent claim

$$C^{(N)} = f\big(S_N^{(N)}\big).$$

Corollary 5.56. *If f is bounded and continuous, the arbitrage-free prices of $C^{(N)}$ calculated under P_N^* converge to a discounted expectation with respect to a log-normal distribution, which is often called the Black–Scholes price. More precisely,*

$$\begin{aligned}\lim_{N\uparrow\infty} E_N^*\left[\frac{C^{(N)}}{(1+r_N)^N}\right] &= e^{-rT} E^*[\, f(S_T)] \\ &= \frac{e^{-rT}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\big(S_0 e^{\sigma\sqrt{T}y + rT - \sigma^2 T/2}\big) e^{-y^2/2}\, dy,\end{aligned} \tag{5.31}$$

where S_T has the form (5.28) *under P^*.*

This convergence result applies in particular to the choice $f(x) = (K - x)^+$ corresponding to a European put option with strike K. Since the put-call parity

$$E_N^*\left[\frac{(S_N^{(N)} - K)^+}{(1+r_N)^N}\right] = E_N^*\left[\frac{(K - S_N^{(N)})^+}{(1+r_N)^N}\right] + S_0 - \frac{K}{(1+r_N)^N}$$

holds for each N, the convergence (5.31) is also true for a European call option with the unbounded payoff profile $f(x) = (x - K)^+$.

Example 5.57 (Black–Scholes formula for the price of a call option). The limit of the arbitrage-free prices of $C^{(N)} = (S_N^{(N)} - K)^+$ is given by $v(S_0, T)$, where

$$v(x, T) = e^{-rT} \int_{-\infty}^{\infty} \big(x e^{\sigma\sqrt{T}y + rT - \sigma^2 T/2} - K\big)^+ e^{-y^2/2}\, dy.$$

The integrand on the right vanishes for

$$y \leq -\frac{\log\frac{x}{K} + \big(r - \frac{1}{2}\sigma^2\big)T}{\sigma\sqrt{T}} =: -d_-(x, T) =: -d_-.$$

Let us also define

$$d_+ := d_+(x, T) := d_-(x, T) + \sigma\sqrt{T} = \frac{\log \frac{x}{K} + \big(r + \frac{1}{2}\sigma^2\big)T}{\sigma\sqrt{T}},$$

and let us denote by $\Phi(z) = (2\pi)^{-1}\int_{-\infty}^{x} e^{-y^2/2}\,dy$ the distribution function of the standard normal distribution. Then

$$v(x, T) = \frac{x}{\sqrt{2\pi}}\int_{-d_-}^{+\infty} e^{-(y-\sigma\sqrt{T})^2/2}\,dy - e^{-rT}K\big(1 - \Phi(-d_-)\big),$$

and we arrive at the *Black–Scholes formula* for the price of a European call option with strike K and maturity T:

$$v(x, T) = x\,\Phi(d_+(x, T)) - e^{-rT}K\,\Phi(d_-(x, T)). \tag{5.32}$$

See Figure 5.5 for the plot of the function $v(x, t)$. ◇

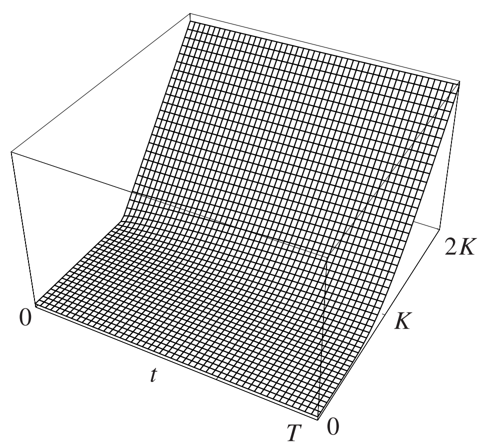


Figure 5.5. The Black–Scholes price $v(x, t)$ of a European call option $(S_T - K)^+$ plotted as a function of the initial spot price $x = S_0$ and the time to maturity t.

Remark 5.58. For fixed x and T, the Black–Scholes price of a European call option increases to the upper arbitrage bound x as $\sigma \uparrow \infty$. In the limit $\sigma \downarrow 0$, we obtain the lower arbitrage bound $(x - e^{-rT}K)^+$; see Remark 1.36. ◇

The following proposition gives a criterion for the convergence (5.31) in case f is not necessarily bounded and continuous. It applies in particular to $f(x) = (x - K)^+$, and so we get an alternative proof for the convergence of call option prices to the Black–Scholes price.

Proposition 5.59. *Let $f:(0,\infty)\to\mathbb{R}$ be measurable, continuous a.e., and such that $|f(x)|\le c\,(1+x)^q$ for some $c\ge 0$ and $0\le q<2$. Then*

$$E_N^*\big[\,f\big(S_N^{(N)}\big)\,\big]\longrightarrow E^*\big[\,f(S_T)\big],$$

where S_T has the form (5.28) *under P^*.*

Proof. Let us note first that by the Taylor expansion (5.29)

$$\begin{aligned}\log E_N^*\big[\,\big(S_N^{(N)}\big)^2\,\big]&=\log\prod_{k=1}^N\big(\operatorname{var}_N\big(1+R_k^{(N)}\big)+E_N^*\big[\,1+R_k^{(N)}\big]^2\big)\\&=\sum_{k=1}^N\log\big(\operatorname{var}_N\big(R_k^{(N)}\big)+(1+r_N)^2\big)\\&\le\sigma_N^2T+2N\,r_N+N\,r_N^2+\widetilde{c}\sum_{k=1}^N\big(\operatorname{var}_N\big(R_k^{(N)}\big)+2|r_N|+r_N^2\big)^2\end{aligned}$$

for a finite constant $\widetilde{c}$. Thus,

$$\sup_N E_N^*\big[\,\big(S_N^{(N)}\big)^2\,\big]<\infty.$$

With this property established, the assertion follows immediately from Theorem 5.53 and the Corollaries A.45 and A.46, but we also give the following more elementary proof. To this end, we may assume that $q>0$, and we define $p:=2/q>1$. Then

$$\sup_N E_N^*\big[\,\big|f\big(S_N^{(N)}\big)\big|^p\,\big]\le c^p\,\sup_N E_N^*\big[\,\big(1+S_N^{(N)}\big)^2\,\big]<\infty,$$

and the assertion follows from Lemma 5.60 below. □

Lemma 5.60. *Suppose $(\mu_N)_{N\in\mathbb{N}}$ is a sequence of probability measures on $\mathbb{R}$ converging weakly to μ. If f is a measurable and μ-a.e. continuous function on $\mathbb{R}$ such that*

$$c:=\sup_{N\in\mathbb{N}}\int|f|^p\,d\mu_N<\infty\quad\textit{for some } p>1,$$

then

$$\int f\,d\mu_N\longrightarrow\int f\,d\mu.$$

Proof. We may assume without loss of generality that $f\ge 0$. Then $f_k:=f\wedge k$ is a bounded and μ-a.e. continuous function for each $k>0$. Clearly,

$$\int f\,d\mu_N=\int f_k\,d\mu_N+\int(f-k)^+\,d\mu_N.$$

Due to part (e) of the portmanteau theorem in the form of Theorem A.38, the first integral on the right converges to $\int f_k \, d\mu$ as $N \uparrow \infty$. Let us consider the second term on the right:

$$\int (f-k)^+ \, d\mu_N \le \int_{\{f>k\}} f \, d\mu_N \le \frac{1}{k^{p-1}} \int f^{p-1} f \, d\mu_N \le \frac{c}{k^{p-1}},$$

uniformly in N. Hence,

$$\begin{aligned}\int f_k \, d\mu &= \lim_{N\uparrow\infty} \int f_k \, d\mu_N \le \liminf_{N\uparrow\infty} \int f \, d\mu_N \\ &\le \limsup_{N\uparrow\infty} \int f \, d\mu_N \le \int f_k \, d\mu + \frac{c}{k^{p-1}}.\end{aligned}$$

Letting $k \uparrow \infty$, we have $\int f_k \, d\mu \nearrow \int f \, d\mu$, and convergence follows. □

Let us now continue the discussion of the Black–Scholes price of a European call option where $f(x) = (x-K)^+$. We are particularly interested how it depends on the various model parameters. The dependence on the spot price $S_0 = x$ can be analyzed via the x-derivatives of the function $v(t,x)$ appearing in the Black–Scholes formula (5.32). The first derivative

$$\Delta(x,t) := \frac{\partial}{\partial x} v(x,t) = \Phi\big(d_+(x,t)\big)$$

is called the option's *Delta*; see Figure 5.6. In analogy to the formula for the hedging strategy in the binomial model obtained in Proposition (5.45), $\Delta(x,t)$ determines the "Delta hedging portfolio" needed for a replication of the call option in continuous time, as explained in (5.37) below.

The *Gamma* of the call option is given by

$$\Gamma(x,t) := \frac{\partial}{\partial x} \Delta(x,t) = \frac{\partial^2}{\partial x^2} v(x,t) = \varphi\big(d_+(x,t)\big) \frac{1}{x\sigma\sqrt{t}};$$

see Figure 5.7. Here $\varphi(x) = \Phi'(x) = e^{-x^2/2}/\sqrt{2\pi}$ stands as usual for the density of the standard normal distribution. Large Gamma values occur in regions where the Delta changes rapidly, corresponding to the need for frequent readjustments of the Delta hedging portfolio. Note that Γ is always strictly positive. It follows that $v(x,t)$ is a strictly convex function of its first argument.

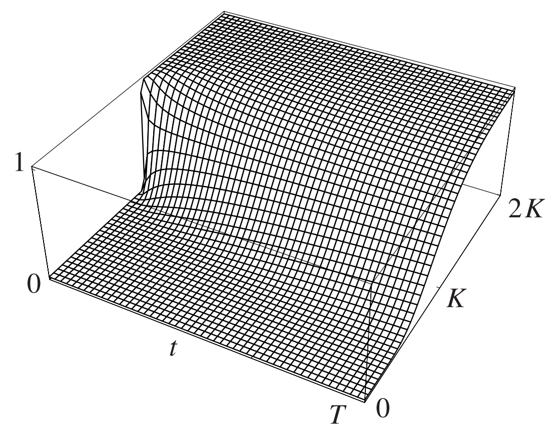


Figure 5.6. The Delta $\Delta(x, t)$ of the Black–Scholes price of a European call option.

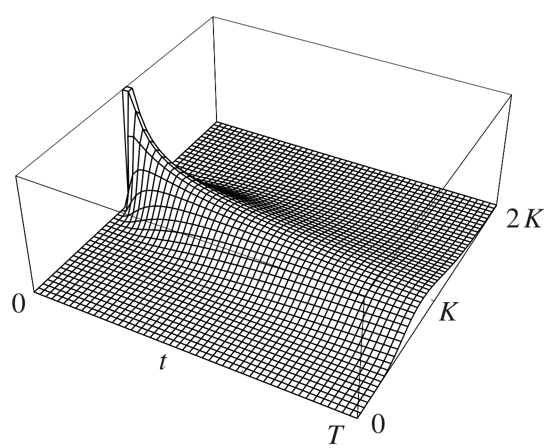


Figure 5.7. The option's Gamma $\Gamma(x, t)$.

Remark 5.61. On the one hand, $0 \leq \Delta(x, t) \leq 1$ implies that

$$|v(x, t) - v(y, t)| \leq |x - y|.$$

Thus, the *total change* of the option values is always less than a corresponding change in the asset prices. On the other hand, the strict convexity of $x \mapsto v(x, t)$ together with (A.1) yields that for $t > 0$ and $z > y$

$$\frac{v(z, t) - v(y, t)}{z - y} > \frac{v(y, t) - v(0, t)}{y - 0} = \frac{v(y, t)}{y}$$

and hence

$$\frac{v(z,t)-v(y,t)}{v(y,t)} > \frac{z-y}{y}.$$

Similarly, one obtains

$$\frac{v(x,t)-v(y,t)}{v(y,t)} < \frac{x-y}{y}$$

for $x < y$. Thus, the *relative change* of option prices is larger in absolute value than the relative change of asset values. This fact can be interpreted as the *leverage effect* for call options; see also Example 1.41. ◇

Another important parameter is the *Theta*

$$\Theta(x,t) := \frac{\partial}{\partial t} v(x,t) = \frac{x\sigma}{2\sqrt{t}}\varphi\big(d_+(x,t)\big) + Kr\,e^{-rt}\,\Phi\big(d_-(x,t)\big)\,;$$

see Figure 5.8. The fact $\Theta > 0$ corresponds to our general observation, made in Example 5.36, that arbitrage-free prices of European call options are typically increasing functions of the maturity.

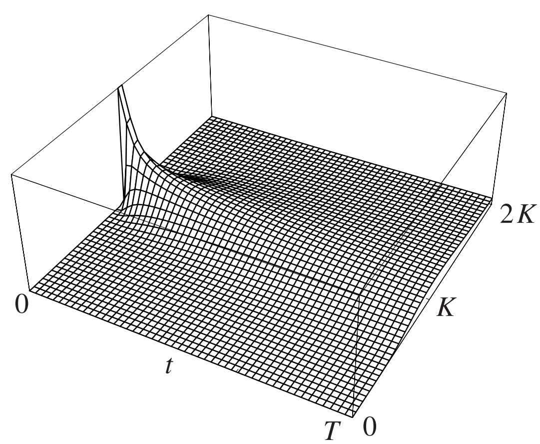


Figure 5.8. The Theta $\Theta(x,t)$.

Note that, for $t > 0$, the parameters Δ, Γ, and Θ are related by the equation

$$\Theta(x,t) = rx\,\Delta(x,t) + \frac{1}{2}\sigma^2x^2\,\Gamma(x,t) - r\,v(x,t).$$

Thus, for $(x,t) \in (0,\infty) \times (0,\infty)$, the function v solves the partial differential equation

$$\frac{\partial v}{\partial t} = rx\,\frac{\partial v}{\partial x} + \frac{1}{2}\sigma^2x^2\,\frac{\partial^2 v}{\partial x^2} - rv, \tag{5.33}$$

often called the *Black–Scholes equation*. Since

$$v(x,t) \longrightarrow f(x) = (x-K)^+ \quad \text{as } t \downarrow 0, \tag{5.34}$$

$v(x,t)$ is a solution of the Cauchy problem defined via (5.33) and (5.34). This fact is not limited to call options, it remains valid for all reasonable payoff profiles f.

Proposition 5.62. *Let f be a continuous function on $(0,\infty)$ such that $|f(x)| \le c(1+x)^p$ for some $c,\ p \ge 0$, and define*

$$u(x,t) := e^{-rt} E^*[\, f(S_t)\,] = \frac{e^{-rt}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\big(xe^{\sigma\sqrt{t}y + rt - \sigma^2 t/2}\big) e^{-y^2/2}\, dy,$$

where $S_t = x\exp(\sigma W_t + rt - \sigma^2 t/2)$ and W_t has law $N(0,t)$ under P^*. Then u solves the Cauchy problem defined by the Black–Scholes equation (5.33) *and the initial condition* $\lim_{t\downarrow 0} u(x,t) = f(x)$*, locally uniformly in x.*

Proof. By using the formula (2.20) for the density of a log-normally distributed random variable, we obtain

$$E^*[\, f(S_t)\,] = \int_0^\infty \frac{1}{y\sigma\sqrt{t}} \varphi\Big(\frac{\log y - rt + \sigma^2 t/2 - \log x}{\sigma\sqrt{t}}\Big) f(y)\, dy,$$

where $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$. The validity of (5.33) can now be checked by differentiating under the integral. The initial condition is verified via Lebesgue's theorem. □

Recall that the Black–Scholes price $v(S_0, T)$ was obtained as the expectation of the *discounted* payoff $e^{-rT}(S_T - K)^+$ under the measure P^*. Thus, at a first glance, it may come as a surprise that the *Rho* of the option,

$$\varrho(x,t) := \frac{\partial}{\partial r} v(x,t) = Kt\, e^{-rt}\, \Phi\big(d_-(x,t)\big),$$

is strictly positive, i.e., the price is *increasing* in r; see Figure 5.9. Note, however, that the measure P^* depends itself on the interest rate r, since $E^*[\, e^{-rT} S_T\,] = S_0$. In a simple one-period model, we have already seen this effect in Example 1.41.

The parameter σ is called the *volatility*. As we have seen, the Black–Scholes price of a European call option is an increasing function of the volatility, and this is reflected in the strict positivity of

$$\mathcal{V}(x,t) := \frac{\partial}{\partial \sigma} v(x,t) = x\sqrt{t}\, \varphi\big(d_+(x,t)\big); \tag{5.35}$$

see Figure 5.10. The function $\mathcal{V}$ is often called the *Vega* of the call option price, and the functions Δ, Γ, Θ, ϱ, and $\mathcal{V}$ are usually called the *Greeks* (although "vega" does not correspond to a letter of the Greek alphabet).

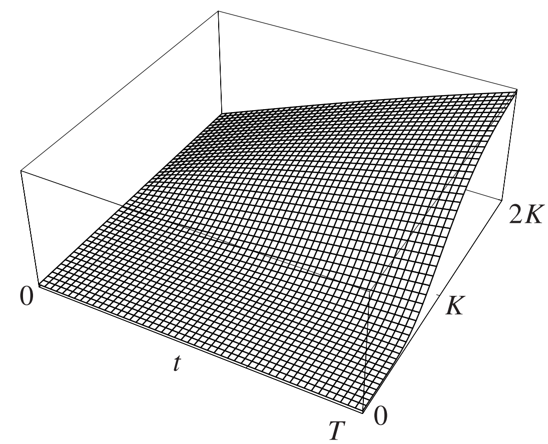


Figure 5.9. The Rho $\varrho(x,t)$ of a call option.

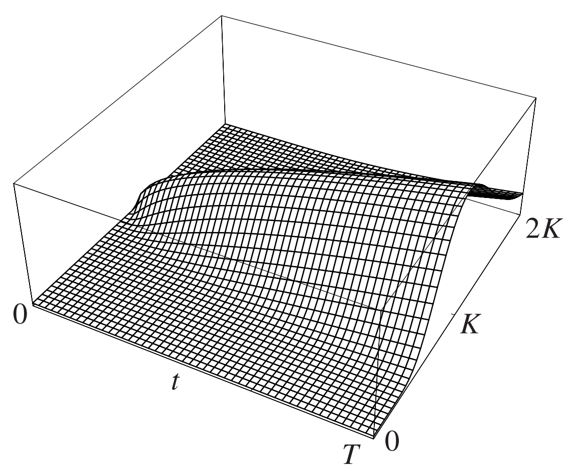


Figure 5.10. The Vega $\mathcal{V}(x,t)$.

Let us conclude this section with some informal comments on the dynamic picture behind the convergence result in Theorem 5.53 and the pricing formulas in Example 5.57 and Proposition 5.59. The constant r is viewed as the interest rate of a riskfree savings account

$$S_t^0 = e^{rt}, \quad 0 \le t \le T.$$

The prices of the risky asset in each discrete-time model are considered as a *continuous* process $\widetilde{S}^{(N)} = \big(\widetilde{S}_t^{(N)}\big)_{0\le t\le T}$, defined as $\widetilde{S}_t^{(N)} := S_k^{(N)}$ at the dates $t = \frac{kT}{N}$, and by linear interpolation in between. Theorem 5.53 shows that the distributions of $\widetilde{S}_t^{(N)}$

converge for each fixed t weakly to the distribution of

$$S_t = S_0 \exp\Big(\sigma\, W_t + \Big(r - \frac{1}{2}\sigma^2\Big)t\Big), \tag{5.36}$$

where W_t has a centered normal distribution with variance t. In fact, one can prove convergence in the much stronger sense of a functional central limit theorem: The laws of the processes $\widetilde{S}^{(N)}$, considered as $C[0, T]$-valued random variables on $(\Omega_N, \mathcal{F}^{(N)}, P_N^*)$, converge weakly to the law of a *geometric Brownian motion* $S = (S_t)_{0\le t\le T}$, where each S_t is of the form (5.36), and where the process $W = (W_t)_{0\le t\le T}$ is a standard *Brownian motion* or *Wiener process*. A Wiener process is characterized by the following properties:

- $W_0 = 0$ almost surely,
- $t \mapsto W_t$ is continuous,
- for each sequence $0 = t_0 < t_1 < \cdots < t_n = T$, the increments

  $$W_{t_1} - W_{t_0}, \dots, W_{t_n} - W_{t_{n-1}}$$

  are independent and have normal distributions $N(0, t_i - t_{i-1})$;

see, e.g., [129]. This multiplicative version of a functional central limit theorem follows as above if we replace the classical central limit theorem by Donsker's invariance principle; for details see, e.g., [73].

Geometric Brownian motion is the classical reference model in continuous-time mathematical finance. In order to describe the model more explicitly, we denote by $W = (W_t)_{0\le t\le T}$ the coordinate process on the canonical path space $\Omega = C[0, T]$, defined by $W_t(\omega) = \omega(t)$, and furthermore by $(\mathcal{F}_t)_{0\le t\le T}$ the filtration given by $\mathcal{F}_t = \sigma(W_s;\ s \le t)$. There is exactly one probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F}_T)$ such that W is a Wiener process under $\mathbb{P}$, and it is called the *Wiener measure*. Let us now model the price process of a risky asset as a geometric Brownian motion S defined by (5.36). The discounted price process

$$X_t := \frac{S_t}{e^{rt}} = S_0 e^{\sigma W_t - \sigma^2 t/2}, \quad 0 \le t \le T,$$

is a martingale under $\mathbb{P}$, since

$$\mathbb{E}[\, X_t \mid \mathcal{F}_s\,] = X_s \mathbb{E}\big[\, e^{\sigma(W_t - W_s) - \sigma^2(t-s)/2}\,\big] = X_s$$

for $0 \le s \le t \le T$. In fact, $\mathbb{P}$ is the only probability measure equivalent to $\mathbb{P}$ with that property.

As in discrete time, uniqueness of the equivalent martingale measure implies completeness of the model. Let us sketch the construction of the replicating strategy for a given European option with reasonable payoff profile $f(S_T)$, for example a call option with strike K. At time t the price of the asset is $S_t(\omega)$, the remaining time to maturity is $T-t$, and the discounted price of the option is given by

$$V_t(\omega) = e^{-rt}u(S_t(\omega), T-t),$$

where u is the function defined in Proposition 5.62. The process $V = (V_t)_{0\le t\le T}$ can be viewed as the value process of the trading strategy $\overline{\xi} = (\xi^0, \xi)$ defined by

$$\xi_t = \Delta(S_t, T-t), \quad \xi_t^0 = e^{-rt}u(S_t, T-t) - \xi_t X_t, \tag{5.37}$$

where $\Delta = \partial u/\partial x$ is the option's Delta. Indeed, if we view ξ as the number of shares in the risky asset S and ξ^0 as the number of shares in the riskfree savings account $S_t^0 = e^{rt}$, then the value of the resulting portfolio in units of the numéraire is given by

$$V_t = \xi_t \cdot X_t + \xi_t^0 = e^{-rt}(\xi_t \cdot S_t + \xi_t^0 \cdot S_t^0).$$

The strategy replicates the option since

$$V_T := \lim_{t\uparrow T} e^{-rt}u(S_t, T-t) = e^{-rT} f(S_T) = \frac{f(S_T)}{S_T^0},$$

due to Proposition 5.62. Moreover, its initial cost is given by the Black–Scholes price

$$V_0 = u(S_0, T) = e^{-rT}\mathbb{E}[\, f(S_T)\,] = \frac{e^{-rT}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\big(xe^{\sigma\sqrt{T}y + rT - \sigma^2 T/2}\big)e^{-y^2/2}\, dy.$$

It remains to show that the strategy is *self-financing* in the sense that changes in the portfolio value are only due to price changes in the underlying assets and do not require any additional capital. To this end, we use *Itô's formula*

$$dF(W_t, t) = \frac{\partial F}{\partial x}(W_t, t)\, dW_t + \Big(\frac{1}{2}\frac{\partial^2 F}{\partial x^2} + \frac{\partial F}{\partial t}\Big)(W_t, t)\, dt$$

for a smooth function F, see, e.g., [129] or, for a strictly pathwise approach, [87]. Applied to the function $F(x,t) = \exp(\sigma x + rt - \sigma^2 t/2)$, it shows that the price process S satisfies the *stochastic differential equation*

$$dS_t = \sigma S_t\, dW_t + rS_t\, dt. \tag{5.38}$$

Thus, the infinitesimal return dS_t/S_t is the sum of the safe return $r\,dt$ and an additional noise term with zero expectation under P^*. The strength of the noise is measured by the *volatility* parameter σ. Similarly, we obtain

$$dX_t = \sigma X_t\, dW_t = e^{-rt}(dS_t - rS_t\, dt). \tag{5.39}$$

Applying Itô's formula to the function

$$F(x,t) = e^{-rt}u\big(\exp(x + rt - \sigma^2 t/2), T-t\big)$$

and using (5.38), we obtain

$$dV_t = e^{-rt}\frac{\partial u}{\partial x}(S_t,t)\,dS_t + e^{-rt}\Big(\frac{1}{2}\sigma^2 S_t^2\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} - ru\Big)(S_t,t)\,dt.$$

The Black–Scholes partial differential equation (5.33) shows that the term in parenthesis is equal to $-rS_t\,\partial u/\partial x$, and we obtain from (5.39) that

$$dV_t = \frac{\partial u}{\partial x}(S_t,t)\,dX_t = \xi_t\,dX_t.$$

More precisely,

$$V_t = V_0 + \int_0^t \xi_s\,dX_s,$$

where the integral with respect to X is defined as an *Itô integral*, i.e., as the limit of non-anticipating Riemann sums

$$\sum_{t_i\in D_n,\,t_i\le t} \xi_{t_i}(X_{t_{i+1}} - X_{t_i})$$

along an increasing sequence (D_n) of partitions of the interval $[0,T]$; see, e.g., [87]. Thus, the Itô integral can be interpreted in financial terms as the cumulative net gain generated by dynamic hedging in the discounted risky asset as described by the hedging strategy ξ. This fact is an analogue of property (c) in Proposition 5.7, and in this sense $\overline{\xi} = (\xi^0,\xi)$ is a self-financing trading strategy in continuous time. Similarly, we obtain the following continuous-time analogue of (5.5), which describes the undiscounted value of the portfolio as a result of dynamic trading both in the undiscounted risky asset and the riskfree asset:

$$e^{rt}V_t = V_0 + \int_0^t \xi_s\,dS_s + \int_0^t \xi_s^0\,dS_s^0.$$

Perfect replication also works for exotic options $C(S)$ defined by reasonable functionals C on the path space $C[0,T]$, due to a general representation theorem for such functionals as Itô integrals of the underlying Brownian motion W or, via (5.39), of the process X. Weak convergence on path space implies, in analogy to Proposition 5.62, that the arbitrage-free prices of the options $C(S^{(N)})$, computed as discounted expectations under the measure P_N^*, converge to the discounted expectation

$$e^{-rT}\mathbb{E}[\,C(S)\,]$$

under the Wiener measure $\mathbb{P}$.

On the other hand, the discussion in Section 5.6 suggests that the prices of certain exotic contingent claims, such as barrier options, can be computed in closed form as the Black–Scholes price for some corresponding payoff profile of the form $f(S_T)$. This is illustrated by the following example, where the price of an up-and-in call is computed in terms of the distribution of the terminal stock price under the equivalent martingale measure.

Example 5.63 (Black–Scholes price of an up-and-in call option). Consider an up-and-in call option

$$C^{\mathrm{call}}_{\mathrm{u\&i}}(S) = \begin{cases} (S_T - K)^+ & \text{if } \max_{0\le t\le T} S_t \ge B, \\ 0 & \text{otherwise,} \end{cases}$$

where $B > S_0 \vee K$ denotes a given barrier, and where $K > 0$ is the strike price. As approximating models we choose the CRR models of Example 5.55. That is, we have interest rates

$$r_N = \frac{r\,T}{N}$$

and parameters a_N and b_N defined by

$$\hat a_N = 1 + a_N = e^{-\sigma\sqrt{T/N}} \quad \text{and} \quad \hat b_N = 1 + b_N = e^{\sigma\sqrt{T/N}}$$

for some given $\sigma > 0$. Applying the formula obtained in Example 5.50 yields

$$\begin{aligned} &E_N^*\big[\, C^{\mathrm{call}}_{\mathrm{u\&i}}(\widetilde S^{(N)})\,\big] \\ &\quad = E_N^*\big[\,(S_N^{(N)} - K)^+;\ S_N^{(N)} \ge B\,\big] \\ &\qquad + \Big(\frac{p_N^*}{1-p_N^*}\Big)^{k_N}\Big(\frac{B_N}{S_0}\Big)^2 E_N^*\big[\,(S_N^{(N)} - \widetilde K_N)^+;\ S_N^{(N)} < B\,\big], \end{aligned}$$

where B_N and $\widetilde K_N$ are given by

$$\widetilde K_N = K\hat b_N^{-2k_N} = K\Big(\frac{S_0}{B_N}\Big)^2,$$

and where

$$k_N = \Big\lceil \frac{\sqrt N}{\sigma\sqrt T}\log\frac{B}{S_0}\Big\rceil$$

is the smallest integer k such that $B \le S_0 \hat b_N^k$. Then we have

$$B_N \searrow B \quad \text{and} \quad \widetilde K_N \nearrow \widetilde K = K\Big(\frac{S_0}{B}\Big)^2.$$

Since $f(x) = (x-K)^+ \mathrm{I}_{\{x\ge B\}}$ is continuous a.e., we obtain

$$E_N^*\big[\,(S_N^{(N)} - K)^+;\ S_T^{(N)} \ge B\,\big] \longrightarrow \mathbb{E}\big[\,(S_T - K)^+;\ S_T \ge B\,\big],$$

due to Proposition 5.59. Combining the preceding argument with the fact that

$$P_N^*[\,\widetilde K_N \le S_N^{(N)} \le \widetilde K\,] \longrightarrow 0$$

also gives the convergence of the second expectation:

$$E_N^*\big[\,(S_N^{(N)}-\widetilde{K}_N)^+;\ S_N^{(N)}<B\,\big]\longrightarrow \mathbb{E}\big[\,(S_T-\widetilde{K})^+;\ S_T<B\,\big].$$

Next we note that for constants $c, d > 0$

$$\lim_{x\downarrow 0}\frac{1}{x}\log\frac{cx^2+1-e^{-dx}}{e^{dx}-1-cx^2}=\frac{2c}{d}-d,$$

due to l'Hôpital's rule. From this fact, one deduces that

$$\Big(\frac{p_N^*}{1-p_N^*}\Big)^{k_N}\longrightarrow\Big(\frac{B}{S_0}\Big)^{\frac{2r}{\sigma^2}-1}.$$

Thus, we may conclude that the arbitrage-free prices

$$\frac{1}{(1+r_N)^N}E_N^*\big[\,C_{\text{u\&i}}^{\text{call}}(\widetilde{S}^{(N)})\,\big]$$

in the N^{th} approximating model converge to

$$e^{-rT}\bigg(\mathbb{E}\big[\,(S_T-\widetilde{K})^+;\ S_T<B\,\big]+\Big(\frac{B}{S_0}\Big)^{\frac{2r}{\sigma^2}+1}\mathbb{E}\big[\,(S_T-\widetilde{K})^+;\ S_T<B\,\big]\bigg).$$

The expectations occurring in this formula are integrals with respect to a log-normal distribution and can be explicitly computed as in Example 5.57. Moreover, our limit is in fact equal to the *Black–Scholes price of the up-and-in call option*: The functional $C_{\text{u\&i}}^{\text{call}}(\,\cdot\,)$ is continuous in each path in $C[0,T]$ whose maximum is different from the value B, and one can show that these paths have full measure for the law of S under $\mathbb{P}$. Hence, $C_{\text{u\&i}}^{\text{call}}(\,\cdot\,)$ is continuous $\mathbb{P}\circ S^{-1}$-a.e., and the functional version of Proposition 5.59 yields

$$E_N^*\big[\,C_{\text{u\&i}}^{\text{call}}(\widetilde{S}^{(N)})\,\big]\longrightarrow\mathbb{E}\big[\,C_{\text{u\&i}}^{\text{call}}(S)\,\big],$$

so that our limiting price must coincide with the discounted expectation on the right. ◇

Remark 5.64. Let us assume, more generally, that the price process S is defined by

$$S_t=S_0e^{\sigma W_t+\alpha t},\qquad 0\le t\le T,$$

for some $\alpha\in\mathbb{R}$. Applying Itô's formula as in (5.38), we see that S is governed by the stochastic differential equation

$$dS_t=\sigma S_t\,dW_t+bS_t\,dt$$

with $b=\alpha+\frac{1}{2}\sigma^2$. The discounted price process is given by

$$X_t=S_0e^{\sigma W_t+(\alpha-r)t}=S_0e^{\sigma W_t^*-\sigma^2t/2}$$

with $W_t^* = W_t + \lambda t$ for $\lambda = (b - r)/\sigma$. The process W^* is a Wiener process under the measure $P^* \approx \mathbb{P}$ defined by the density

$$\frac{dP^*}{d\mathbb{P}} = e^{-\lambda W_T - \lambda^2 T/2}.$$

In fact, P^* is the unique equivalent martingale measure for X. We can now repeat the arguments above to conclude that the cost of perfect replication for a contingent claim $C(S)$ is given by

$$e^{-rT} E^*[\, C(S) \,].$$ ◇

Even in the context of simple diffusion models such as geometric Brownian motion, however, completeness is lost as soon as the future behavior of the volatility parameter σ is unknown. If, for instance, volatility itself is modeled as a stochastic process, we are facing incompleteness. Thus, the problems of pricing and hedging in discrete-time incomplete markets as discussed in this book reappear in continuous time. Other versions of the invariance principle may lead to other classes of continuous-time models with discontinuous paths, for instance to geometric Poisson or Lévy processes. Discontinuity of paths is another important source of incompleteness. In fact, this has already been illustrated in this book, since discrete-time models can be regarded as stochastic processes in continuous time, where jumps occur at predictable dates.

Chapter 6
American contingent claims

So far, we have studied European contingent claims whose payoff is due at a fixed maturity date. In the case of American options, the buyer can claim the payoff at any time up to the expiration of the contract.

First, we take the point of view of the *seller*, whose aim is to hedge against all possible claims of the buyer. In Section 6.1, this problem is solved under the assumption of market completeness, using the Snell envelope of the contingent claim. The *buyer* tries to choose the best date for exercising the claim, contingent on the information available up to that time. Since future prices are usually unknown, a formulation of this problem will typically involve subjective preferences. If preferences are expressed in terms of expected utility, the choice of the best exercise date amounts to solving an *optimal stopping problem*. In the special case of a complete market model, any exercise strategy which maximizes the expected payoff under the unique equivalent martingale measure turns out to be optimal even in an almost sure sense.

In Section 6.3, we characterize the set of all arbitrage-free prices of an American contingent claim in an incomplete market model. This involves a lower Snell envelope of the claim, which is analyzed in Section 6.5, using the fact that the class of equivalent martingale measures is stable under pasting. This notion of stability under pasting is discussed in Section 6.4 in a general context, and in Section 6.5 we explain its connection with the time-consistency of dynamic risk measures. The results on lower Snell envelopes can also be regarded as a solution to the buyer's optimal stopping problem in the case where preferences are described by robust Savage functionals. Moreover, these results will be used in the theory of superhedging of Chapter 7.

6.1 Hedging strategies for the seller

We start this section by introducing the Doob decomposition of an adapted process and the notion of a supermartingale.

Proposition 6.1. *Let Q be a probability measure on $(\Omega, \mathcal{F}_T)$, and suppose that Y is an adapted process such that $Y_t \in \mathcal{L}^1(Q)$ for all t. Then there exists a unique decomposition*

$$Y = M - A, \tag{6.1}$$

where M is a Q-martingale and A is a process such that $A_0 = 0$ and $(A_t)_{t=1,\dots,T}$ is predictable. The decomposition (6.1) *is called the* Doob decomposition *of Y with respect to the probability measure Q.*

Proof. Define A by

$$A_t - A_{t-1} := -E_Q[\, Y_t - Y_{t-1} \mid \mathcal{F}_{t-1}\,] \quad \text{for } t = 1, \dots, T. \tag{6.2}$$

Then A is predictable and $M_t := Y_t + A_t$ is a Q-martingale. Clearly, any process A with the required properties must satisfy (6.2), so the uniqueness of the decomposition follows. □

Definition 6.2. Let Q be a probability measure on $(\Omega, \mathcal{F}_T)$ and suppose that Y is an adapted process such that $Y_t \in \mathcal{L}^1(Q)$ for all t. Denote by $Y = M - A$ the Doob decomposition of Y.

(a) Y is called a *Q-supermartingale* if A is increasing.

(b) Y is called a *Q-submartingale* if A is decreasing.

Clearly, a process is a martingale if and only if it is both a supermartingale and a submartingale, i.e., if and only if $A \equiv 0$. The following result gives equivalent characterizations of the supermartingale property of a process Y. Its proof is elementary and left to the reader.

Proposition 6.3. *Let Y be an adapted process with $Y_t \in \mathcal{L}^1(Q)$ for all t. Then the following conditions are equivalent:*

(a) *Y is a Q-supermartingale.*

(b) $Y_s \geq E_Q[\, Y_t \mid \mathcal{F}_s\,]$ *for* $0 \leq s \leq t \leq T$.

(c) $Y_{t-1} \geq E_Q[\, Y_t \mid \mathcal{F}_{t-1}\,]$ *for* $t = 1, \dots, T$.

(d) *$-Y$ is a Q-submartingale.*

We now return to the market model introduced in Section 5.1. An American option, or American contingent claim, corresponds to a contract which is issued at time 0 and which obliges the seller to pay a certain amount $C_\tau \geq 0$ if the buyer decides at time τ to exercise the option. The choice of the exercise time τ is entirely up to the buyer, except that the claim is automatically exercised at the "expiration date" of the claim. The American contingent claim can be exercised only once: It becomes invalid as soon as the payoff has been claimed by the buyer. This concept is formalized as follows:

Definition 6.4. An *American contingent claim* is a non-negative adapted process $C = (C_t)_{t=0,\dots,T}$ on the filtered space $(\Omega, (\mathcal{F}_t)_{t=0,\dots,T})$.

For each t, the random variable C_t is interpreted as the payoff of the American contingent claim if the claim is exercised at time t. The time horizon T plays the role of the expiration date of the claim. The possible exercise times for C are not limited to fixed deterministic times $t \in \{0. \dots, T\}$; the buyer may exercise the claim in a way which depends on the scenario $\omega \in \Omega$ of the market evolution.

Definition 6.5. An *exercise strategy* for an American contingent claim C is an $\mathcal{F}_T$-measurable random variable τ taking values in $\{0, \dots, T\}$. The payoff obtained by using τ is equal to

$$C_\tau(\omega) := C_{\tau(\omega)}(\omega), \quad \omega \in \Omega.$$

Example 6.6. An *American put option* on the i^{th} asset and with strike $K > 0$ pays the amount

$$C_t^{\text{put}} := (K - S_t^i)^+$$

if it is exercised at time t. The payoff at time t of the corresponding *American call option* is given by

$$C_t^{\text{call}} := (S_t^i - K)^+.$$

Clearly, the American call option is "out of the money" (i.e., has zero payoff) if the corresponding American put is "in the money" (i.e., has non-zero payoff). It is therefore a priori clear that the respective owners of C^{put} and C^{call} will usually exercise their claims at different times. In particular, there will be no put-call parity for American options. ◇

Similarly, one defines American versions of most options mentioned in the examples of Section 5.3. Clearly, the value of an American option is at least as high as the value of the corresponding European option with maturity T.

Remark 6.7. It should be emphasized that the concept of American contingent claims can be regarded as a generalization of European contingent claims: If C^E is a European contingent claim, then we can define a corresponding American claim C^A by

$$C_t^A = \begin{cases} 0 & \text{if } t < T, \\ C^E & \text{if } t = T. \end{cases} \tag{6.3}$$

◇

Example 6.8. A *Bermuda option* can be exercised by its buyer at each time of a predetermined subset $\mathbb{T} \subset \{0, \dots, T\}$. For instance, a Bermuda call option pays the amount $(S_t^i - K)^+$ if it is exercised at some time $t \in \mathbb{T}$. Thus, a Bermuda option is a financial instrument "between" an American option with $\mathbb{T} = \{0, \dots, T\}$ and a European option with $\mathbb{T} = \{T\}$, just as Bermuda lies between America and Europe; hence the name "Bermuda option". A Bermuda option can be regarded as a particular American option C that pays the amount $C_t = 0$ for $t \notin \mathbb{T}$. ◇

The process

$$H_t = \frac{C_t}{S_t^0}, \quad t = 0, \dots, T,$$

of discounted payoffs of C will be called the *discounted American claim* associated with C. As far as the mathematical theory is concerned, the discounted American claim H will be the primary object. For certain examples it will be helpful to keep track of the numéraire and, thus, of the payoffs C_t prior to discounting.

In this section, we will analyze the theory of hedging American claims in a *complete* market model. We will therefore assume throughout this section that the set $\mathcal{P}$ of equivalent martingale measures consists of one single element P^*:

$$\mathcal{P} = \{ P^* \}.$$

Under this assumption, we will construct a suitable trading strategy that permits the seller of an American claim to hedge against the buyer's discounted claim H_τ. Let us first try to characterize the minimal amount of capital U_t which will be needed at time $t \in \{0, \dots, T\}$. Since the choice of the exercise time τ is entirely up to the buyer, the seller must be prepared to pay at any time t the current payoff H_t of the option. This amounts to the condition $U_t \geq H_t$. Moreover, the amount U_t must suffice to cover the purchase of the hedging portfolio for the possible payoffs H_u for $u > t$. Since the latter condition is void at maturity, we require

$$U_T = H_T.$$

At time $T - 1$, our first requirement on U_{T-1} reads $U_{T-1} \geq H_{T-1}$. The second requirement states that the amount U_{T-1} must suffice for hedging the claim H_T in case the option is not exercised before time T. Due to our assumption of market completeness, the latter amount equals

$$E^*[H_T \mid \mathcal{F}_{T-1}] = E^*[U_T \mid \mathcal{F}_{T-1}].$$

Thus,

$$U_{T-1} := H_{T-1} \vee E^*[U_T \mid \mathcal{F}_{T-1}]$$

is the minimal amount that fulfills both requirements. Iterating this argument leads to the following recursive scheme for U_t:

$$U_T := H_T, \quad U_t := H_t \vee E^*[U_{t+1} \mid \mathcal{F}_t] \quad \text{for } t = T-1, \dots, 0. \tag{6.4}$$

Definition 6.9. The process $U^{P^*} := U$ defined by the recursion (6.4) is called the *Snell envelope* of the process H with respect to the measure P^*.

Example 6.10. Let H^E be a discounted European claim. Then the Snell envelope with respect to P^* of the discounted American claim H^A associated with H^E via (6.3) satisfies

$$U_t^{P^*} = E^*[H_T^A \mid \mathcal{F}_t] = E^*[H^E \mid \mathcal{F}_t].$$

Thus, U is equal to the value process of a replicating strategy for H^E. ◇

Clearly, a Snell envelope U^Q can be defined for any probability measure Q on $(\Omega, \mathcal{F}_T)$ and for any adapted process H that satisfies the following integrability condition:

$$H_t \in \mathcal{L}^1(Q) \quad \text{for } t = 0, \dots, T. \tag{6.5}$$

In our finite-time setting, this condition is equivalent to

$$E_Q\big[\max_{t\leq T}|H_t|\big]<\infty.$$

For later applications, the following proposition is stated for a general measure Q.

Proposition 6.11. *Let H be an adapted process such that* (6.5) *holds. Then the Snell envelope U^Q of H with respect to Q is the smallest Q-supermartingale dominating H: If $\widetilde{U}$ is another Q-supermartingale such that $\widetilde{U}_t \geq H_t$ Q-a.s. for all t, then $\widetilde{U}_t \geq U_t^Q$ Q-a.s. for all t.*

Proof. It follows immediately from the definition of U^Q that $U_{t-1}^Q \geq E_Q[\,U_t^Q \mid \mathcal{F}_{t-1}\,]$ so that U^Q is indeed a supermartingale. If $\widetilde{U}$ is another supermartingale dominating H, then $\widetilde{U}_T \geq H_T = U_T^Q$. We now proceed by backward induction on t. If we already know that $\widetilde{U}_t \geq U_t^Q$, then

$$\widetilde{U}_{t-1} \geq E_Q[\,\widetilde{U}_t \mid \mathcal{F}_{t-1}\,] \geq E_Q[\,U_t^Q \mid \mathcal{F}_{t-1}\,].$$

Adding our assumption $\widetilde{U}_{t-1} \geq H_{t-1}$ yields that

$$\widetilde{U}_{t-1} \geq H_{t-1} \vee E_Q[\,U_t^Q \mid \mathcal{F}_{t-1}\,] = U_{t-1}^Q,$$

and the result follows. □

Proposition 6.11 illustrates how the seller can (super-) hedge a discounted American claim H by using the Doob decomposition

$$U_t^{P^*} = M_t - A_t, \quad t = 0, \dots, T,$$

of the Snell envelope U^{P^*} with respect to P^*. Then M is a P^*-martingale, A is increasing, and $(A_t)_{t=1,\dots,T}$ is predictable. Since we assume the completeness of the market model, Theorem 5.39 yields the representation of the martingale M as the "stochastic integral" of a suitable d-dimensional predictable process ξ:

$$M_t = U_0^{P^*} + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}), \quad t = 0, \dots, T. \tag{6.6}$$

It follows that

$$M_t \geq U_t^{P^*} \geq H_t \quad \text{for all } t.$$

By adding a numéraire component ξ^0 such that $\overline{\xi} = (\xi^0, \xi)$ becomes a self-financing trading strategy with initial investment $U_0^{P^*}$, we obtain a (super-) hedge for H, namely a self-financing trading strategy whose value process V satisfies

$$V_t \geq H_t \quad \text{for all } t. \tag{6.7}$$

Thus, $U_t^{P^*}$ may be viewed as the resulting capital at each time t if we use the self-financing strategy $\overline{\xi}$, combined with a refunding scheme where we withdraw successively the amounts defined by the increments of A. In fact, $U_t^{P^*}$ is the minimal investment at time t for which one can purchase a hedging strategy such that (6.7) holds. This follows from our next result.

Theorem 6.12. *Let H be a discounted American claim with Snell envelope U^{P^*}. Then there exists a d-dimensional predictable process ξ such that*

$$U_t^{P^*} + \sum_{k=t+1}^{u} \xi_k \cdot (X_k - X_{k-1}) \geq H_u \quad \text{for all } u \geq t. \tag{6.8}$$

Moreover, any $\mathcal{F}_t$-measurable random variable $\widetilde{U}_t$ which, for some predictable ξ, satisfies (6.8) *in place of $U_t^{P^*}$ is such that*

$$\widetilde{U}_t \geq U_t^{P^*} \quad P^*\text{-a.s.}$$

Thus, $U_t^{P^*}$ is the minimal amount of capital which is necessary to hedge H from time t up to maturity.

Proof. Clearly, U^{P^*} satisfies (6.8) for ξ as in (6.6). Now suppose that $\widetilde{U}_t$ is $\mathcal{F}_t$-measurable, that $\widetilde{\xi}$ is predictable, and that

$$V_u := \widetilde{U}_t + \sum_{k=t+1}^{u} \widetilde{\xi}_k \cdot (X_k - X_{k-1}) \geq H_u \quad \text{for all } u \geq t.$$

We show $V_u \geq U_u^{P^*}$ for all $u \geq t$ by backward induction. $V_T \geq H_T = U_T^{P^*}$ holds by assumption, so assume $V_{u+1} \geq U_{u+1}^{P^*}$ for some u. Since our market model is complete, Theorem 5.38 implies that $\widetilde{\xi}$ is bounded. Hence, we get

$$E^*[\, V_{u+1} - V_u \mid \mathcal{F}_u\,] = E^*[\, \widetilde{\xi}_{u+1} \cdot (X_{u+1} - X_u) \mid \mathcal{F}_u\,] = 0 \quad P\text{-a.s.}$$

It follows that

$$V_u = E^*[\, V_{u+1} \mid \mathcal{F}_u\,] \geq H_u \vee E^*[\, U_{u+1}^{P^*} \mid \mathcal{F}_u\,] = U_u^{P^*}. \qquad \square$$

6.2 Stopping strategies for the buyer

In this section, we take the point of view of the buyer of an American contingent claim. Thus, our aim is to optimize the exercise strategy. It is natural to assume that the decision to exercise the claim at a particular time t depends only on the market information which is available at t. This constraint can be formulated as follows:

Definition 6.13. A function $\tau : \Omega \to \{0, 1, \dots, T\} \cup \{+\infty\}$ is called a *stopping time* if $\{\tau = t\} \in \mathcal{F}_t$ for $t = 0, \dots, T$.

In particular, the constant function $\tau \equiv t$ is a stopping time for fixed $t \in \{0, \dots, T\}$.

Remark 6.14. A function $\tau : \Omega \to \{0, 1, \dots, T\} \cup \{+\infty\}$ is a stopping time if and only if $\{\tau \le t\} \in \mathcal{F}_t$ for each t. Moreover, if τ and σ are two stopping times, then the following functions are also stopping times:

$$\tau \wedge \sigma, \quad \tau \vee \sigma, \quad (\tau + \sigma) \wedge T.$$

The proof is easy and left to the reader. ◇

Example 6.15. A typical example of a non-trivial stopping time is the first time at which an adapted process Y exceeds a certain level c:

$$\tau(\omega) := \inf\{t \ge 0 \mid Y_t(\omega) \ge c\}.$$

In fact,

$$\{\tau \le t\} = \bigcup_{s=0}^{t} \{Y_s \ge c\} \in \mathcal{F}_t$$

for $t = 0, \dots, T$. This example also illustrates the role of the value $+\infty$ in Definition 6.13: We have $\tau(\omega) = +\infty$ if, for this particular ω, the criterion that triggers τ is not met for any $t \in \{0, \dots, T\}$. ◇

Definition 6.16. For any stochastic process Y and each stopping time τ we denote by Y^τ the *process stopped in* τ:

$$Y_t^\tau(\omega) := Y_{t \wedge \tau(\omega)}(\omega) \quad \text{for } \omega \in \Omega \text{ and for all } t \in \{0, \dots, T\}.$$

It follows from the definition of a stopping time that Y^τ is an adapted process if Y is. Informally, the following basic theorem states that a martingale cannot be turned into a favorable game by using a clever stopping strategy. This result is often called *Doob's stopping theorem* or the *optional sampling theorem*. Recall that we assume $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

Theorem 6.17. *Let M be an adapted process such that $M_t \in \mathcal{L}^1(Q)$ for each t. Then the following conditions are equivalent:*

(a) *M is a Q-martingale.*

(b) *For any stopping time τ the stopped process M^τ is a Q-martingale.*

(c) $E_Q[M_{\tau \wedge T}] = M_0$ *for any stopping time τ.*

Proof. (a)$\Rightarrow$(b): Note that

$$M^\tau_{t+1} - M^\tau_t = (M_{t+1} - M_t)\,\mathrm{I}_{\{\tau>t\}}.$$

Since $\{\tau > t\} \in \mathcal{F}_t$, we obtain that

$$E_Q[\, M^\tau_{t+1} - M^\tau_t \mid \mathcal{F}_t \,] = E_Q[\, M_{t+1} - M_t \mid \mathcal{F}_t \,] \cdot \mathrm{I}_{\{\tau>t\}} = 0.$$

(b)$\Rightarrow$(c): This follows simply from the fact that the expectation of M^τ_t is constant in t.

(c)$\Rightarrow$(a): We need to show that if $t < T$, then

$$E_Q[\, M_T;\ A\,] = E_Q[\, M_t;\ A\,] \tag{6.9}$$

for each $A \in \mathcal{F}_t$. Fix such an A and define a stopping time τ as

$$\tau(\omega) := \begin{cases} t & \text{if } \omega \in A, \\ T & \text{if } \omega \notin A. \end{cases}$$

We obtain that

$$M_0 = E_Q[\, M_{T\wedge\tau}\,] = E_Q[\, M_t;\ A\,] + E_Q[\, M_T;\ A^c\,].$$

Using the constant stopping time T instead of τ yields that

$$M_0 = E_Q[\, M_T\,] = E_Q[\, M_T;\ A\,] + E_Q[\, M_T;\ A^c\,].$$

Subtracting the latter identity from the previous one yields (6.9). □

Corollary 6.18. *Let U be an adapted process such that $U_t \in \mathcal{L}^1(Q)$ for each t. Then the following conditions are equivalent:*

(a) *U is a Q-supermartingale.*

(b) *For any stopping time τ, the stopped process U^τ is a Q-supermartingale.*

Proof. If $U = M - A$ is the Doob decomposition of U, then one checks that

$$U^\tau = M^\tau - A^\tau \tag{6.10}$$

is the Doob decomposition of U^τ. This observation and Theorem 6.17 yield the equivalence of (a) and (b). □

Let us return to the problem of finding an optimal exercise time τ for a discounted American claim H. We assume that the buyer chooses the possible exercise times from the set

$$\mathcal{T} := \big\{\, \tau \mid \tau \text{ is a stopping time with } \tau \le T \,\big\}$$

of all stopping times which do not take the value $+\infty$. Assume that the aim of the buyer is to choose a payoff from the class $\{ H_\tau \mid \tau \in \mathcal{T} \}$ which is optimal in the sense that it has maximal expectation. Thus, the problem is:

$$\text{Maximize } E[\, H_\tau \,] \text{ among all } \tau \in \mathcal{T}. \tag{6.11}$$

The analysis of the *optimal stopping problem* (6.11) does not require any properties of the underlying market model, not even the absence of arbitrage. We may also drop the positivity assumption on H: All we have to assume is that H is an adapted process which satisfies

$$H_t \in \mathcal{L}^1(\Omega, \mathcal{F}_t, P) \quad \text{for all } t. \tag{6.12}$$

This relaxed assumption will be useful in Chapter 9, and it allows us to include the interpretation of the optimal stopping problem in terms of the following utility maximization problem:

Remark 6.19. Suppose the buyer uses a preference relation on $\mathcal{X} := \{ H_\tau \mid \tau \in \mathcal{T} \}$ which can be represented in terms of a *Savage representation*

$$U(H_\tau) = E_Q[\, u(H_\tau)\,]$$

where Q is a probability measure on $(\Omega, \mathcal{F})$, and u is a measurable or continuous function; see Section 2.5. Then a natural goal is to maximize the utility $U(H_\tau)$ among all $\tau \in \mathcal{T}$. This is equivalent to the optimal stopping problem (6.11) for the transformed process $\widetilde{H}_t := u(H_t)$, and with respect to the measure Q instead of P. This utility maximization problem is covered by the discussion in this section as long as $\widetilde{H}_t \in \mathcal{L}^1(Q)$ for all t. In Remark 6.51 we will discuss the problem of maximizing the more general utility functionals which appear in a *robust* Savage representation. ◇

Under the assumption (6.12), we can construct the Snell envelope $U := U^P$ of H with respect to P, i.e., U is defined via the recursive formula

$$U_T := H_T \quad \text{and} \quad U_t := H_t \vee E[\, U_{t+1} \mid \mathcal{F}_t\,], \quad t = T-1, \dots, 0.$$

Let us define a stopping time $\tau_{\min}$ by

$$\tau_{\min} := \min\{\, t \ge 0 \mid U_t = H_t \,\}.$$

Note that $\tau_{\min} \le T$ since $U_T = H_T$. As we will see in the following theorem, $\tau_{\min}$ maximizes the expectation of H_τ among all $\tau \in \mathcal{T}$. In other words, $\tau_{\min}$ is a solution to our optimal stopping problem (6.11). Similarly, we let

$$\tau_{\min}^{(t)} := \min\{\, u \ge t \mid U_u = H_u \},$$

which is a member of the set

$$\mathcal{T}_t := \{\, \tau \in \mathcal{T} \mid \tau \ge t \,\}.$$

Theorem 6.20. *The Snell envelope U of H satisfies*

$$U_t = E[\, H_{\tau^{(t)}_{\min}} \mid \mathcal{F}_t \,] = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E[\, H_\tau \mid \mathcal{F}_t \,].$$

In particular,

$$U_0 = E[\, H_{\tau_{\min}} \,] = \sup_{\tau \in \mathcal{T}} E[\, H_\tau \,].$$

Proof. Since U is a supermartingale under P, Corollary 6.18 shows that for $\tau \in \mathcal{T}_t$

$$U_t \geq E[\, U_\tau \mid \mathcal{F}_t \,] \geq E[\, H_\tau \mid \mathcal{F}_t \,].$$

Therefore,

$$U_t \geq \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E[\, H_\tau \mid \mathcal{F}_t \,].$$

Hence, the theorem will be proved if we can show that $U_t = E[\, H_{\tau^{(t)}_{\min}} \mid \mathcal{F}_t \,]$, which is in turn implied by the identity

$$U_t = E[\, U_{\tau^{(t)}_{\min}} \mid \mathcal{F}_t \,]. \tag{6.13}$$

In order to prove (6.13), let $U^{(t)}$ denote the stopped process

$$U^{(t)}_s := U_{s \wedge \tau^{(t)}_{\min}},$$

and fix some s between t and T. Then $U_s > H_s$ on $\{\tau^{(t)}_{\min} > s\}$. Hence, P-a.s. on $\{\tau^{(t)}_{\min} > s\}$:

$$U^{(t)}_s = U_s = H_s \vee E[\, U_{s+1} \mid \mathcal{F}_s \,] = E[\, U_{s+1} \mid \mathcal{F}_s \,] = E[\, U^{(t)}_{s+1} \mid \mathcal{F}_s \,].$$

On the set $\{\tau^{(t)}_{\min} \leq s\}$ one has $U^{(t)}_{s+1} = U_{\tau^{(t)}_{\min}} = U^{(t)}_s$, hence $U^{(t)}_s = E[\, U^{(t)}_{s+1} \mid \mathcal{F}_s \,]$. Thus, $U^{(t)}$ is a martingale from time t on:

$$U^{(t)}_s = E[\, U^{(t)}_{s+1} \mid \mathcal{F}_s \,] \quad \text{for all } s \in \{t, t+1, \dots, T\}.$$

It follows that

$$E[\, U_{\tau^{(t)}_{\min}} \mid \mathcal{F}_t \,] = E[\, U^{(t)}_T \mid \mathcal{F}_t \,] = U^{(t)}_t = U_t.$$

This proves the claim (6.13). □

Definition 6.21. A stopping time $\tau^* \in \mathcal{T}$ is called *optimal* (*with respect to* P) if

$$E[\, H_{\tau^*} \,] = \sup_{\tau \in \mathcal{T}} E[\, H_\tau \,].$$

In particular, $\tau_{\min}$ is an optimal stopping time in the sense of this definition. The following result implies that $\tau_{\min}$ is in fact the minimal optimal stopping time.

Proposition 6.22. *A stopping time $\tau \in \mathcal{T}$ is optimal if and only if $H_\tau = U_\tau$ P-a.s., and if the stopped process U^τ is a martingale. In particular, any optimal stopping time τ satisfies $\tau \geq \tau_{\min}$.*

Proof. First note that $\tau \in \mathcal{T}$ is optimal if it satisfies the two conditions of the assertion, because then Theorem 6.20 implies that

$$\sup_{\sigma \in \mathcal{T}} E[\, H_\sigma \,] = U_0 = E[\, U_T^\tau \,] = E[\, U_\tau \,] = E[\, H_\tau \,].$$

For the converse implication, we apply the assumption of optimality, the fact that $H_\tau \leq U_\tau$, and the stopping theorem for supermartingales to obtain that

$$U_0 = E[\, H_\tau \,] \leq E[\, U_\tau \,] \leq U_0,$$

so that all inequalities are in fact equalities. It follows in particular that $H_\tau = U_\tau$ P-almost surely. Moreover, the identity $E[\, U_\tau \,] = U_0$ implies that the stopped process U^τ is a supermartingale with constant expectation U_0, and hence is a martingale. □

In general, there can be many different optimal stopping times. The largest optimal stopping time admits an explicit description: It is the first time before T for which the Snell envelope U loses the martingale property:

$$\begin{aligned} \tau_{\max} &:= \inf\big\{\, t \geq 0 \mid E[\, U_{t+1} - U_t \mid \mathcal{F}_t \,] \neq 0 \big\} \wedge T \\ &= \inf\{\, t \geq 0 \mid A_{t+1} \neq 0\} \wedge T. \end{aligned}$$

Here, A denotes the increasing process obtained from the Doob decomposition of U under P.

Theorem 6.23. *The stopping time $\tau_{\max}$ is the largest optimal stopping time. Moreover, a stopping time τ is optimal if and only if P-a.s. $\tau \leq \tau_{\max}$ and $U_\tau = H_\tau$.*

Proof. Let $U = M - A$ be the Doob decomposition of U. Recall from (6.10) that $U^\tau = M^\tau - A^\tau$ is the Doob decomposition of U^τ for any stopping time τ. Thus, U^τ is a martingale if and only if $A_\tau = 0$, because A is increasing. Therefore, U^τ is a martingale if and only if $\tau \leq \tau_{\max}$, and so the second part of the assertion follows from Proposition 6.22. It remains to prove that $\tau_{\max}$ itself is optimal, i.e., that $U_{\tau_{\max}} = H_{\tau_{\max}}$. This is clear on the set $\{\, \tau_{\max} = T \,\}$. On the set $\{\, \tau_{\max} = t \,\}$ for $t < T$ one has $A_t = 0$ and $A_{t+1} > 0$. Hence,

$$E[\, U_{t+1} - U_t \mid \mathcal{F}_t \,] = -(A_{t+1} - A_t) = -A_{t+1} < 0 \quad \text{on } \{\, \tau_{\max} = t \,\}.$$

Thus, $U_t > E[\, U_{t+1} \mid \mathcal{F}_t \,]$ and the definition of the Snell envelope yields that $U_t = H_t \vee E[\, U_{t+1} \mid \mathcal{F}_t \,] = H_t$ on $\{\, \tau_{\max} = t \,\}$. □

Let us now return to our complete financial market model, where H_t is the discounted payoff of an American contingent claim. Thus, an optimal stopping strategy for H maximizes the *expected* payoff $E[\,H_\tau\,]$. But a stopping time turns out to be the best choice even in a *pathwise* sense, provided that it is optimal with respect to the unique equivalent martingale measure P^* in a complete market model. In order to explain this fact, let us first recall from Section 6.1 the construction of a perfect hedge of H from the seller's perspective. Let

$$U^{P^*} = M - A$$

denote the Doob decomposition of the Snell envelope U^{P^*} of H with respect to P^*. Since P^* is the unique equivalent martingale measure in our model, the martingale M has the representation

$$M_t = U_0^{P^*} + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}), \quad t = 0, \dots, T,$$

for a d-dimensional predictable process ξ. Clearly, M is equal to the value process of the self-financing strategy constructed from ξ and the initial investment $U_0^{P^*}$. Since M dominates H, this yields a perfect hedge of H from the perspective of the seller: If the buyer exercises the option at some stopping time τ, then the seller makes a profit $M_\tau - H_\tau \geq 0$. The following corollary states that the buyer can in fact meet the value of the seller's hedging portfolio, and that this happens if and only if the option is exercised at an optimal stopping time with respect to P^*. In this sense, $U_0^{P^*}$ can be regarded as the unique *arbitrage-free price* of the discounted American claim H.

Corollary 6.24. *With the above notation,*

$$H_\tau \leq M_\tau = U_0^{P^*} + \sum_{k=1}^{\tau} \xi_k \cdot (X_k - X_{k-1}), \quad P^*\text{-a.s. for all } \tau \in \mathcal{T},$$

and equality holds P^*-almost surely if and only if τ is optimal with respect to P^*.

Proof. At time τ,

$$H_\tau \leq U_\tau^{P^*} = M_\tau - A_\tau \leq M_\tau.$$

Moreover, by Theorem 6.23, both $H_\tau = U_\tau^{P^*}$ and $A_\tau = 0$ hold P^*-a.s. if and only if τ is optimal with respect to P^*. □

Let us now compare a discounted American claim H to the corresponding discounted European claim H_T, i.e., to the contract which is obtained from H by restricting the exercise time to be T. In particular, we are interested in the relation between American and European put or call options. Let

$$V_t := E^*[\,H_T \mid \mathcal{F}_t\,]$$

denote the amount needed at time t to hedge H_T. Since our market model is complete, V_t can also be regarded as the unique arbitrage-free price of the discounted claim H_T at time t. From the seller's perspective, $U_t^{P^*}$ plays a similar role for the American option. It is intuitively clear that an American claim should be more expensive than the corresponding European one. This is made mathematically precise in the following statement.

Proposition 6.25. *With the above notation, $U_t^{P^*} \geq V_t$ for all t. Moreover, if V dominates H, then U^{P^*} and V coincide.*

Proof. The first statement follows immediately from the supermartingale property of U^{P^*}:

$$U_t^{P^*} \geq E^*[\, U_T^{P^*} \mid \mathcal{F}_t \,] = E^*[\, H_T \mid \mathcal{F}_t \,] = V_t.$$

Next, if the P^*-martingale V dominates H, then it also dominates the corresponding Snell envelope U^{P^*} by Proposition 6.11. Thus V and U^{P^*} must coincide. □

Remark 6.26. The situation in which V dominates H occurs, in particular, when the process H is a P^*-submartingale. This happens, for instance, if H is obtained by applying a convex function $f : \mathbb{R}^d \to [0, \infty)$ to the discounted price process X. Indeed, in this case, Jensen's inequality for conditional expectations implies that

$$E^*[\, f(X_{t+1}) \mid \mathcal{F}_t \,] \geq f\big(E^*[\, X_{t+1} \mid \mathcal{F}_t \,]\big) = f(X_t). \qquad \diamond$$

Example 6.27. The discounted payoff of an American call option $C_t^{\text{call}} = (S_t^1 - K)^+$ is given by

$$H_t^{\text{call}} = \left(X_t^1 - \frac{K}{S_t^0} \right)^+.$$

Under the hypothesis that S_t^0 is increasing in t, (5.21) states that

$$E^*[\, H_{t+1}^{\text{call}} \mid \mathcal{F}_t \,] \geq H_t^{\text{call}} \qquad P^*\text{-a.s. for } t = 0, \dots, T-1.$$

In other words, H^{call} is a submartingale, and the Snell envelope U^{P^*} of H^{call} coincides with the value process

$$V_t = E^*\left[\left(X_T^1 - \frac{K}{S_T^0} \right)^+ \,\middle|\, \mathcal{F}_t \right]$$

of the corresponding European call option with maturity T. In particular, we have $U_0^{P^*} = V_0$, i.e., the unique arbitrage-free price of the American call option is equal to its European counterpart. Moreover, Theorem 6.23 implies that the maximal optimal stopping time with respect to P^* is given by $\tau_{\max} \equiv T$. This suggests that, in a complete model, an American call should not be exercised before maturity. $\diamond$

Example 6.28. For an *American put option* $C_t^{\text{put}} := (K - S_t^1)^+$ the situation is different, because the argument in (5.21) fails unless S^0 is decreasing. If S^0 is an increasing bond, then the time value

$$W_t := S_t^0 \, E^*\Big[\frac{(K - S_T^1)^+}{S_T^0} \;\Big|\; \mathcal{F}_t \Big] - (K - S_t^1)^+$$

of a European put $(K - S_T^1)^+$ typically becomes negative at a certain time t, corresponding to an *early exercise premium* $-W_t$; see Figure 5.3. Thus, the early exercise premium is the surplus which an owner of the American put option would have over the value of the European put $(K - S_T^1)^+$.

The relation between the price of a put option and its intrinsic value can be illustrated in the context of the CRR model. With the notation of Section 5.5, the price process of the risky asset $S_t = S_t^1$ can be written as

$$S_t = S_0 \, \Lambda_t \quad \text{for} \quad \Lambda_t := \prod_{k=1}^{t} (1 + R_k)$$

and with the constant $S_0 \geq 0$. Recall that the returns R_k can take only two possible values a and b with $-1 < a < b$, and that the market model is arbitrage-free if and only if the riskless interest rate r satisfies $a < r < b$. In this case, the model is complete, and the unique equivalent martingale measure P^* is characterized by the fact that it makes $R_1, \dots, R_T$ independent with common distribution

$$P^*[\, R_k = b \,] = p^* = \frac{r - a}{b - a}. \tag{6.14}$$

Let

$$\pi(x) := \sup_{\tau \in \mathcal{T}} E^*\Big[\frac{(K - x\Lambda_\tau)^+}{(1 + r)^\tau} \Big]$$

denote the price of C^{put} regarded as a function of $x := S_0$. Clearly, $\pi(x)$ is a convex and decreasing function in x. Let us assume that $r > 0$ and that the parameter a is strictly negative. A trivial situation occurs if the option is "far out of the money" in the sense that

$$x \geq \frac{K}{(1 + a)^T},$$

because then $S_t = x\Lambda_t \geq K$ for all t, and the payoff of C^{put} is always zero. In particular, $\pi(x) = 0$. If

$$x \leq \frac{K}{(1 + b)^T} \tag{6.15}$$

then $S_t = x\Lambda_t \leq K$ for all t, and hence

$$\pi(x) = \sup_{\tau \in \mathcal{T}} \Big(E^*\Big[\frac{K}{(1 + r)^\tau} \Big] - x \Big) = K - x.$$

In this case, the price of the American put option is equal to its intrinsic value $(K-x)^+$ at time $t = 0$, and an optimal strategy for the owner would simply consist in exercising the option immediately, i.e., there is no demand for the option in the regime (6.15).

Now consider the case

$$K \leq x < \frac{K}{(1+a)^T}$$

of a put option which is "at the money" or "not too far out of the money". For large enough $t > 0$, the probability $P^*[\, C_t^{\text{put}} > 0\,]$ of a non-zero payoff is strictly positive, while the intrinsic value $(K - x)^+$ vanishes. It follows that the price $\pi(x)$ is strictly higher than the intrinsic value, and so it is not optimal for the buyer to exercise the option immediately.

Summarizing our observations, we can say that there exists a value x^* with

$$\frac{K}{(1+b)^T} < x^* \leq K$$

such that

$$\begin{aligned} \pi(x) &= (K-x)^+ && \text{for } x \leq x^*, \\ \pi(x) &> (K-x)^+ && \text{for } x^* < x < K/(1+a)^T, \text{ and} \\ \pi(x) &= 0 && \text{for } x \geq K/(1+a)^T; \end{aligned}$$

see Figure 6.1. ◇

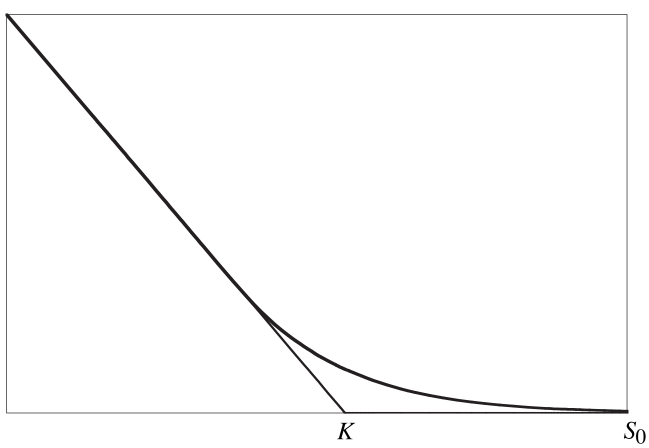


Figure 6.1. The price of an American put option as a function of S_0 compared to the option's intrinsic value $(K - S_0)^+$.

Remark 6.29. In the context of an arbitrage-free CRR model, we consider a discounted American claim H whose payoff is determined by a function of time and of the current spot price, i.e.,

$$H_t = h_t(S_t) \quad \text{for all } t.$$

Clearly, this setting includes American call and put options as special cases. By using the same arguments as in the derivation of (5.25), we get that the Snell envelope U^{P^*} of H is of the form

$$U_t^{P^*} = u_t(S_t), \quad t = 0, \dots, T,$$

where the functions u_t are determined by the recursion

$$u_T(x) = h_T(x) \quad \text{and} \quad u_t(x) = h_t(x) \vee \big(u_{t+1}(x\,\hat{b})\,p^* + u_{t+1}(x\,\hat{a})\,(1-p^*)\big).$$

Here p^* is defined as in (6.14), and the parameters $\hat{a}$ and $\hat{b}$ are given by $\hat{a} = 1 + a$ and $\hat{b} = 1 + b$. Thus, the space $[0, T] \times [0, \infty)$ can be decomposed into the two regions

$$\mathcal{R}_c := \{\,(t, x) \mid u_t(x) > h_t(x)\,\} \quad \text{and} \quad \mathcal{R}_s := \{\,(t, x) \mid u_t(x) = h_t(x)\,\},$$

and the minimal optimal stopping time $\tau_{\min}$ can be described as the first exit time of the space time process (t, S_t) from the *continuation region* $\mathcal{R}_c$ or, equivalently, as the first entrance time into the *stopping region* $\mathcal{R}_s$:

$$\tau_{\min} = \min\{\,t \ge 0 \mid (t, S_t) \notin \mathcal{R}_c\,\} = \min\{\,t \ge 0 \mid (t, S_t) \in \mathcal{R}_s\,\}. \qquad \diamond$$

6.3 Arbitrage-free prices

In this section, we drop the condition of market completeness, and we develop the notion of an arbitrage-free price π for a discounted American claim H in a general incomplete framework. The basic idea consists in reducing the problem to the determination of the arbitrage-free price for the payoff H_τ which arises from H by fixing the exercise strategy τ. The following remark explains that H_τ can be treated like the discounted payoff of a European contingent claim, whose set of arbitrage-free prices is given by

$$\Pi(H_\tau) = \big\{\, E^*[\,H_\tau\,] \mid P^* \in \mathcal{P},\ E^*[\,H_\tau\,] < \infty \,\big\}. \tag{6.16}$$

Remark 6.30. As observed in Remark 5.34, a discounted payoff $\widetilde{H}_t$ which is received at time $t < T$ can be regarded as a discounted European claim $\widetilde{H}^E$ maturing at T. $\widetilde{H}^E$ is obtained from $\widetilde{H}_t$ by investing at time t the payoff $S_t^0\widetilde{H}_t$ into the numéraire, i.e., by buying $\widetilde{H}_t$ shares of the 0^{th} asset, and by considering the discounted terminal value of this investment:

$$\widetilde{H}^E = \frac{1}{S_T^0}\,(\,S_T^0\,\widetilde{H}_t\,) = \widetilde{H}_t.$$

In the case of our discounted American claim H which is payed off at the random time τ, we can either apply this argument to each payoff

$$\widetilde{H}_t := H_\tau \, \mathrm{I}_{\{\tau=t\}} = H_t \, \mathrm{I}_{\{\tau=t\}},$$

or directly use a stopping time version of this argument. We conclude that H_τ can be regarded as a discounted European claim, whose arbitrage-free prices are given by (6.16). ◇

Now suppose that H is offered at time $t = 0$ for a price $\pi \geq 0$. From the buyer's point of view there should be at least one exercise strategy τ such that the proposed price π is not too high in the sense that $\pi \leq \pi'$ for some $\pi' \in \Pi(H_\tau)$. From the seller's point of view the situation looks different: There should be no exercise strategy τ' such that the proposed price π is too low in the sense that $\pi < \pi'$ for all $\pi' \in \Pi(H_{\tau'})$. By adding the assumption that the buyer only uses stopping times in exercising the option, we obtain the following formal definition.

Definition 6.31. A real number π is called an *arbitrage-free price* of a discounted American claim H if the following two conditions are satisfied.

- The price π is not too high in the sense that there exists some $\tau \in \mathcal{T}$ and $\pi' \in \Pi(H_\tau)$ such that $\pi \leq \pi'$.
- The price π is not too low in the sense that there exists no $\tau' \in \mathcal{T}$ such that $\pi < \pi'$ for all $\pi' \in \Pi(H_{\tau'})$.

The set of all arbitrage-free prices of H is denoted $\Pi(H)$, and we define

$$\pi_{\inf}(H) := \inf \Pi(H) \quad \text{and} \quad \pi_{\sup}(H) := \sup \Pi(H).$$

Recall from Remark 6.7 that every discounted European claim H^E can be regarded as a discounted American claim H^A whose payoff is zero if H^A is exercised before T, and whose payoff at T equals H^E. Clearly, the two sets $\Pi(H^E)$ and $\Pi(H^A)$ coincide, and so the two Definitions 5.29 and 6.31 are consistent with each other.

Remark 6.32. It follows from the definition that any arbitrage-free price π for H must be an arbitrage-free price for some H_τ. Hence, (6.16) implies that $\pi = E^*[\, H_\tau \,]$ for some $P^* \in \mathcal{P}$. Similarly, we obtain from the second condition in Definition 6.31 that $\pi \geq \inf_{P^* \in \mathcal{P}} E^*[\, H_\tau \,]$ for all $\tau \in \mathcal{T}$. It follows that

$$\sup_{\tau \in \mathcal{T}} \inf_{P^* \in \mathcal{P}} E^*[\, H_\tau \,] \leq \pi \leq \sup_{\tau \in \mathcal{T}} \sup_{P^* \in \mathcal{P}} E^*[\, H_\tau \,] \quad \text{for all } \pi \in \Pi(H). \tag{6.17}$$

In particular,

$$\sup_{\tau \in \mathcal{T}} E^*[\, H_\tau \,]$$

is the unique arbitrage-free price of H if P^* is the unique equivalent martingale measure in a complete market model, and so Definition 6.31 is consistent with the results of the Section 6.1 and 6.2. ◇

Our main goal in this section is to characterize the set $\Pi(H)$, and to identify the upper and lower bounds in (6.17) with the quantities $\pi_{\sup}(H)$ and $\pi_{\inf}(H)$. We will work under the simplifying assumption that

$$H_t \in \mathcal{L}^1(P^*) \text{ for all } t \text{ and each } P^* \in \mathcal{P}. \tag{6.18}$$

Note that (6.18) implies the condition

$$\inf_{P^*\in\mathcal{P}} \sup_{\tau\in\mathcal{T}} E^*[\, H_\tau \,] < \infty.$$

For each $P^* \in \mathcal{P}$ we denote by U^{P^*} the corresponding Snell envelope of H, i.e.,

$$U_t^{P^*} = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}_t} E^*[\, H_\tau \mid \mathcal{F}_t \,].$$

With this notation, the right-hand bound in (6.17) can be written as

$$\sup_{\tau\in\mathcal{T}} \sup_{P^*\in\mathcal{P}} E^*[\, H_\tau \,] = \sup_{P^*\in\mathcal{P}} \sup_{\tau\in\mathcal{T}} E^*[\, H_\tau \,] = \sup_{P^*\in\mathcal{P}} U_0^{P^*}.$$

In fact, a similar relation also holds for the lower bound in (6.17):

$$\sup_{\tau\in\mathcal{T}} \inf_{P^*\in\mathcal{P}} E^*[\, H_\tau \,] = \inf_{P^*\in\mathcal{P}} \sup_{\tau\in\mathcal{T}} E^*[\, H_\tau \,] = \inf_{P^*\in\mathcal{P}} U_0^{P^*}. \tag{6.19}$$

The proof that the above interchange of infimum and supremum is indeed justified under assumption (6.18) is postponed to the next section; see Theorem 6.47.

Theorem 6.33. *Under condition* (6.18)*, the set of arbitrage-free prices for H is a real interval with endpoints*

$$\pi_{\inf}(H) = \inf_{P^*\in\mathcal{P}} \sup_{\tau\in\mathcal{T}} E^*[\, H_\tau \,] = \sup_{\tau\in\mathcal{T}} \inf_{P^*\in\mathcal{P}} E^*[\, H_\tau \,]$$

and

$$\pi_{\sup}(H) = \sup_{P^*\in\mathcal{P}} \sup_{\tau\in\mathcal{T}} E^*[\, H_\tau \,] = \sup_{\tau\in\mathcal{T}} \sup_{P^*\in\mathcal{P}} E^*[\, H_\tau \,].$$

Moreover, $\Pi(H)$ either consists of one single point or does not contain its upper endpoint $\pi_{\sup}(H)$.

Proof. Let τ^* be a stopping time which is optimal with respect to a given $P^* \in \mathcal{P}$. Then $U_0^{P^*} = E^*[\, H_{\tau^*} \,] = \sup_{\tau'\in\mathcal{T}} E^*[\, H_{\tau'} \,]$, and consequently $U_0^{P^*} \in \Pi(H)$. Together with the a priori bounds (6.17), we obtain the inclusions

$$\{\, U_0^{P^*} \mid P^* \in \mathcal{P} \,\} \subset \Pi(H) \subset [a, b], \tag{6.20}$$

where

$$a := \sup_{\tau\in\mathcal{T}} \inf_{P^*\in\mathcal{P}} E^*[\, H_\tau \,] \quad \text{and} \quad b := \sup_{\tau\in\mathcal{T}} \sup_{P^*\in\mathcal{P}} E^*[\, H_\tau \,].$$

Moreover, the minimax identity (6.19) shows that

$$a = \inf_{P^* \in \mathcal{P}} \sup_{\tau \in \mathcal{T}} E^*[\, H_\tau \,] = \inf_{P^* \in \mathcal{P}} U_0^{P^*} \quad \text{and} \quad b = \sup_{P^* \in \mathcal{P}} U_0^{P^*}.$$

Together with (6.20), this yields the identification of $\pi_{\inf}(H)$ and $\pi_{\sup}(H)$ as a and b.

Now we claim that $\{\, U_0^{P^*} \mid P^* \in \mathcal{P} \,\}$ is an interval, which, in view of the preceding step, will prove that $\Pi(H)$ is also an interval. Take $P_0, P_1 \in \mathcal{P}$ and define $P_\alpha \in \mathcal{P}$ by $P_\alpha := \alpha P_1 + (1-\alpha) P_0$ for $0 \le \alpha \le 1$. By Theorem 6.20, $f(\alpha) := U_0^{P_\alpha}$ is the supremum of the affine functions

$$\alpha \mapsto E_\alpha[\, H_\tau \,] = \alpha E_1[\, H_\tau \,] + (1-\alpha) E_0[\, H_\tau \,], \quad \tau \in \mathcal{T}.$$

Thus, f is convex and lower semicontinuous on [0, 1], hence continuous; see part (a) of Proposition A.4. Since $\mathcal{P}$ is convex, this proves our claim.

It remains to exclude the possibility that b belongs to $\Pi(H)$ in case $a < b$. Suppose by way of contradiction that $b \in \Pi(H)$. Then there exist $\hat{\tau} \in \mathcal{T}$ and $\hat{P} \in \mathcal{P}$ such that

$$\hat{E}[\, H_{\hat{\tau}} \,] = b = \sup_{\tau \in \mathcal{T}} \sup_{P^* \in \mathcal{P}} E^*[\, H_\tau \,].$$

In particular, $\hat{P}$ attains the supremum of $E^*[\, H_{\hat{\tau}} \,]$ for $P^* \in \mathcal{P}$. Theorem 5.33 implies that the discounted European claim $H_{\hat{\tau}}$ is attainable and that $E^*[\, H_{\hat{\tau}} \,]$ is in fact independent of $P^* \in \mathcal{P}$. Hence,

$$b = \hat{E}[\, H_{\hat{\tau}} \,] = \inf_{P^* \in \mathcal{P}} E^*[\, H_{\hat{\tau}} \,] \le \sup_{\tau \in \mathcal{T}} \inf_{P^* \in \mathcal{P}} E^*[\, H_\tau \,],$$

and we end up with the contradiction $b \le a$. Thus, b cannot belong to $\Pi(H)$. □

Comparing the previous result with Theorem 5.33, one might wonder whether $\Pi(H)$ contains its lower bound if $\pi_{\inf}(H) < \pi_{\sup}(H)$. At a first glance, it may come as a surprise that both cases

$$\pi_{\inf}(H) \in \Pi(H) \quad \text{and} \quad \pi_{\inf}(H) \notin \Pi(H)$$

can occur, as is illustrated by the following simple example.

Example 6.34. Consider a complete market model with $T = 2$, defined on some probability space $(\Omega_0, \mathcal{F}_0, P_0)$. This model will be enlarged by adding two external states ω^+ and ω^-, i.e., we define $\Omega := \Omega_0 \times \{\omega^+, \omega^-\}$ and

$$P[\, \{(\omega_0, \omega^\pm)\} \,] := \frac{1}{2} P_0[\, \{\omega_0\} \,], \quad \omega_0 \in \Omega_0.$$

The enlarged financial market model will then be incomplete, and the corresponding set $\mathcal{P}$ of equivalent martingale measures is given by

$$\mathcal{P} = \{\, P_p^* \mid 0 < p < 1 \,\},$$

where P_p^* is determined by $P_p^*[\,\Omega_0 \times \{\omega^+\}\,] = p$. Consider the discounted American claim H defined as

$$H_0 \equiv 0, \quad H_1 \equiv 1, \quad \text{and} \quad H_2(\omega) := \begin{cases} 2 & \text{if } \omega = (\omega_0, \omega^+), \\ 0 & \text{if } \omega = (\omega_0, \omega^-). \end{cases}$$

Clearly, $\tau_2 \equiv 2$ is an optimal stopping time for P_p^* if $p > \frac{1}{2}$, while $\tau_1 \equiv 1$ is optimal for $p \le \frac{1}{2}$. Hence,

$$\Pi(H) = [1, 2),$$

and the lower bound $\pi_{\inf}(H) = 1$ is an arbitrage-free price for H. Now consider the discounted American claim $\widetilde{H}$ defined by $\widetilde{H}_t = H_t$ for $t = 0, 2$ and by $\widetilde{H}_1 \equiv 0$. In this case, we have

$$\Pi(\widetilde{H}) = (0, 2). \qquad \diamondsuit$$

Theorem 6.33 suggests that an American claim H which admits a unique arbitrage-free price should be attainable in an appropriate sense. Corollary 6.24, our hedging result in the case of a complete market, suggests the following definition of attainability.

Definition 6.35. A discounted American claim H is called *attainable* if there exists a stopping time $\tau \in \mathcal{T}$ and a self-financing trading strategy $\overline{\xi}$ whose value process V satisfies P-a.s.

$$V_t \ge H_t \text{ for all } t, \text{ and } V_\tau = H_\tau.$$

The trading strategy $\overline{\xi}$ is called a *hedging strategy* for H.

If H is attainable, then a hedging strategy protects the seller not only against those claims H_τ which arise from stopping times τ. The seller is on the safe side even if the buyer would have full knowledge of future prices and would exercise H at an arbitrary $\mathcal{F}_T$-measurable random time σ. For instance, the buyer even could choose σ such that

$$H_\sigma = \max_{0 \le t \le T} H_t.$$

In fact, we will see in Remark 7.12 that H is attainable in the sense of Definition 6.35 if and only if $V_t \ge H_t$ for all t and $V_\sigma = H_\sigma$ for some $\mathcal{F}_T$-measurable random time σ.

If the market model is *complete*, then every American claim H is attainable. Moreover, Theorem 6.12 and Corollary 6.24 imply that the minimal initial investment needed for the purchase of a hedging strategy for H is equal to the unique arbitrage-free price of H. In a general market model, every *attainable* discounted American claim H satisfies our integrability condition (6.18) and has a unique arbitrage-free price which is equal to the initial investment of a hedging strategy for H. This follows from Theorem 5.26. In fact, the converse implication is also true:

Theorem 6.36. *For a discounted American claim H satisfying* (6.18)*, the following conditions are equivalent:*

(a) *H is attainable.*

(b) *H admits a unique arbitrage-free price $\pi(H)$, i.e., $\Pi(H) = \{\pi(H)\}$.*

(c) $\pi_{\sup}(H) \in \Pi(H)$.

Moreover, if H is attainable, then $\pi(H)$ is equal to the initial investment of any hedging strategy for H.

The equivalence of (b) and (c) is an immediate consequence of Theorem 6.33. The remainder of the proof of Theorem 6.36 is postponed to Remark 7.10 because it requires the technique of superhedging, which will be introduced in Section 7.

6.4 Stability under pasting

In this section we define the pasting of two equivalent probability measures at a given stopping time. This operation will play an important role in the analysis of lower and upper Snell envelopes as developed in Section 6.5. In particular, we will prepare for the proof of the minimax identity (6.17), which was used in the characterization of arbitrage-free prices of an American contingent claim. Let us start with a few preparations.

Definition 6.37. Let τ be a stopping time. The σ-algebra of events which are observable up to time τ is defined as

$$\mathcal{F}_\tau := \Big\{ A \in \mathcal{F} \mid A \cap \{\tau \le t\} \in \mathcal{F}_t \text{ for all } t \Big\}.$$

It is straightforward to check that $\mathcal{F}_\tau$ is indeed a σ-algebra and that

$$\mathcal{F}_\tau = \Big\{ A \in \mathcal{F} \mid A \cap \{\tau = t\} \in \mathcal{F}_t \text{ for all } t \Big\}.$$

In particular, $\mathcal{F}_\tau$ coincides with $\mathcal{F}_t$ if $\tau \equiv t$. Moreover, if σ is a stopping time with $\sigma(\omega) \le \tau(\omega)$ for all $\omega \in \Omega$, then $\mathcal{F}_\sigma \subset \mathcal{F}_\tau$.

The following result is an addendum to Doob's stopping theorem; see Theorem 6.17:

Proposition 6.38. *For an adapted process M in $\mathcal{L}^1(Q)$ the following conditions are equivalent:*

(a) *M is a Q-martingale.*

(b) *$E_Q[\, M_\tau \mid \mathcal{F}_\sigma \,] = M_{\tau\wedge\sigma}$ for all $\tau \in \mathcal{T}$ and all stopping times σ.*

Proof. (a)$\Rightarrow$(b): Take a set $A \in \mathcal{F}_\sigma$ and let us write

$$E_Q[\,M_\tau;\ A\,] = E_Q\big[\,M_\tau;\ A \cap \{\tau \le \sigma\}\,\big] + E_Q\big[\,M_\tau;\ A \cap \{\tau > \sigma\}\,\big].$$

Condition (b) will follow if we may replace M_τ by M_σ in the rightmost expectation. To this end, note that

$$A \cap \{\sigma = t\} \cap \{\tau > \sigma\} = A \cap \{\sigma = t\} \cap \{\tau > t\} \in \mathcal{F}_t.$$

Thus, since the stopped process M^τ is a martingale by Theorem 6.17,

$$\begin{aligned} E_Q\big[\,M_\tau;\ A \cap \{\tau > \sigma\}\,\big] &= \sum_{t=0}^{T} E_Q\big[\,M_T^\tau;\ A \cap \{\sigma = t\} \cap \{\tau > \sigma\}\,\big] \\ &= \sum_{t=0}^{T} E_Q\big[\,M_t^\tau;\ A \cap \{\sigma = t\} \cap \{\tau > \sigma\}\,\big] \\ &= E_Q\big[\,M_\sigma;\ A \cap \{\tau > \sigma\}\,\big]. \end{aligned}$$

(b)$\Rightarrow$(a): This follows by taking $\tau \equiv t$ and $\sigma \equiv s \le t$. □

Example 6.39. Let Z be the density process of a probability measure $\widetilde{Q}$ that is absolutely continuous with respect to Q; see Remark 5.13. If σ is a stopping time, then we have $\widetilde{Q} \ll Q$ on $\mathcal{F}_\sigma$, and Propositions A.11 and 6.38 show that the corresponding density is given by

$$\left.\frac{d\widetilde{Q}}{dQ}\right|_{\mathcal{F}_\sigma} = E_Q[\,Z_T \mid \mathcal{F}_\sigma\,] = Z_{T\wedge\sigma}. \qquad \diamond$$

We have also the following extension of Theorem 6.20. It provides the solution to the optimal stopping problem posed at any stopping time $\tau \le T$.

Proposition 6.40. *Let H be an adapted process in $\mathcal{L}^1(\Omega, \mathcal{F}, Q)$, and define for $\tau \in \mathcal{T}$*

$$\mathcal{T}_\tau := \{\,\sigma \in \mathcal{T} \mid \sigma \ge \tau\,\}.$$

Then the Snell envelope U^Q of H satisfies Q-a.s.

$$U_\tau^Q = \operatorname*{ess\,sup}_{\sigma \in \mathcal{T}_\tau} E_Q[\,H_\sigma \mid \mathcal{F}_\tau\,],$$

and the essential supremum is attained for

$$\sigma_{\min}^{(\tau)} := \min\{\,t \ge \tau \mid H_t = U_t^Q\,\}.$$

Proof. It is not difficult to reduce the assertion to Theorem 6.20 by using the identity

$$E_Q[\,H_\sigma \mid \mathcal{F}_\tau\,] = E_Q[\,H_\sigma \mid \mathcal{F}_t\,] \quad Q\text{-a.s. on } \{\tau = t\}, \tag{6.21}$$

which is proved by checking the defining properties of the conditional expectation. □

Definition 6.41. Let Q_1 and Q_2 be two equivalent probability measures and take $\sigma \in \mathcal{T}$. The probability measure

$$\widetilde{Q}[\,A\,] := E_{Q_1}\big[\,Q_2[\,A \mid \mathcal{F}_\sigma\,]\big], \quad A \in \mathcal{F}_T,$$

is called the *pasting of Q_1 and Q_2 in σ*.

The monotone convergence theorem for conditional expectations guarantees that $\widetilde{Q}$ is indeed a probability measure and that

$$E_{\widetilde{Q}}[\,Y\,] = E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,]\big]$$

for all $\mathcal{F}_T$-measurable $Y \ge 0$. Note that $\widetilde{Q}$ coincides with Q_1 on $\mathcal{F}_\sigma$, i.e.,

$$E_{\widetilde{Q}}[\,Y\,] = E_{Q_1}[\,Y\,] \quad \text{for all } \mathcal{F}_\sigma\text{-measurable } Y \ge 0.$$

Lemma 6.42. *For $Q_1 \approx Q_2$, their pasting in $\sigma \in \mathcal{T}$ is equivalent to Q_1 and satisfies*

$$\frac{d\widetilde{Q}}{dQ_1} = \frac{Z_T}{Z_\sigma},$$

where Z is the density process of Q_2 with respect to Q_1.

Proof. For $Y \ge 0$,

$$\begin{aligned} E_{\widetilde{Q}}[\,Y\,] &= E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,]\big] \\ &= E_{Q_1}\Big[\,\frac{1}{Z_\sigma} E_{Q_1}[\,Y Z_T \mid \mathcal{F}_\sigma\,]\Big] \\ &= E_{Q_1}\Big[\,\frac{Z_T}{Z_\sigma}\,Y\,\Big], \end{aligned}$$

where we have used the martingale property of Z and the fact that $Z_\sigma > 0$ Q_1-almost surely. The equivalence of $\widetilde{Q}$ and Q_1 follows from $Z_T > 0$ Q_1-almost surely. □

Lemma 6.43. *For $Q_1 \approx Q_2$, let $\widetilde{Q}$ be their pasting in $\sigma \in \mathcal{T}$. Then, for all stopping times τ and $\mathcal{F}_T$-measurable $Y \ge 0$,*

$$E_{\widetilde{Q}}[\,Y \mid \mathcal{F}_\tau\,] = E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_{\sigma\vee\tau}\,] \mid \mathcal{F}_\tau\,\big].$$

Proof. If $\varphi \ge 0$ is $\mathcal{F}_\tau$-measurable, then $\varphi I_{\{\tau \le \sigma\}}$ is $\mathcal{F}_\sigma \cap \mathcal{F}_\tau$-measurable. Hence,

$$\begin{aligned} E_{\widetilde{Q}}[\,Y\varphi;\ \tau \le \sigma\,] &= E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,]\varphi;\ \tau \le \sigma\,\big] \\ &= E_{Q_1}\big[\,E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,] \mid \mathcal{F}_\tau\,\big]\varphi;\ \tau \le \sigma\,\big] \\ &= E_{\widetilde{Q}}\big[\,E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,] \mid \mathcal{F}_\tau\,\big]\varphi;\ \tau \le \sigma\,\big], \end{aligned}$$

where we have used the fact that $\widetilde{Q}$ coincides with Q_1 on $\mathcal{F}_\sigma$. On the other hand,

$$\begin{aligned} E_{\widetilde{Q}}[\,Y\varphi;\ \tau > \sigma\,] &= E_{Q_1}\big[\,E_{Q_2}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\tau\,]\varphi \mid \mathcal{F}_\sigma\,\big];\ \tau > \sigma\,\big] \\ &= E_{\widetilde{Q}}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\tau\,]\varphi;\ \tau > \sigma\,\big]. \end{aligned}$$

It follows that

$$E_{\widetilde{Q}}[\,Y \mid \mathcal{F}_\tau\,] = E_{Q_1}\big[\,E_{Q_2}[\,Y \mid \mathcal{F}_\sigma\,] \mid \mathcal{F}_\tau\,\big]\mathrm{I}_{\{\tau\le\sigma\}} + E_{Q_2}[\,Y \mid \mathcal{F}_\tau\,]\mathrm{I}_{\{\tau>\sigma\}},$$

and this coincides with the right-hand side of the asserted identity. □

Definition 6.44. A set $\mathcal{Q}$ of equivalent probability measures on $(\Omega, \mathcal{F})$ is called *stable* if, for any $Q_1, Q_2 \in \mathcal{Q}$ and $\sigma \in \mathcal{T}$, also their pasting in σ is contained in $\mathcal{Q}$.

The condition of stability in the preceding definition is sometimes also called *fork convexity*, *m-stability*, or *stability under pasting*. For the purposes of this book, the most important example of a stable set is the class $\mathcal{P}$ of all equivalent martingale measures, but in Section 6.5 we will also discuss the connection between stable sets and dynamic risk measures.

Proposition 6.45. *$\mathcal{P}$ is stable.*

Proof. Take $P_1, P_2 \in \mathcal{P}$ and denote by $\widetilde{P}$ their pasting a given $\sigma \in \mathcal{T}$. Doob's stopping theorem in the form of Proposition 6.38 and Lemma 6.43 applied with $Y := X_t^i \ge 0$ and $\tau \equiv s$ yield that

$$\widetilde{E}[\,X_t \mid \mathcal{F}_s\,] = E_1\big[\,E_2[\,X_t \mid \mathcal{F}_{\sigma\vee s}\,] \mid \mathcal{F}_s\,\big] = E_1[\,X_{\sigma\vee s} \mid \mathcal{F}_s\,] = X_s.$$

It follows in particular that each component X_t^i is in $\mathcal{L}^1(\widetilde{P})$ since $\widetilde{E}[\,X_t^i\,] = X_0^i < \infty$, concluding the proof of $\widetilde{P} \in \mathcal{P}$. □

6.5 Lower and upper Snell envelopes

Our main goal in this section is to provide a proof of the minimax identity (6.19), that was used in the characterization of the set of arbitrage-free prices of an American contingent claim. The techniques and results which we develop here will help to characterize the time-consistency of dynamic coherent risk measures and they will also be needed in Chapter 7. Moreover, they can be interpreted in terms of an optimal stopping problem for general utility functionals which appear in a robust Savage representation of preferences on payoff profiles. Let us now fix a set $\mathcal{Q}$ of equivalent probability measures and an adapted process H such that

$$H_t \in \mathcal{L}^1(Q) \text{ for all } t \text{ and each } Q \in \mathcal{Q}.$$

Recall that this condition implies

$$\inf_{Q\in\mathcal{Q}} \sup_{\tau\in\mathcal{T}} E_Q[\, H_\tau \,] = \inf_{Q\in\mathcal{Q}} U_0^Q < \infty,$$

where U^Q denotes the Snell envelope of H with respect to $Q \in \mathcal{Q}$. Let us also assume that

$$\mathcal{Q} \text{ is stable.}$$

Definition 6.46. The *lower Snell envelope* of H is defined as

$$U_t^{\downarrow} := \operatorname*{ess\,inf}_{Q\in\mathcal{Q}} U_t^Q = \operatorname*{ess\,inf}_{Q\in\mathcal{Q}} \operatorname*{ess\,sup}_{\tau\in\mathcal{T}_t} E_Q[\, H_\tau \mid \mathcal{F}_t \,], \quad t = 0, \dots, T.$$

The *upper Snell envelope* of H is defined as

$$U_t^{\uparrow} := \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} U_t^Q = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}_t} \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\, H_\tau \mid \mathcal{F}_t \,], \quad t = 0, \dots, T.$$

We will first study the lower Snell envelope. The following "minimax theorem" states that the essential infimum and the essential supremum occurring in the definition of $U^{\downarrow}$ may be interchanged if $\mathcal{Q}$ is stable. Applied at $t = 0$ and combined with Proposition 6.45, this gives the identity (6.19), which was used in our characterization of the arbitrage-free prices of H.

Theorem 6.47. *The lower Snell envelope of H satisfies*

$$U_t^{\downarrow} = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}_t} \operatorname*{ess\,inf}_{Q\in\mathcal{Q}} E_Q[\, H_\tau \mid \mathcal{F}_t \,] \quad \textit{for each } t. \tag{6.22}$$

In particular,

$$U_0^{\downarrow} = \inf_{Q\in\mathcal{Q}} \sup_{\tau\in\mathcal{T}} E_Q[\, H_\tau \,] = \sup_{\tau\in\mathcal{T}} \inf_{Q\in\mathcal{Q}} E_Q[\, H_\tau \,].$$

The inequality $\geq$ in (6.22) is obvious. Its converse is an immediate consequence of the next theorem, which solves the following optimal stopping problem that is formulated with respect to the *nonadditive expectation operator* $\inf_{Q\in\mathcal{Q}} E_Q[\,\cdot\,]$:

$$\text{maximize } \inf_{Q\in\mathcal{Q}} E_Q[\, H_\tau \,] \text{ among all } \tau \in \mathcal{T}.$$

Theorem 6.48. *Define a stopping time $\tau_t \in \mathcal{T}_t$ by*

$$\tau_t := \min\{\, u \geq t \mid U_u^{\downarrow} = H_u \,\}.$$

Then, P-a.s.,

$$U_t^{\downarrow} = \operatorname*{ess\,inf}_{Q\in\mathcal{Q}} E_Q[\, H_{\tau_t} \mid \mathcal{F}_t \,]. \tag{6.23}$$

In particular,

$$\sup_{\tau\in\mathcal{T}} \inf_{Q\in\mathcal{Q}} E_Q[\, H_\tau \,] = \inf_{Q\in\mathcal{Q}} E_Q[\, H_{\tau_0} \,] = U_0^{\downarrow}.$$

For the proof of Theorem 6.48, we need some preparations.

Lemma 6.49. *Suppose that we are given $Q_1, Q_2 \in \mathcal{Q}$, a stopping time $\tau \in \mathcal{T}$, and a set $B \in \mathcal{F}_\tau$. Let $\widetilde{Q} \in \mathcal{Q}$ be the pasting of Q_1 and Q_2 in the stopping time*

$$\sigma := \tau\, \mathrm{I}_B + T\, \mathrm{I}_{B^c}.$$

Then the Snell envelopes associated with these three measures are related as follows:

$$U_\tau^{\widetilde{Q}} = U_\tau^{Q_2}\, \mathrm{I}_B + U_\tau^{Q_1}\, \mathrm{I}_{B^c} \quad P\text{-a.s.} \tag{6.24}$$

Proof. With Proposition 6.40 and its notation, we have

$$U_\tau^{\widetilde{Q}} = \operatorname*{ess\,sup}_{\rho \in \mathcal{T}_\tau} E_{\widetilde{Q}}[\, H_\rho \mid \mathcal{F}_\tau \,].$$

To compute the conditional expectation on the right, note first that

$$E_{Q_2}[\, H_\rho \mid \mathcal{F}_{\sigma \vee \tau} \,] = E_{Q_2}[\, H_\rho \mid \mathcal{F}_\tau \,]\, \mathrm{I}_B + H_\rho\, \mathrm{I}_{B^c}.$$

Hence, Lemma 6.43 yields that

$$E_{\widetilde{Q}}[\, H_\rho \mid \mathcal{F}_\tau \,] = E_{Q_2}[\, H_\rho \mid \mathcal{F}_\tau \,]\, \mathrm{I}_B + E_{Q_1}[\, H_\rho \mid \mathcal{F}_\tau \,]\, \mathrm{I}_{B^c}.$$

Moreover, whenever $\rho_1, \rho_2 \in \mathcal{T}_\tau$, then

$$\widetilde{\rho} := \rho_1\, \mathrm{I}_B + \rho_2\, \mathrm{I}_{B^c}$$

is also a stopping time in $\mathcal{T}_\tau$. Thus,

$$U_\tau^{\widetilde{Q}} = \operatorname*{ess\,sup}_{\rho \in \mathcal{T}_\tau} E_{Q_2}[\, H_\rho \mid \mathcal{F}_\tau \,]\, \mathrm{I}_B + \operatorname*{ess\,sup}_{\rho \in \mathcal{T}_\tau} E_{Q_1}[\, H_\rho \mid \mathcal{F}_\tau \,]\, \mathrm{I}_{B^c},$$

and (6.24) follows. □

Lemma 6.50. *For any $Q \in \mathcal{Q}$ and $\tau \in \mathcal{T}$ there exist $Q_k \in \mathcal{Q}$ such that $Q_k = Q$ on $\mathcal{F}_\tau$ and*

$$U_\tau^{Q_k} \searrow \operatorname*{ess\,inf}_{\hat{Q} \in \mathcal{Q}} U_\tau^{\hat{Q}} = U_\tau^{\downarrow}.$$

Similarly, there exist $Q^k \in \mathcal{Q}$ such that $Q^k = Q$ on $\mathcal{F}_\tau$ and

$$U_\tau^{Q^k} \nearrow \operatorname*{ess\,sup}_{\hat{Q} \in \mathcal{Q}} U_\tau^{\hat{Q}} =: U_\tau^{\uparrow}.$$

Proof. For $Q_1, Q_2 \in \mathcal{Q}$, $B := \{\, U_\tau^{Q_1} > U_\tau^{Q_2} \,\}$, take $\widetilde{Q} \in \mathcal{Q}$ as in Lemma 6.49. Then

$$U_\tau^{\widetilde{Q}} = U_\tau^{Q_1} \cdot \mathrm{I}_{B^c} + U_\tau^{Q_2} \cdot \mathrm{I}_B = U_\tau^{Q_1} \wedge U_\tau^{Q_2}. \tag{6.25}$$

Moreover, if $Q_1 = Q$ on $\mathcal{F}_\tau$ then also $\widetilde{Q} = Q$ on $\mathcal{F}_\tau$. Hence, the set

$$\Phi := \big\{ U_\tau^{\hat{Q}} \mid \hat{Q} \in \mathcal{Q} \text{ and } \hat{Q} = Q \text{ on } \mathcal{F}_\tau \big\}$$

is such that $U_\tau^\downarrow = \operatorname{ess\,inf} \Phi$. Moreover, (6.25) implies that Φ is directed downwards, and the second part of Theorem A.32 states the existence of the desired sequence $(Q_k) \subset \mathcal{Q}$. The proof for the essential supremum is analogous. □

Proof of Theorem 6.48. To prove (6.23), observe first that $U_t^Q \geq E_Q[\, H_{\tau_t} \mid \mathcal{F}_t \,]$ for each $Q \in \mathcal{Q}$, so that $\geq$ holds in (6.23). For the proof of the converse inequality, note that

$$\tau_t \leq \min\{\, u \geq t \mid U_u^Q = H_u \,\} =: \tau_t^Q \quad \text{for } Q \in \mathcal{Q}.$$

It was shown in Theorem 6.20 that τ_t^Q is the minimal optimal stopping time after time t and with respect to Q. It was also shown in the proof of Theorem 6.20 that the stopped process $(U^Q)^{\tau_t^Q}$ is a Q-martingale from time t on. In particular,

$$U_t^Q = E_Q[\, U_{\tau_t}^Q \mid \mathcal{F}_t \,] \quad \text{for all } Q \in \mathcal{Q}. \tag{6.26}$$

Let us now fix some $Q \in \mathcal{Q}$. Lemma 6.50 yields $Q_k \in \mathcal{Q}$ with $Q_k = Q$ on $\mathcal{F}_{\tau_t}$ such that $U_{\tau_t}^{Q_k}$ decreases to $U_{\tau_t}^\downarrow$. We obtain

$$\begin{aligned} E_Q[\, H_{\tau_t} \mid \mathcal{F}_t \,] &= E_Q[\, U_{\tau_t}^\downarrow \mid \mathcal{F}_t \,] = E_Q\Big[\lim_{k\uparrow\infty} U_{\tau_t}^{Q_k} \mid \mathcal{F}_t \Big] \\ &= \lim_{k\uparrow\infty} E_Q[\, U_{\tau_t}^{Q_k} \mid \mathcal{F}_t \,] = \lim_{k\uparrow\infty} E_{Q_k}[\, U_{\tau_t}^{Q_k} \mid \mathcal{F}_t \,] \\ &= \lim_{k\uparrow\infty} U_t^{Q_k} \geq U_t^\downarrow. \end{aligned}$$

Here we have used the facts that $H_{\tau_t} \leq U_{\tau_t}^{Q_k} \leq U_{\tau_t}^{Q_1}$ and $E_Q[\, |U_{\tau_t}^{Q_1}| \,] = E_{Q_1}[\, |U_{\tau_t}^{Q_1}| \,] < \infty$ together with dominated convergence in the third step, the fact that $Q_k = Q$ on $\mathcal{F}_{\tau_t} \supset \mathcal{F}_t$ in the fourth, and (6.26) in the fifth identity. □

Remark 6.51. Suppose the buyer of an American option uses a utility functional of the form

$$\inf_{Q\in\mathcal{Q}} E_Q[\, u(Z) \,],$$

where $\mathcal{Q}$ is a set of probability measures and u is a measurable function. This may be viewed as a robust Savage representation of a preference relation on discounted asset payoffs; see Section 2.5. Thus, the aim of the buyer is to maximize the utility

$$\inf_{Q\in\mathcal{Q}} E_Q[\, u(H_\tau) \,]$$

of the discounted payoff H_τ among all stopping times $\tau \in \mathcal{T}$. This generalized utility maximization problem can be solved with the results developed in this section, provided that the set $\mathcal{Q}$ is a *stable* set of equivalent probability measures. Indeed, assume

$$\widetilde{H}_t := u(H_t) \in \mathcal{L}^1(Q) \quad \text{for all } t \text{ and each } Q \in \mathcal{Q},$$

and let U^Q be the Snell envelope of $\widetilde{H}_t$ with respect to $Q \in \mathcal{Q}$. Theorem 6.48 states that the generalized optimal stopping problem is solved by the stopping time

$$\tau^* := \min\{\, t \geq 0 \mid \operatorname*{ess\,inf}_{Q \in \mathcal{Q}} U_t^Q = \widetilde{H}_t \,\},$$

i.e.,

$$\inf_{Q \in \mathcal{Q}} \sup_{\tau \in \mathcal{T}} E_Q[\, u(H_\tau)\,] = U_0^{\downarrow} = \inf_{Q \in \mathcal{Q}} E_Q[\, u(H_{\tau^*})\,]. \qquad \diamondsuit$$

Let us now turn to the analysis of the upper Snell envelope

$$U_t^{\uparrow} := \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} U_t^Q = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} E_Q[\, H_\tau \mid \mathcal{F}_t\,], \quad t = 0, \dots, T.$$

In order to simplify the presentation, we will assume from now on that

$$\sup_{Q \in \mathcal{Q}} E_Q[\, |H_t|\,] < \infty \quad \text{for all } t.$$

This condition implies that

$$U_0^{\uparrow} = \sup_{Q \in \mathcal{Q}} U_0^Q \leq \sup_{\tau \in \mathcal{T}} \sup_{Q \in \mathcal{Q}} E_Q[\, |H_\tau|\,] < \infty.$$

Our main result on upper Snell envelopes states that, for stable sets $\mathcal{Q}$, the upper Snell envelope $U^{\uparrow}$ satisfies a recursive scheme that is similar to the one for ordinary Snell envelopes. In contrast to (6.4), however, it involves the *nonadditive conditional expectation operators* $\operatorname{ess\,sup}_Q E_Q[\,\cdot \mid \mathcal{F}_t\,]$.

Theorem 6.52. *$U^{\uparrow}$ satisfies the following recursive scheme:*

$$U_T^{\uparrow} = H_T \;\; \textit{and} \;\; U_t^{\uparrow} = H_t \vee \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} E_Q[\, U_{t+1}^{\uparrow} \mid \mathcal{F}_t\,], \quad t = T-1 \dots, 0. \tag{6.27}$$

Proof. The definition of the Snell envelope U^Q implies that

$$U_t^{\uparrow} = \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} U_t^Q = H_t \vee \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} E_Q[\, U_{t+1}^Q \mid \mathcal{F}_t\,]. \tag{6.28}$$

Next, we fix $Q \in \mathcal{Q}$ and denote by $\mathcal{Q}_{t+1}(Q)$ the set of all $\hat{Q} \in \mathcal{Q}$ which coincide with Q on $\mathcal{F}_{t+1}$. According to Lemma 6.50, there are $Q^k \in \mathcal{Q}_{t+1}(Q)$ such that

$U^{Q^k}_{t+1} \nearrow U^{\uparrow}_{t+1}$. The fact that $E_Q[\,|U^{Q_1}_{t+1}|\,] = E_{Q_1}[\,|U^{Q_1}_{t+1}|\,] < \infty$ combined with monotone convergence for conditional expectations shows that

$$\begin{aligned}
\operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,U^{\uparrow}_{t+1} \mid \mathcal{F}_t\,] &\geq \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,U^{Q}_{t+1} \mid \mathcal{F}_t\,] \\
&= \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} \operatorname*{ess\,sup}_{\hat{Q}\in\mathcal{Q}_{t+1}(Q)} E_Q[\,U^{\hat{Q}}_{t+1} \mid \mathcal{F}_t\,] \\
&\geq \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} \liminf_{k\uparrow\infty} E_Q[\,U^{Q^k}_{t+1} \mid \mathcal{F}_t\,] \\
&= \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,U^{\uparrow}_{t+1} \mid \mathcal{F}_t\,].
\end{aligned} \tag{6.29}$$

In particular, all inequalities are in fact identities. Together with (6.28) we obtain the recursive scheme for $U^{\uparrow}$. □

The following result shows that the nonadditive conditional expectation operators $\operatorname{ess\,sup}_Q E_Q[\,\cdot \mid \mathcal{F}_t\,]$ associated with a stable set $\mathcal{Q}$ enjoy a consistency property that is similar to the martingale property for ordinary conditional expectations.

Theorem 6.53. *Let $\mathcal{Q}$ be a set of equivalent probability measures and*

$$V^{\uparrow}_t := \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T,$$

for some $\mathcal{F}_T$-measurable $H \geq 0$ such that $V^{\uparrow}_0 < \infty$. If $\mathcal{Q}$ is stable then

$$V^{\uparrow}_\sigma = \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,V^{\uparrow}_\tau \mid \mathcal{F}_\sigma\,] \quad \textit{for } \sigma, \tau \in \mathcal{T} \textit{ with } \sigma \leq \tau.$$

If, in addition, $\mathcal{Q}$ is convex and the set of densities $\{\,dQ/dP \mid Q \in \mathcal{Q}\,\}$ is closed in $L^1(P)$ for some equivalent reference measure P, then also the following converse implication holds: if for all $\tau \in \mathcal{T}$ and $H \in L^\infty(P)$,

$$\sup_{Q\in\mathcal{Q}} E_Q[\,H\,] = \sup_{Q\in\mathcal{Q}} E_Q\Big[\operatorname*{ess\,sup}_{Q'\in\mathcal{Q}} E_{Q'}[\,H \mid \mathcal{F}_\tau\,]\Big], \tag{6.30}$$

then $\mathcal{Q}$ is stable.

Remark 6.54. Note that, for H as in the theorem and $\tau \in \mathcal{T}$,

$$\begin{aligned}
V^{\uparrow}_\tau &= \sum_{t=0}^{T} \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,H \mid \mathcal{F}_t\,]\mathrm{I}_{\{\tau=t\}} \\
&= \sum_{t=0}^{T} \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,H \mid \mathcal{F}_\tau\,]\mathrm{I}_{\{\tau=t\}} \\
&= \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\,H \mid \mathcal{F}_\tau\,],
\end{aligned}$$

where we have used (6.21) in the second identity. ◇

Proof of Theorem 6.53. By Remark 6.54,

$$V_\sigma^\uparrow = \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\, H \mid \mathcal{F}_\sigma\,] = \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q\big[\, E_Q[\, H \mid \mathcal{F}_\tau\,] \,\big|\, \mathcal{F}_\sigma\,\big].$$

The proof that the right-hand side is equal to $\operatorname{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\, V_\tau^\uparrow \mid \mathcal{F}_\sigma\,]$ is done by first noting that $V^\uparrow$ is equal to the upper Snell envelope of the process H_t given by $H_T = H$ and $H_t = 0$ for $t < T$. Then the same argument as in (6.29) applies. All one has to do is to replace $t + 1$ by τ.

To prove the converse implication, let us assume by way of contradiction that the pasting $\widetilde{Q}$ of some measures $Q_1, Q_2 \in \mathcal{Q}$ in a stopping time $\tau \in \mathcal{T}$ does not belong to $\mathcal{Q}$. By the Hahn–Banach theorem, there exists some $H \in L^\infty(P)$ such that

$$\sup_{Q\in\mathcal{Q}} E_Q[\, H\,] < E_{\widetilde{Q}}[\, H\,]. \tag{6.31}$$

Using the definition of $\widetilde{Q}$ and our assumption (6.30), we obtain

$$\begin{aligned} E_{\widetilde{Q}}[\, H\,] &= E_{Q_1}\big[\, E_{Q_2}[\, H \mid \mathcal{F}_\tau\,]\big] \\ &\le E_{Q_1}\big[\, \operatorname*{ess\,sup}_{Q'\in\mathcal{Q}} E_{Q'}[\, H \mid \mathcal{F}_\tau\,]\big] \\ &\le \sup_{Q\in\mathcal{Q}} E_Q\big[\, \operatorname*{ess\,sup}_{Q'\in\mathcal{Q}} E_{Q'}[\, H \mid \mathcal{F}_\tau\,]\big] \\ &= \sup_{Q\in\mathcal{Q}} E_Q[\, H\,], \end{aligned}$$

in contradiction to (6.31). □

Let us conclude this section by pointing out the connection between stability under pasting and the time-consistency of dynamic coherent risk measures. Let

$$\rho(Y) := \sup_{Q\in\mathcal{Q}} E_Q[\, -Y\,], \quad Y \in L^\infty(P),$$

be a coherent risk measure on $L^\infty(P)$ defined in terms of a set $\mathcal{Q}$ of probability measures equivalent to P. In the context of a dynamic financial market model, it is natural to update the initial risk assessment at later times $t > 0$. If one continues to use $\mathcal{Q}$ as a basis to compute the risk but takes into account the available information, one is led to consider the *conditional risk measures*

$$\rho_t(Y) = \operatorname*{ess\,sup}_{Q\in\mathcal{Q}} E_Q[\, -Y \mid \mathcal{F}_t\,], \quad t = 0, \dots, T. \tag{6.32}$$

The sequence $\rho_0 \dots, \rho_T$ can be regarded as a *dynamic coherent risk measure*. Often, such a dynamic risk measure is called *time-consistent* or *dynamically consistent* if

$$\rho_s(-\rho_t(Y)) = \rho_s(Y) \quad \text{for } 0 \le s \le t \le T. \tag{6.33}$$

Corollary 6.55. *Let $\mathcal{Q}$ be a set of equivalent probability measures. The dynamic risk measure $\rho_0, \dots, \rho_T$ arising from $\mathcal{Q}$ via* (6.32) *is time-consistent as soon as $\mathcal{Q}$ is stable. Conversely, stability of $\mathcal{Q}$ is necessary if time-consistency is required also for stopping times and if $\mathcal{Q}$ coincides with $\mathcal{Q}_{\max}$, the maximal representing subset of $\mathcal{M}_1(P)$ for ρ.*

Proof. This follows from Theorem 6.53, since

$$\left\{ \frac{dQ}{dP} \,\Big|\, Q \in \mathcal{Q}_{\max} \right\} = \bigcap_{X \in \mathcal{A}_\rho} \left\{ \varphi \in L^1_+(P) \,|\, E[\,\varphi X\,] \geq 0,\ E[\,\varphi\,] = 1 \right\}$$

is convex and closed in $L^1(P)$. □

If one requires time-consistency only in the weak form (6.33), one can show by using the arguments in this section that it is necessary and sufficient for $\mathcal{Q}$ to be stable under pasting in stopping times that take at most one value different from T.

Chapter 7
Superhedging

The idea of superhedging is to find a self-financing trading strategy with minimal initial investment which covers any possible future obligation resulting from the sale of a contingent claim. If the contingent claim is not attainable, the proof of the existence of such a "superhedging strategy" requires new techniques, and in particular a new uniform version of the Doob decomposition. We will develop this theory for general American contingent claims. In doing so, we will also obtain new results for European contingent claims. In the first three sections of this chapter, we assume that our market model is arbitrage-free or, equivalently, that the set of equivalent martingale measures satisfies

$$\mathcal{P} \neq \emptyset.$$

In the final Section 7.4, we discuss liquid options in a setting where no probabilistic model is fixed a priori. Such options may be used for the construction of specific martingale measures, and also for the purpose of hedging illiquid exotic derivatives.

7.1 $\mathcal{P}$-supermartingales

In this section, H denotes a discounted American claim with

$$\sup_{P^* \in \mathcal{P}} E^*[\, H_t \,] < \infty \quad \text{for all } t. \tag{7.1}$$

Our aim in this chapter is to find the minimal amount of capital U_t that will be needed at time t in order to purchase a self-financing trading strategy whose value process satisfies $V_u \geq H_u$ for all $u \geq t$. In analogy to our derivation of the recursive scheme (6.4), we will now heuristically derive a formula for U_t. At time T, the minimal amount needed is clearly given by

$$U_T = H_T.$$

At time $T-1$, a first requirement is to have $U_{T-1} \geq H_{T-1}$. Moreover, the amount U_{T-1} must suffice to purchase an $\mathcal{F}_{T-1}$-measurable portfolio $\overline{\xi}_T$ such that $\overline{\xi}_T \cdot X_T \geq H_T$ almost surely. An informal application of Theorem 1.31, conditional on $\mathcal{F}_{T-1}$, shows that

$$U_{T-1} \geq \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, H_T \mid \mathcal{F}_{T-1} \,].$$

Hence, the minimal amount U_{T-1} is equal to the maximum of H_{T-1} and this essential supremum. An iteration of this argument yields the recursive scheme

$$U_T = H_T \quad \text{and} \quad U_t = H_t \vee \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, U_{t+1} \mid \mathcal{F}_t \,]$$

for $t = T-1, \dots, 0$. By combining Proposition 6.45 and Theorem 6.52, we can identify U as the *upper Snell envelope*

$$U_t^{\uparrow} = \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} U_t^{P^*} = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, H_\tau \mid \mathcal{F}_t \,]$$

of H with respect to the stable set $\mathcal{P}$, where U^{P^*} denotes the Snell envelope of H with respect to P^*. In the first three sections of this chapter, we will in particular give a rigorous version of the heuristic argument above.

Note first that condition (7.1) implies that

$$\pi_{\sup}(H) = \sup_{P^* \in \mathcal{P}} U_0^{P^*} = \sup_{P^* \in \mathcal{P}} \sup_{\tau \in \mathcal{T}} E^*[\, H_\tau \,] < \infty,$$

where we have used the identification of the upper bound $\pi_{\sup}(H)$ of the arbitrage-free prices of H given in Theorem 6.33. It will turn out that the following definition applies to the upper Snell envelope if we choose $\mathcal{Q} = \mathcal{P}$.

Definition 7.1. Suppose that $\mathcal{Q}$ is a non-empty set of probability measures on $(\Omega, \mathcal{F}_T)$. An adapted process is called a $\mathcal{Q}$*-supermartingale* if it is a supermartingale with respect to each $Q \in \mathcal{Q}$. Analogously, we define the notions of a $\mathcal{Q}$*-submartingale* and of a $\mathcal{Q}$*-martingale*.

In Theorem 5.26, we have already encountered an example of a $\mathcal{P}$-martingale, namely the value process of the replicating strategy of an attainable discounted European claim.

Theorem 7.2. *The upper Snell envelope $U^{\uparrow}$ of H is the smallest $\mathcal{P}$-supermartingale that dominates H.*

Proof. For each $P^* \in \mathcal{P}$ the recursive scheme (6.27) implies that P^*-a.s.

$$U_t^{\uparrow} \ge H_t \vee E^*[\, U_{t+1}^{\uparrow} \mid \mathcal{F}_t \,] \ge E^*[\, U_{t+1}^{\uparrow} \mid \mathcal{F}_t \,].$$

Since $U_0^{\uparrow}$ is a finite constant due to our integrability assumption (7.1), induction on t shows that $U_t^{\uparrow}$ is integrable with respect to each $P^* \in \mathcal{P}$ and hence is a $\mathcal{P}$-supermartingale dominating H.

If $\widetilde{U}$ is another $\mathcal{P}$-supermartingale which dominates H, then $\widetilde{U}_T \ge H_T = U_T^{\uparrow}$. Moreover, if $\widetilde{U}_{t+1} \ge U_{t+1}^{\uparrow}$ for some t, then

$$\widetilde{U}_t \ge H_t \vee E^*[\, \widetilde{U}_{t+1} \mid \mathcal{F}_t \,] \ge H_t \vee E^*[\, U_{t+1}^{\uparrow} \mid \mathcal{F}_t \,].$$

Thus,

$$\widetilde{U}_t \geq H_t \vee \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, U_{t+1}^{\uparrow} \mid \mathcal{F}_t \,] = U_t^{\uparrow},$$

and backward induction shows that $\widetilde{U}$ dominates $U^{\uparrow}$. □

For European claims, Theorem 7.2 takes the following form.

Corollary 7.3. *Let H^E be a discounted European claim such that*

$$\sup_{P^* \in \mathcal{P}} E^*[\, H^E \,] < \infty.$$

Then

$$V_t^{\uparrow} := \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, H^E \mid \mathcal{F}_t \,], \quad t = 0, \dots, T,$$

is the smallest $\mathcal{P}$-supermartingale whose terminal value dominates H^E.

Remark 7.4. Note that the proof of Theorem 7.2 did not use any special properties of the set $\mathcal{P}$. Thus, if $\mathcal{Q}$ is an arbitrary set of equivalent probability measures, the process U defined by the recursion

$$U_T = H_T \quad \text{and} \quad U_t = H_t \vee \operatorname*{ess\,sup}_{Q \in \mathcal{Q}} E_Q[\, U_{t+1} \mid \mathcal{F}_t \,]$$

is the smallest $\mathcal{Q}$-supermartingale dominating the adapted process H. ◇

7.2 Uniform Doob decomposition

The aim of this section is to give a complete characterization of all non-negative $\mathcal{P}$-supermartingales. It will turn out that an integrable and non-negative process U is a $\mathcal{P}$-supermartingale if and only if it can be written as the difference of a $\mathcal{P}$-martingale N and an increasing adapted process B satisfying $B_0 = 0$. This decomposition may be viewed as a uniform version of the Doob decomposition since it involves simultaneously the whole class $\mathcal{P}$. It will turn out that the $\mathcal{P}$-martingale N has a special structure: It can be written as a "stochastic integral" of the underlying process X, which defines the class $\mathcal{P}$. On the other hand, the increasing process B is only adapted, not predictable as in the Doob decomposition with respect to a single measure.

Theorem 7.5. *For an adapted, non-negative process U, the following two statements are equivalent.*

(a) *U is a $\mathcal{P}$-supermartingale.*

(b) *There exists an adapted increasing process B with $B_0 = 0$ and a d-dimensional predictable process ξ such that*

$$U_t = U_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) - B_t \quad P\text{-a.s. for all } t.$$

Proof. First, we prove the easier implication (b) ⇒(a). Fix $P^* \in \mathcal{P}$ and note that

$$V_T := U_0 + \sum_{k=1}^{T} \xi_k \cdot (X_k - X_{k-1}) \geq U_T \geq 0.$$

Hence, V is a $\mathcal{P}$-martingale by Theorem 5.15. It follows that $U_t \in \mathcal{L}^1(P^*)$ for all t. Moreover, for $P^* \in \mathcal{P}$

$$E^*[\, U_{t+1} \mid \mathcal{F}_t \,] = E^*[\, V_{t+1} - B_{t+1} \mid \mathcal{F}_t \,] \leq V_t - B_t = U_t,$$

and so U is a $\mathcal{P}$-supermartingale.

The proof of the implication (a) ⇒(b) is similar to the proof of Theorem 5.33. We must show that for any given $t \in \{1, \dots, T\}$, there exist $\xi_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$ and $R_t \in L^0_+(\Omega, \mathcal{F}_t, P)$ such that

$$U_t - U_{t-1} = \xi_t \cdot (X_t - X_{t-1}) - R_t.$$

This condition can be written as

$$U_t - U_{t-1} \in \mathcal{K}_t - L^0_+(\Omega, \mathcal{F}_t, P),$$

where $\mathcal{K}_t$ is as in (5.12). There is no loss of generality in assuming that P is itself a martingale measure. In this case, $U_t - U_{t-1}$ is contained in $L^1(\Omega, \mathcal{F}_t, P)$ by the definition of a $\mathcal{P}$-supermartingale. Assume that

$$U_t - U_{t-1} \notin \mathcal{C} := \big(\mathcal{K}_t - L^0_+(\Omega, \mathcal{F}_t, P)\big) \cap L^1(P).$$

Since our model is arbitrage-free, Lemma 1.67 implies that $\mathcal{C}$ is closed in $L^1(\Omega, \mathcal{F}_t, P)$. Hence, Theorem A.56 implies the existence of some $Z \in L^\infty(\Omega, \mathcal{F}_t, P)$ such that

$$\alpha := \sup_{W \in \mathcal{C}} E[\, Z\, W \,] < E[\, Z\,(U_t - U_{t-1}) \,] =: \delta < \infty. \tag{7.2}$$

In fact, we have $\alpha = 0$ since $\mathcal{C}$ is a cone containing the constant function 0. Lemma 1.57 implies that such a random variable Z must be non-negative and must satisfy

$$E[\, (X_t - X_{t-1})\, Z \mid \mathcal{F}_{t-1}] = 0. \tag{7.3}$$

In fact, we can always modify Z such that it is bounded below by some $\varepsilon > 0$ and still satisfies (7.2). To see this, note first that every $W \in \mathcal{C}$ is dominated by a term of the form $\xi_t \cdot (X_t - X_{t-1})$. Hence, our assumption $P \in \mathcal{P}$, the integrability of W, and an application of Fatou's lemma as yield that

$$E[\,W\,] \le E[\,\xi_t \cdot (X_t - X_{t-1})\,] \le \liminf_{c\uparrow\infty} E\big[\,\mathrm{I}_{\{|\xi_t|\le c\}}\xi_t \cdot (X_t - X_{t-1})\,\big] \le 0.$$

Thus, if we let $Z^\varepsilon := \varepsilon + Z$, then Z^ε also satisfies $E[\,Z^\varepsilon\,W\,] \le 0$ for all $W \in \mathcal{C}$. If we chose ε small enough, then $E[\,Z^\varepsilon\,(U_t - U_{t-1})\,]$ is still larger than 0; i.e., Z^ε also satisfies (7.2) and in turn (7.3). Therefore, we may assume from now on that our Z with (7.2) is bounded below by some constant $\varepsilon > 0$.

Let

$$Z_{t-1} := E[\,Z \mid \mathcal{F}_{t-1}\,],$$

and define a new measure $\widetilde{P} \approx P$ by

$$\frac{d\widetilde{P}}{dP} := \frac{Z}{Z_{t-1}}.$$

We claim that $\widetilde{P} \in \mathcal{P}$. To prove this, note first that $X_k \in L^1(\widetilde{P})$ for all k, because the density $d\widetilde{P}/dP$ is bounded. Next, let

$$\varphi_k := E\Big[\,\frac{Z}{Z_{t-1}}\,\Big|\,\mathcal{F}_k\,\Big], \qquad k = 0, \dots, T.$$

If $k \ne t$, then $\varphi_{k-1} = \varphi_k$; this is clear for $k > t$, and for $k < t$ it follows from

$$\varphi_k = E\Big[\,\frac{E[\,Z \mid \mathcal{F}_{t-1}\,]}{Z_{t-1}}\,\Big|\,\mathcal{F}_k\,\Big] = 1.$$

Thus, for $k \ne t$

$$\begin{aligned}
\widetilde{E}[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] &= \frac{1}{\varphi_{k-1}}\,E[\,(X_k - X_{k-1})\,\varphi_k \mid \mathcal{F}_{k-1}\,] \\
&= E[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] \\
&= 0.
\end{aligned}$$

If $k = t$, then (7.3) yields that

$$\widetilde{E}[\,X_k - X_{k-1} \mid \mathcal{F}_{k-1}\,] = \frac{1}{Z_{t-1}}\,E[\,(X_t - X_{t-1})\,Z \mid \mathcal{F}_{t-1}\,] = 0.$$

Hence $\widetilde{P} \in \mathcal{P}$.

Since $\widetilde{P} \in \mathcal{P}$, we have $\widetilde{E}[\,U_t - U_{t-1} \mid \mathcal{F}_{t-1}\,] \le 0$, and we get

$$\begin{aligned}
0 &\ge \widetilde{E}\big[\,\widetilde{E}[\,U_t - U_{t-1} \mid \mathcal{F}_{t-1}\,]\,Z_{t-1}\,\big] \\
&= \widetilde{E}[\,(U_t - U_{t-1})\,Z_{t-1}\,] \\
&= E[\,(U_t - U_{t-1})\,Z\,] \\
&= \delta.
\end{aligned}$$

This, however, contradicts the fact that $\delta > 0$. □

Remark 7.6. The decomposition in part (b) of Theorem 7.5 is sometimes called the *optional decomposition* of the $\mathcal{P}$-supermartingale U. The existence of such a decomposition was first proved by El Karoui and Quenez [80] and D. Kramkov [135] in a continuous-time framework where B is an "optional" process; this explains the terminology. ◇

7.3 Superhedging of American and European claims

Let H be a discounted American claim such that

$$\sup_{P^* \in \mathcal{P}} E^*[\, H_t \,] < \infty \quad \text{for all } t,$$

which is equivalent to the condition that the upper bound of the arbitrage-free prices of H is finite:

$$\pi_{\sup}(H) = \sup_{P^* \in \mathcal{P}} \sup_{\tau \in \mathcal{T}} E^*[\, H_\tau \,] < \infty.$$

Our aim in this section is to construct self-financing trading strategies such that the seller of H stays on the safe side in the sense that the corresponding portfolio value is always above H.

Definition 7.7. Any self-financing trading strategy $\overline{\xi}$ whose value process V satisfies

$$V_t \geq H_t \quad P\text{-a.s. for all } t$$

is called a *superhedging strategy* for H.

Sometimes, a superhedging strategy is also called a *superreplication strategy*. According to Definition 6.35, H is attainable if and only if there exist $\tau \in \mathcal{T}$ and a superhedging strategy whose value process satisfies $V_\tau = H_\tau$ P-almost surely.

Lemma 7.8. *If H is not attainable, then the value process V of any superhedging strategy satisfies*

$$P[\, V_t > H_t \text{ for all } t \,] > 0.$$

Proof. We introduce the stopping time

$$\tau := \inf\{\, t \geq 0 \mid H_t = V_t \,\}.$$

Then $P[\, \tau = \infty \,] = P[\, V_t > H_t \text{ for all } t \,]$. Suppose that $P[\, \tau = \infty \,] = 0$. In this case, $V_\tau = H_\tau$ P-a.s so that we arrive at the contradiction that H must be an attainable American claim. □

Let us now turn to the question whether superhedging strategies *exist*. In Section 6.1, we have already seen how one can use the Doob decomposition of the Snell envelope U^{P^*} of H together with the martingale representation of Theorem 5.39 in order to obtain a superhedging strategy for the price $U_0^{P^*}$, where P^* denotes the unique equivalent martingale measure in a *complete* market model. We have also seen that $U_0^{P^*}$ is the minimal amount for which such a superhedging strategy is available, and that $U_0^{P^*}$ is the unique arbitrage-free price of H. The same is true of any *attainable* American claim in an incomplete market model.

In the context of a *non-attainable* American claim H in an *incomplete* financial market model, the P^*-Snell envelope will be replaced with the upper Snell envelope $U^\uparrow$ of H. The uniform Doob decomposition will take over the roles played by the usual Doob decomposition and the martingale representation theorem. Since $U^\uparrow$ is a $\mathcal{P}$-supermartingale by Theorem 7.2, the uniform Doob decomposition states that $U^\uparrow$ takes the form

$$\begin{aligned} U_t^\uparrow &= U_0^\uparrow + \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}) - B_t \\ &\geq H_t \end{aligned} \tag{7.4}$$

for some predictable process ξ and some increasing process B. Thus, the self-financing trading strategy $\overline{\xi} = (\xi^0, \xi)$ defined by ξ and the initial capital

$$\overline{\xi}_1 \cdot \overline{X}_0 = U_0^\uparrow = \pi_{\sup}(H)$$

is a superhedging strategy for H. Moreover, if $\widetilde{V}$ is the value process of any superhedging strategy, then Lemma 7.8 implies that $\widetilde{V}_0 > E^*[\, H_\tau \,]$ for all $\tau \in \mathcal{T}$ and each $P^* \in \mathcal{P}$. In particular, $\widetilde{V}_0$ is larger than any arbitrage-free price for H, and it follows that $\widetilde{V}_0 \geq \pi_{\sup}(H)$. Thus, we have proved:

Corollary 7.9. *There exists a superhedging strategy with initial investment* $\pi_{\sup}(H)$*, and this is the minimal amount needed to implement a superhedging strategy.*

We will call $\pi_{\sup}(H)$ the *cost of superhedging* of H. Sometimes, a superhedging strategy is also called a *superreplication strategy*, and one says that $\pi_{\sup}(H)$ is the *cost of superreplication* or the *upper hedging price* of H. Recall, however, that $\pi_{\sup}(H)$ is typically not an arbitrage-free price for H. In particular, the seller cannot expect to receive the amount $\pi_{\sup}(H)$ for selling H.

On the other hand, the process B in the decomposition (7.4) can be interpreted as a *refunding scheme*: Using the superhedging strategy $\overline{\xi}$, the seller may withdraw successively the amounts defined by the increments of B. With this capital flow, the hedging portfolio at time t has the value $U_t^\uparrow \geq H_t$. Thus, the seller is on the safe side at no matter when the buyer decides to exercise the option. As we are going to show in Theorem 7.13 below, this procedure is optimal in the sense that, if started at any time t, it requires a minimal amount of capital.

Remark 7.10. Suppose $\pi_{\sup}(H)$ belongs to the set $\Pi(H)$ of arbitrage-free prices for H. By Theorem 6.33, this holds if and only if $\pi_{\sup}(H)$ is the only element of $\Pi(H)$.

In this case, the definition of $\Pi(H)$ yields a stopping time $\tau \in \mathcal{T}$ and some $P^* \in \mathcal{P}$ such that

$$\pi_{\sup}(H) = E^*[\, H_\tau \,].$$

Now let V be the value process of a superhedging strategy bought at $V_0 = \pi_{\sup}(H)$. It follows that $E^*[\, V_\tau \,] = \pi_{\sup}(H)$. Hence, $V_\tau = H_\tau$ P-a.s., so that H is attainable in the sense of Definition 6.35. This observation completes the proof of Theorem 6.36. ◇

Remark 7.11. If the American claim H is not attainable, then $\pi_{\sup}(H)$ is not an arbitrage-free price of H. Thus, one may expect the existence of arbitrage opportunities if H would be traded at the price $\pi_{\sup}(H)$. Indeed, selling H for $\pi_{\sup}(H)$ and buying a superhedging strategy $\overline{\xi}$ creates such an arbitrage opportunity: The balance at $t = 0$ is zero, but Lemma 7.8 implies that the value process V of $\overline{\xi}$ cannot be reached by any exercise strategy σ, i.e., we always have

$$V_\sigma \geq H_\sigma \quad \text{and} \quad P[\, V_\sigma > H_\sigma \,] > 0. \tag{7.5}$$

Note that (7.5) is not limited to exercise strategies which are stopping times but holds for arbitrary $\mathcal{F}_T$-measurable random times $\sigma : \Omega \to \{0, \dots, T\}$. In other words, $\pi_{\sup}(H)$ is too expensive even if the buyer of H would have full information about the future price evolution. ◇

Remark 7.12. The argument of Remark 7.11 implies that an American claim H is attainable if and only if there exists an $\mathcal{F}_T$-measurable random time $\sigma : \Omega \to \{0, \dots, T\}$ such that $H_\sigma = V_\sigma$, where V the value process of a superhedging strategy. In other words, the notion of attainability of American claims does not need the restriction to stopping times. ◇

We already know that $\pi_{\sup}(H)$ is the smallest amount for which one can buy a superhedging strategy at time 0. The following "superhedging duality theorem" extends this result to times $t > 0$. To this end, denote by $\mathcal{U}_t^{\uparrow}(H)$ the set of all $\mathcal{F}_t$-measurable random variables $\widetilde{U}_t \geq 0$ for which there exists a d-dimensional predictable process $\widetilde{\xi}$ such that

$$\widetilde{U}_t + \sum_{k=t+1}^{u} \widetilde{\xi}_k \cdot (X_k - X_{k-1}) \geq H_u \quad \text{for all } u \geq t \ P\text{-a.s.} \tag{7.6}$$

Theorem 7.13. *The upper Snell envelope $U_t^{\uparrow}$ of H is the minimal element of $\mathcal{U}_t^{\uparrow}(H)$. More precisely:*

(a) $U_t^{\uparrow} \in \mathcal{U}_t^{\uparrow}(H)$,

(b) $U_t^{\uparrow} = \operatorname{ess\,inf}\ \mathcal{U}_t^{\uparrow}(H)$.

Proof. Assertion (a) follows immediately from the uniform Doob decomposition of the $\mathcal{P}$-supermartingale $U^\uparrow$. As to part (b), we clearly get $U_t^\uparrow \geq \operatorname{ess\,inf} \mathcal{U}_t^\uparrow(H)$ from (a). For the proof of the converse inequality, take $\widetilde{U}_t \in \mathcal{U}_t^\uparrow(H)$ and choose a predictable process $\widetilde{\xi}$ for which (7.6) holds. We must show that the set $B := \{ U_t^\uparrow \leq \widetilde{U}_t \}$ satisfies $P[\,B\,] = 1$. Let

$$\hat{U}_t := U_t^\uparrow \wedge \widetilde{U}_t = U_t^\uparrow \cdot \mathrm{I}_B + \widetilde{U}_t \cdot \mathrm{I}_{B^c}.$$

Then $\hat{U}_t \leq U_t^\uparrow$, and our claim will follow if we can show that $U_t^\uparrow \leq \hat{U}_t$. Let ξ denote the predictable process obtained from the uniform Doob decomposition of the $\mathcal{P}$-supermartingale $U^\uparrow$, and define

$$\hat{\xi}_s := \begin{cases} \xi_s & \text{if } s \leq t, \\ \xi_s \cdot \mathrm{I}_B + \widetilde{\xi}_s \cdot \mathrm{I}_{B^c} & \text{if } s > t. \end{cases}$$

With this choice, $\hat{U}_t$ satisfies (7.6), i.e., $\hat{U}_t \in \mathcal{U}_t^\uparrow(H)$. Let

$$\hat{V}_s := U_0^\uparrow + \sum_{k=1}^{s} \hat{\xi}_k \cdot (X_k - X_{k-1}).$$

Then $\hat{V}_s \geq U_s^\uparrow$ for all $s \leq t$. In particular $\hat{V}_t \geq \hat{U}_t$, and hence $\hat{V}_T \geq H_T$, which implies that $\hat{V}$ is a $\mathcal{P}$-martingale; see Theorem 5.26. Hence, Doob's stopping theorem implies

$$\begin{aligned} U_t^\uparrow &= \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E^*[\, H_\tau \mid \mathcal{F}_t \,] \\ &\leq \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} \operatorname*{ess\,sup}_{\tau \in \mathcal{T}} E^*\Big[\, \hat{U}_t + \sum_{k=t+1}^{\tau} \hat{\xi}_k \cdot (X_k - X_{k-1}) \;\Big|\; \mathcal{F}_t \,\Big] \\ &= \hat{U}_t, \end{aligned}$$

which concludes the proof. □

We now take the point of view of the buyer of the American claim H. The buyer allocates an initial investment π to purchase H, and then receives the amount $H_\tau \geq 0$. The objective is to find an exercise strategy and a self-financing trading strategy $\overline{\eta}$ with initial investment $-\pi$, such that the portfolio value is covered by the payoff of the claim. In other words, find $\tau \in \mathcal{T}$ and a self-financing trading strategy with value process V such that $V_0 = -\pi$ and $V_\tau + H_\tau \geq 0$. As shown below, the maximal π for which this is possible is equal to

$$\pi_{\text{inf}}(H) = \sup_{\tau \in \mathcal{T}} \inf_{P^* \in \mathcal{P}} E^*[\, H_\tau \,] = \inf_{P^* \in \mathcal{P}} \sup_{\tau \in \mathcal{T}} E^*[\, H_\tau \,] = U_0^\downarrow,$$

where

$$\begin{aligned} U_t^{\downarrow} &= \operatorname*{ess\,inf}_{P^* \in \mathcal{P}} U_t^{P^*} \\ &= \operatorname*{ess\,inf}_{P^* \in \mathcal{P}} \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E^*[\, H_\tau \mid \mathcal{F}_t \,] \\ &= \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} \operatorname*{ess\,inf}_{P^* \in \mathcal{P}} E^*[\, H_\tau \mid \mathcal{F}_t \,] \end{aligned}$$

is the lower Snell envelope of H with respect to the stable set $\mathcal{P}$. More generally, we will consider the buyer's problem for arbitrary $t \geq 0$. To this end, denote by $\mathcal{U}_t^{\downarrow}(H)$ the set of all $\mathcal{F}_t$-measurable random variables $\widetilde{U}_t \geq 0$ for which there exists a d-dimensional predictable process $\widetilde{\eta}$ and a stopping time $\sigma \in \mathcal{T}_t$ such that

$$\widetilde{U}_t - \sum_{k=t+1}^{\sigma} \widetilde{\eta}_k \cdot (X_k - X_{k-1}) \leq H_\sigma \qquad P\text{-a.s.}$$

Theorem 7.14. *$U_t^{\downarrow}$ is the maximal element of $\mathcal{U}_t^{\downarrow}(H)$. More precisely:*

(a) $U_t^{\downarrow} \in \mathcal{U}_t^{\downarrow}(H)$,

(b) $U_t^{\downarrow} = \operatorname{ess\,sup}\, \mathcal{U}_t^{\downarrow}(H)$.

Proof. (a): Let $\overline{\xi}$ be a superhedging strategy for H with initial investment $\pi_{\sup}(H)$, and denote by V the value process of $\overline{\xi}$. The main idea of the proof is to use that $V_t - H_t \geq 0$ can be regarded as a new discounted American claim, to which we can apply Theorem 7.13. However, we must take care of the basic asymmetry of the hedging problem for American options: The seller of H must hedge against *all* possible exercise strategies, while the buyer must find only *one* suitable stopping time. It will turn out that a suitable stopping time is given by $\tau_t := \inf\{u \geq t \mid U_u^{\downarrow} = H_u\}$. With this choice, let us define a modified discounted American claim $\widetilde{H}$ by

$$\widetilde{H}_u = (V_u - H_u) \cdot \mathrm{I}_{\{u=\tau_t\}}, \quad u = 0, \dots, T.$$

Clearly $\widetilde{H}_\sigma \leq \widetilde{H}_{\tau_t}$ for all $\sigma \in \mathcal{T}_t$. It follows that

$$\begin{aligned} \operatorname*{ess\,sup}_{\sigma \in \mathcal{T}_t} \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, \widetilde{H}_\sigma \mid \mathcal{F}_t \,] &= \operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, \widetilde{H}_{\tau_t} \mid \mathcal{F}_t \,] \\ &= V_t - \operatorname*{ess\,inf}_{P^* \in \mathcal{P}} E^*[\, H_{\tau_t} \mid \mathcal{F}_t \,] \\ &= V_t - U_t^{\downarrow}, \end{aligned}$$

where we have used that V is a $\mathcal{P}$-martingale in the second and Theorem 6.48 in the third step. Thus, $V_t - U_t^{\downarrow}$ is equal to the upper Snell envelope $\widetilde{U}^{\uparrow}$ of $\widetilde{H}$ at time t. Let $\widetilde{\xi}$ be the d-dimensional predictable process obtained from the uniform Doob decomposition of $\widetilde{U}^{\uparrow}$. Then, due to part (a) of Theorem 7.13,

$$V_t - U_t^{\downarrow} + \sum_{k=t+1}^{u} \widetilde{\xi}_k \cdot (X_k - X_{k-1}) \geq \widetilde{H}_u = (V_u - H_u) \cdot \mathrm{I}_{\{u=\tau_t\}} \quad \text{for all } u \geq t.$$

Thus, $\eta := \widetilde{\xi} - \xi$ is as desired.

(b): Part (a) implies the inequality $\leq$ in (b). To prove its converse, take $\widetilde{U}_t \in \mathcal{U}_t^{\downarrow}$, a d-dimensional predictable process $\widetilde{\eta}$, and $\sigma \in \mathcal{T}_t$ such that

$$\widetilde{U}_t - \sum_{k=t+1}^{\sigma} \widetilde{\eta}_k \cdot (X_k - X_{k-1}) \leq H_\sigma \qquad P\text{-a.s.}$$

We will show below that

$$E^*\Big[\sum_{k=t+1}^{\sigma} \widetilde{\eta}_k \cdot (X_k - X_{k-1}) \,\Big|\, \mathcal{F}_t \Big] = 0 \quad \text{for all } P^* \in \mathcal{P}. \tag{7.7}$$

Given this fact, we obtain that

$$\widetilde{U}_t \leq E^*[\, H_\sigma \mid \mathcal{F}_t \,] \leq \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E^*[\, H_\tau \mid \mathcal{F}_t \,]$$

for all $P^* \in \mathcal{P}$. Taking the essential infimum over $P^* \in \mathcal{P}$ thus yields $\widetilde{U}_t \leq U_t^{\downarrow}$ and in turn (b).

To prove (7.7), let

$$\widetilde{G}_s := \mathrm{I}_{\{s \geq t+1\}} \sum_{k=t+1}^{s} \mathrm{I}_{\{k \leq \sigma\}} \widetilde{\eta}_k \cdot (X_k - X_{k-1}), \qquad s = 0, \dots, T.$$

Then $\widetilde{G}_T \geq \widetilde{U}_t - H_\sigma \geq -H_\sigma \in \mathcal{L}^1(P^*)$ for all P^*, and Theorem 5.15 implies that $\widetilde{G}$ is a $\mathcal{P}$-martingale. Hence (7.7) follows. □

We conclude this section by stating explicitly the corresponding results for European claims. Recall from Remark 6.7 that every discounted European claim H^E can be regarded as the discounted American claim. Therefore, the results we have obtained so far include the corresponding "European" counterparts as special cases.

Corollary 7.15. *For any discounted European claim H^E such that*

$$\sup_{P^* \in \mathcal{P}} E^*[\, H^E \,] < \infty,$$

there exist two d-dimensional predictable processes ξ and η such that P-a.s.

$$\operatorname*{ess\,sup}_{P^*\in\mathcal{P}} E^*[\,H^E \mid \mathcal{F}_t\,] + \sum_{k=t+1}^{T} \xi_k \cdot (X_k - X_{k-1}) \geq H^E, \tag{7.8}$$

$$\operatorname*{ess\,inf}_{P^*\in\mathcal{P}} E^*[\,H^E \mid \mathcal{F}_t\,] - \sum_{k=t+1}^{T} \eta_k \cdot (X_k - X_{k-1}) \leq H^E. \tag{7.9}$$

Remark 7.16. For $t = 0$, (7.8) takes the form

$$\sup_{P^*\in\mathcal{P}} E^*[\,H^E\,] + \sum_{k=1}^{T} \xi_k \cdot (X_k - X_{k-1}) \geq H^E \quad P\text{-a.s.}$$

Thus, the self-financing trading strategy $\overline{\xi}$ arising from ξ and the initial investment $\overline{\xi}_1 \cdot \overline{X}_0 = \sup_{P^*\in\mathcal{P}} E^*[\,H^E\,]$ allows the seller to cover all possible obligations without any downside risk. Similarly, (7.9) yields an interpretation of the self-financing trading strategy $\overline{\eta}$ which arises from η and the initial investment $\overline{\eta}_1 \cdot \overline{X}_0 = -\inf_{P^*\in\mathcal{P}} E^*[\,H^E\,]$. The latter quantity corresponds to the largest loan the buyer can take out and still be sure that, by using the trading strategy $\overline{\eta}$, this debt will be covered by the payoff H^E. ◇

Remark 7.17. Let H be a discounted European claim such that

$$\sup_{P^*\in\mathcal{P}} E^*[\,H^E\,] < \infty.$$

Suppose that $\hat{P} \in \mathcal{P}$ is such that $\hat{E}[\,H\,] = \pi_{\sup}(H)$. If $\overline{\xi} = (\xi^0, \xi)$ is a superhedging strategy for H, then

$$\hat{H} := \hat{E}[\,H\,] + \sum_{k=1}^{T} \xi_k \cdot (X_k - X_{k-1})$$

satisfies $\hat{H} \geq H \geq 0$. Hence, $\hat{H}$ is an attainable discounted claim, and it follows from Theorem 5.26 that

$$\hat{E}[\,\hat{H}\,] = \hat{E}[\,H\,].$$

This shows that $\hat{H}$ and H are identical and that H is attainable. We have thus obtained another proof of Theorem 5.33. ◇

As the last result in this section, we formulate the following "superhedging duality theorem", which states that the bounds in (7.8) and (7.9) are optimal.

Corollary 7.18. *Suppose that H^E is a discounted European claim with*

$$\sup_{P^*\in\mathcal{P}} E^*[\,H^E\,] < \infty.$$

Denote by $\mathcal{U}_t^{\uparrow}(H^E)$ *the set of all* $\mathcal{F}_t$ *measurable random variables* $\widetilde{U}_t$ *for which there exists a* d*-dimensional predictable process* $\widetilde{\xi}$ *such that*

$$\widetilde{U}_t + \sum_{k=t+1}^{T} \widetilde{\xi}_k \cdot (X_k - X_{k-1}) \geq H^E \quad P\text{-a.s.}$$

Then

$$\operatorname*{ess\,sup}_{P^* \in \mathcal{P}} E^*[\, H^E \mid \mathcal{F}_t \,] = \operatorname{ess\,inf} \mathcal{U}_t^{\uparrow}(H^E).$$

By $\mathcal{U}_t^{\downarrow}(H^E)$ *we denote the set of all* $\mathcal{F}_t$ *measurable random variables* $\widetilde{U}_t$ *for which there exists a* d*-dimensional predictable process* $\widetilde{\eta}$ *such that*

$$\widetilde{U}_t - \sum_{k=t+1}^{T} \widetilde{\eta}_k \cdot (X_k - X_{k-1}) \leq H^E \quad P\text{-a.s.}$$

Then

$$\operatorname*{ess\,inf}_{P^* \in \mathcal{P}} E^*[\, H^E \mid \mathcal{F}_t \,] = \operatorname{ess\,sup} \mathcal{U}_t^{\downarrow}(H^E).$$

Remark 7.19. Define $\mathcal{A}$ as the set of financial positions $Z \in L^\infty(\Omega, \mathcal{F}_T, P)$ which are acceptable in the sense that there exists a d-dimensional predictable process ξ such that

$$Z + \sum_{k=1}^{T} \xi_k \cdot (X_k - X_{k-1}) \geq 0 \quad P\text{-a.s.}$$

As in Section 4.8, this set $\mathcal{A}$ induces a coherent measure of risk ρ on $L^\infty(\Omega, \mathcal{F}_T, P)$:

$$\rho(Z) = \inf\{\, m \in \mathbb{R} \mid m + Z \in \mathcal{A} \,\}, \qquad Z \in L^\infty(\Omega, \mathcal{F}_T, P).$$

Corollary 7.18 implies that ρ can be represented as

$$\rho(Z) = \sup_{P^* \in \mathcal{P}} E^*[\, -Z \,].$$

We therefore obtain a multiperiod version of Proposition 4.93. ◇

Remark 7.20. Often, the superhedging strategy in a given incomplete model can be identified as the perfect hedge in an associated "extremal" model. As an example, consider a one-period model with d discounted risky assets given by bounded random variables $X^1, \dots, X^d$. Denote by μ the distribution of $X = (X^1, \dots, X^d)$ and by $\Gamma(\mu)$ the convex hull of the support of μ. The closure $K := \overline{\Gamma}(\mu)$ of $\Gamma(\mu)$ is convex and compact. We know from Section 1.5 that the model is arbitrage-free if and only if the price system $\pi = (\pi^1, \dots, \pi^d)$ is contained in the relative interior of $\Gamma(\mu)$, and the equivalent martingale measures can be identified with the measures $\mu^* \approx \mu$ with barycenter π. Consider a derivative $H = h(X)$ given by a *convex* function h on K. The cost of superhedging is given by

$$\sup_{\mu^*} \int h \, d\mu^* = \inf \left\{ \alpha(\pi) \mid \alpha \text{ affine on } K, \alpha \geq h \ \mu\text{-a.s.} \right\},$$

which is a special case of the duality result of Theorem 1.31. Since $\{\alpha \geq h\}$ is convex and closed, the condition $\mu(\alpha \geq h) = 1$ implies $\alpha \geq h$ on K. Denote by $\mathcal{M}(\pi)$ the class of all probability measures on K with barycenter π. For any affine function α with $\alpha \geq h$ on K, and for any $\widetilde{\mu} \in \mathcal{M}(\pi)$ we have

$$\int h \, d\widetilde{\mu} \leq \int \alpha \, d\widetilde{\mu} = \alpha(\pi).$$

Thus,

$$\hat{h}(\pi) = \sup_{\widetilde{\mu} \in \mathcal{M}(\pi)} \int h \, d\widetilde{\mu} \tag{7.10}$$

where we define for $f \in C(K)$

$$\hat{f} := \inf \left\{ \alpha \mid \alpha \text{ affine on } K, \alpha \geq f \ \mu\text{-a.s.} \right\}.$$

The supremum in (7.10) is attained since $\mathcal{M}(\pi)$ is weakly compact. More precisely, it is attained by any measure $\hat{\mu} \in \mathcal{M}(\pi)$ on K which is maximal with respect to the balayage order $\succcurlyeq_{\text{bal}}$ defined for measures on K as in (2.18); see Théorème X. 41 in [65]. But such a maximal measure is supported by the set of extreme points of the convex compact set K, i.e., by the *Choquet boundary* of K. This follows from a general integral representation theorem of Choquet; see, e.g., Théorème X. 43 of [65]. In our finite-dimensional setting, $\hat{\mu}$ can in fact be chosen to have a support consisting of at most $d + 1$ points, due to a theorem of Carathéodory and the representation of K as the convex hull of its extreme points; see [166], Theorems 17.1 and 18.5. But this means that $\hat{\mu}$ can be identified with a *complete* model, due to Proposition 1.39. Thus, the cost of superhedging $\hat{h}(\pi)$ can be identified with the canonical price

$$\hat{\pi} := \int h \, d\hat{\mu}$$

of the derivative H, computed in the complete model $\hat{\mu}$. Note that $\hat{\mu}$ sits on the Choquet boundary of $K = \overline{\Gamma}(\mu)$, but typically it will no longer be equivalent or absolutely continuous with respect to the original measure μ. As a simple illustration, consider a one-period model with one risky asset X^1. If X^1 is bounded, then the distribution μ of X^1 has bounded support, and $\Gamma(\mu)$ is of the form $[a, b]$. In this case, the cost of superhedging $H = h(X^1)$ for a convex function h is given by the price

$$p^* h(b) + (1 - p^*) h(a),$$

computed in the binary model in which X^1 takes only the values a and b, and where $p^* \in (0, 1)$ is determined by

$$p^* b + (1 - p^*) a = \pi^1. \qquad \diamond$$

The following example illustrates that a superhedging strategy is typically too expensive from a practical point of view. However, we will see in Chapter 8 how superhedging strategies can be used in order to construct other hedging strategies which are efficient in terms of cost and shortfall risk.

Example 7.21. Consider a simple one-period model where S_1^1 has under P a Poisson distribution and where $S^0 \equiv 1$. Let $H := (S_1^1 - K)^+$ be a call option with strike $K > 0$. We have seen in Example 1.37 that $\pi_{\text{inf}}(H)$ and $\pi_{\text{sup}}(H)$ coincide with the universal arbitrage bounds of Remark 1.36:

$$(S_0^1 - K)^+ = \pi_{\text{inf}}(H) \quad \text{and} \quad \pi_{\text{sup}}(H) = S_0^1.$$

Thus, the superhedging strategy for the seller consists in the trivial hedge of buying the asset at time 0, while the corresponding strategy for the buyer is a short-sale of the asset in case the option is in the money, i.e., if $S_0^1 > K$. ◇

7.4 Superhedging with liquid options

In practice, some derivatives such as put or call options are traded so frequently that their prices are quoted just like those of the primary assets. The prices of such *liquid options* can be regarded as an additional source of information on the expectations of the market as to the future evolution of asset prices. This information can be exploited in various ways. First, it serves to single out those martingale measures P^* which are compatible with the observed options prices, in the sense that the observed prices coincide with the expectations of the discounted payoff under P^*. Second, liquid options may be used as instruments for hedging more exotic options.

Our aim in this section is to illustrate these ideas in a simple setting. Assume that there is only one risky asset S^1 such that S_0^1 is a positive constant, and that S^0 is a riskless bond with interest rate $r = 0$. Thus, the discounted price process of the risky asset is given by $X_t = S_t^1 \geq 0$ for $t = 0, \dots, T$. As the underlying space of scenarios, we use the product space

$$\Omega := [0, \infty)^T.$$

We define $X_t(\omega) = x_t$ for $\omega = (x_1, \dots, x_T) \in \Omega$, and denote by $\mathcal{F}_t$ the σ-algebra generated by $X_0, \dots, X_t$; note that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. No probability measure P is given a priori. Let us now introduce a linear space $\mathcal{X}$ of $\mathcal{F}_T$-measurable functions as the smallest linear space such that the following conditions are satisfied:

(a) $1 \in \mathcal{X}$.

(b) $(X_t - X_s)\,\mathrm{I}_A \in \mathcal{X}$ for $0 \leq s < t \leq T$ and $A \in \mathcal{F}_s$.

(c) $(X_t - K)^+ \in \mathcal{X}$ for $K \geq 0$ and $t = 1, \dots, T$.

The functions in the space $\mathcal{X}$ will be interpreted as (discounted) payoffs of liquid derivatives. The constant 1 in (a) corresponds to a unit investment into the riskless bond. The function $X_t - X_s$ in (b) corresponds to the payoff of a forward contract on the risky asset, issued at time s for the price X_s and expiring at time t. The decision to buy such a forward contract at time s may depend on the market situation at time s; this is taken into account by allowing for payoffs $(X_t - X_s)\,\mathrm{I}_A$ with $A \in \mathcal{F}_s$. Linearity of $\mathcal{X}$ together with conditions (a) and (b) implies that

$$X_t \in \mathcal{X} \qquad \text{for all } t.$$

Finally, condition (c) states that call options with any possible strike and any maturity up to time T can be used as liquid securities.

Suppose that a linear pricing rule Φ is given on $\mathcal{X}$. The value $\Phi(Y)$ will be interpreted as the market price of the liquid security $Y \in \mathcal{X}$. The price of a liquid call option with strike K and maturity t will be denoted by

$$C_t(K) := \Phi\big((X_t - K)^+\big).$$

Assumption 7.22. *We assume that $\Phi : \mathcal{X} \to \mathbb{R}$ is a linear functional which satisfies the following conditions.*

(a) $\Phi(1) = 1$.

(b) $\Phi(Y) \geq 0$ *if* $Y \geq 0$.

(c) $\Phi\big((X_t - X_s)\,\mathrm{I}_A\big) = 0$ *for all* $0 \leq s < t \leq T$ *and* $A \in \mathcal{F}_s$.

(d) $C_t(K) = \Phi\big((X_t - K)^+\big) \to 0$ *as* $K \uparrow \infty$ *for all* t.

The first two conditions must clearly be satisfied if the pricing rule Φ shall not create arbitrage opportunities. Condition (c) states that X_s is the fair price for a forward contract issued at time s. This condition is quite natural in view of Theorem 5.30. In our present setting, it can also be justified by the following simple replication argument. At time s, take out a loan $X_s(\omega)$ and use it for buying the asset. At time t, the asset is worth $X_t(\omega)$ and the loan must be paid back, which results in a balance $X_t(\omega) - X_s(\omega)$. Since this investment strategy requires zero initial capital, the price of the corresponding payoff should also be zero. The continuity condition (d) is also quite natural.

Our first goal is to show that any such pricing rule Φ is compatible with the paradigm that arbitrage-free prices can be identified as expectations with respect to some martingale measure for X. More precisely, we are going to construct a martingale measure P^* such that $\Phi(Y) = E^*[\,Y\,]$ for all $Y \in \mathcal{X}$. On the one hand, this will imply regularity properties of Φ. On the other hand, this will yield an extension of our pricing rule Φ to a larger space of payoffs including path-dependent exotic options. As a first step in this direction, we have the following result.

Lemma 7.23. *For each t, there exists a unique probability measure μ_t on* $[0,\infty)$ *such that for all* $K \geq 0$

$$C_t(K) = \Phi\big((X_t - K)^+\big) = \int (x - K)^+ \,\mu_t(dx).$$

In particular, μ_t has the mean

$$\int x\,\mu_t(dx) = X_0.$$

Proof. Since $K \mapsto (X_t - K)^+$ is convex and decreasing, linearity and positivity of Φ imply that the function $\phi_t(K) := \Phi\big((X_t - K)^+\big)$ is convex and decreasing as well. Hence, there exists a decreasing right-continuous function $f : [0,\infty) \to [0,\infty)$ such that

$$\begin{aligned} C_t(K) &= C_t(0) - \int_0^K f(x)\,dx \\ &= X_0 - \int_0^K f(x)\,dx, \end{aligned}$$

i.e., $-f(K)$ is equal to the right-hand derivative of $C_t(K)$ at K. Our fourth condition on Φ yields

$$\int_0^\infty f(x)\,dx = X_0,$$

so that $f(x) \searrow 0$ as $x \uparrow \infty$. Hence, there exists a positive measure μ_t on $(0,\infty)$ such that

$$f(x) = \mu_t\big((x,\infty)\big) \quad \text{for } x > 0.$$

Fubini's theorem implies

$$\int_{(0,\infty)} x\,\mu_t(dx) = \int_0^\infty f(y)\,dy = X_0$$

and

$$\begin{aligned} C_t(K) &= X_0 - \int_0^K \int_{(0,\infty)} \mathrm{I}_{\{y<x\}}\,\mu_t(dx)\,dy \\ &= \int_{(0,\infty)} (x - K)^+\,\mu_t(dx). \end{aligned}$$

It remains to show that μ_t can be extended to a probability measure on $[0,\infty)$, i.e., we must show that $\mu_t\big((0,\infty)\big) \leq 1$. To this end, we will use the "put-call parity"

$$C_t(K) = X_0 - K + \Phi\big((K - X_t)^+\big),$$

which follows from our assumptions on Φ. Thus,

$$\Phi\big((K - X_t)^+\big) = \int_0^K g(x)\,dx,$$

where $g(x) = 1 - f(x)$. Since $K \mapsto \Phi\big((K - X_t)^+\big)$ is increasing, g must be non-negative, and we obtain $1 \geq f(0) = \mu_t\big((0,\infty)\big)$. □

The following lemma shows that the measures μ_t constructed in Lemma 7.23 are related to each other by the *balayage order* $\succcurlyeq_{\text{bal}}$, defined by

$$\mu \succcurlyeq_{\text{bal}} \nu \quad \Longleftrightarrow \quad \int f\,d\mu \geq \int f\,d\nu \quad \text{for all convex functions } f$$

for probability measures with finite expectation; see Remark 2.65.

Lemma 7.24. *The map $t \mapsto \mu_t$ is increasing with respect to the balayage order $\succcurlyeq_{\text{bal}}$:*

$$\mu_{t+1} \succcurlyeq_{\text{bal}} \mu_t \quad \textit{for all } t.$$

Proof. Note that

$$\begin{aligned}
(X_{t+1} - K)^+ &- (X_{t+1} - X_t)\,\mathrm{I}_{\{X_t > K\}} - (X_t - K)^+ \\
&= (X_{t+1} - K)^+ - (X_{t+1} - K)^+\,\mathrm{I}_{\{X_t > K\}} \\
&\geq 0.
\end{aligned}$$

Since the price of the forward contract $(X_{t+1} - X_t)\,\mathrm{I}_{\{X_t > K\}}$ vanishes under our pricing rule Φ, we must have that for all $K \geq 0$

$$\begin{aligned}
\int (x - K)^+\,\mu_{t+1}(dx) &- \int (x - K)^+\,\mu_t(dx) \\
&= C_{t+1}(K) - C_t(K) \\
&\geq 0.
\end{aligned}$$

An application of Corollary 2.62 concludes the proof. □

Let us introduce the class

$$\mathcal{P}_\Phi = \big\{\, P^* \in \mathcal{M}_1(\Omega, \mathcal{F}) \mid E^*[\,Y\,] = \Phi(Y) \text{ for all } Y \in \mathcal{X} \,\big\}$$

of all probability measures P^* on $(\Omega, \mathcal{F})$ which coincide with Φ on $\mathcal{X}$. Note that for any $P^* \in \mathcal{P}_\Phi$,

$$E^*[\,(X_t - X_s)\,\mathrm{I}_A\,] = 0 \quad \text{for } s < t \text{ and } A \in \mathcal{F}_s,$$

so that $\mathcal{P}_\Phi$ consists of *martingale measures* for X. Our first main result in this section can be regarded as a version of the "fundamental theorem of asset pricing" without an a priori measure P.

Theorem 7.25. *Under Assumption* 7.22*, the class* $\mathcal{P}_\Phi$ *is non-empty. Moreover, there exists* $P^* \in \mathcal{P}_\Phi$ *with the* Markov property: *For* $0 \le s \le t \le T$ *and each bounded measurable function* f*,*

$$E^*[\, f(X_t) \mid \mathcal{F}_s\,] = E^*[\, f(X_t) \mid X_s\,].$$

Proof. Since $\mu_{t+1} \succcurlyeq_{\mathrm{bal}} \mu_t$, Corollary 2.62 yields the existence of stochastic kernel Q_{t+1} such that $\int y\, Q_{t+1}(x, dy) = x$ and $\mu_{t+1} = \mu_t Q_{t+1}$. Let us define

$$P^* := \mu_1 \otimes Q_2 \otimes \cdots \otimes Q_T,$$

i.e., for each measurable set $A \subset \Omega = [0, \infty)^T$

$$P^*[\, A\,] = \int \mu_1(dx_1) \int Q_2(x_1, dx_2) \dots \int Q_T(x_{T-1}, dx_T)\, \mathrm{I}_A(x_1, x_2, \dots, x_T).$$

Clearly, μ_t is the law of X_t under P^*. In particular, all call options are priced correctly by calculating their expectation with respect to P^*. Then one checks that

$$E^*[\, f(X_{t+1}) \mid \mathcal{F}_t\,] = \int f(y)\, Q_{t+1}(X_t, dy) = E^*[\, f(X_{t+1}) \mid X_t\,] \qquad P^*\text{-a.s.} \tag{7.11}$$

The first identity above implies $E^*[\, X_{t+1} - X_t \mid \mathcal{F}_t\,] = 0$. In particular, P^* is a martingale measure, and the expectation of $(X_t - X_s)\,\mathrm{I}_A$ vanishes for $s < t$ and $A \in \mathcal{F}_s$. It follows that $E^*[\, Y\,] = \Phi(Y)$ for all $Y \in \mathcal{X}$. Finally, an induction argument applied to (7.11) yields the Markov property. □

So far, we have assumed that our space $\mathcal{X}$ of liquidly traded derivatives contains call options with *all* possible strike prices and maturities. From now on, we will simplify our setting by assuming that only call options with maturity T are liquidly traded. Thus, we replace $\mathcal{X}$ by the smaller space $\mathcal{X}_T$ which is defined as the linear hull of the constants, of all forward contracts

$$(X_t - X_s)\,\mathrm{I}_A, \quad 0 \le s < t \le T,\ A \in \mathcal{F}_s,$$

and of all call options

$$(X_T - K)^+, \qquad K \ge 0,$$

with maturity T. The observed market prices of derivatives in $\mathcal{X}_T$ are as before modeled by a linear pricing rule

$$\Phi_T : \mathcal{X}_T \to \mathbb{R}.$$

We assume that Φ_T satisfies Assumption 7.22 in the sense that condition (d) is only required for $t = T$:

(d′) $C_T(K) := \Phi_T\big((X_T - K)^+\big) \to 0$ as $K \uparrow \infty$.

By

$$\mathcal{P}_{\Phi_T} = \left\{ P^* \in \mathcal{M}_1(\Omega, \mathcal{F}) \mid E^*[\, Y \,] = \Phi_T(Y) \text{ for all } Y \in \mathcal{X}_T \right\}$$

we denote the class of all probability measures P^* on $(\Omega, \mathcal{F})$ which coincide with Φ_T on $\mathcal{X}_T$. As before, it follows from condition (c) of Assumption 7.22 that any $P^* \in \mathcal{P}_{\Phi_T}$ will be a *martingale measure* for the price process X. Obviously, any linear pricing rule Φ which is defined on the full space $\mathcal{X}$ and which satisfies Assumption 7.22 can be restricted to $\mathcal{X}_T$, and this restriction satisfies the above assumptions. Thus, we have $\mathcal{P}_{\Phi_T} \supset \mathcal{P}_\Phi \neq \emptyset$.

Proposition 7.26. *Under the above assumptions, $\mathcal{P}_{\Phi_T}$ is non-empty.*

Proof. Let μ_T be the measure constructed in Lemma 7.23 from the call prices with maturity T. Now consider the measure $\widetilde{P}$ on $(\Omega, \mathcal{F})$ defined as

$$\widetilde{P} := \delta_{X_0} \otimes \cdots \otimes \delta_{X_0} \otimes \mu_T,$$

i.e., under $\widetilde{P}$ we have $X_t = X_0$ $\widetilde{P}$-a.s. for $t < T$, and the law of X_T is μ_T. Clearly, we have $\widetilde{P} \in \mathcal{P}_{\Phi_T}$. □

A measure $P^* \in \mathcal{P}_{\Phi_T}$ can be regarded as an extension of the pricing rule Φ_T to the larger space $\mathcal{L}^1(P^*)$, and the expectation $E^*[\, H \,]$ of some European claim $H \geq 0$ can be regarded as an *arbitrage-free price* for H. Our aim is to obtain upper and lower bounds for $E^*[\, H \,]$ which hold simultaneously for all $P^* \in \mathcal{P}_\Phi$. We will derive such bounds for various exotic options; this will amount to the construction of certain superhedging strategies in terms of liquid securities.

As a first example, we consider the following *digital option*

$$H^{\text{dig}} := \begin{cases} 1 & \text{if } \max\limits_{0 \le t \le T} X_t \geq B \\ 0 & \text{otherwise,} \end{cases}$$

which has a unit payoff if the price processes reaches a given upper barrier $B > X_0$. If we denote by

$$\tau_B := \inf\{\, t \geq 0 \mid X_t \geq B \,\}$$

the first hitting time of the barrier B, then the payoff of the digital option can also be described as

$$H^{\text{dig}} = \mathrm{I}_{\{\tau_B \le T\}}.$$

For simplicity, we will assume from now on that

$$C_T(B) > 0,$$

so that in particular $\mu_T\big((B, \infty)\big) > 0$.

Theorem 7.27. *The following upper bound on the arbitrage-free prices of the digital option holds:*

$$\max_{P^*\in\mathcal{P}_{\Phi_T}} E^*[\, H^{\mathrm{dig}}\,] = \min_{0\le K<B} \frac{C_T(K)}{B-K}. \tag{7.12}$$

Proof. For $0 \le K < B$, we have $X_{\tau_B} \ge B$. Hence,

$$H^{\mathrm{dig}} = \mathrm{I}_{\{\tau_B\le T\}} \le \frac{(X_T-K)^+}{B-K} + \frac{X_{\tau_B}-X_T}{B-K}\,\mathrm{I}_{\{\tau_B\le T\}}.$$

Taking expectations with respect to some $P^* \in \mathcal{P}_{\Phi_T}$ yields

$$E^*[\, H^{\mathrm{dig}}\,] \le \frac{C_T(K)}{B-K} + \frac{1}{B-K} E^*[\,(X_{\tau_B}-X_T)\,\mathrm{I}_{\{\tau_B\le T\}}\,].$$

Since P^* is a martingale measure, the stopping theorem in the form of Proposition 6.38 implies

$$E^*[\, X_T\,\mathrm{I}_{\{\tau_B\le T\}}\,] = E^*[\, X_{\tau_B}\,\mathrm{I}_{\{\tau_B\le T\}}\,].$$

This shows that

$$\sup_{P^*\in\mathcal{P}_{\Phi_T}} E^*[\, H^{\mathrm{dig}}\,] \le \inf_{0\le K<B} \frac{C_T(K)}{B-K}.$$

The proof will be completed by Lemmas 7.28 and 7.29 below. □

Lemma 7.28. *If we let*

$$\lambda := 1 - \inf_{0\le K<B} \frac{C_T(K)}{B-K} \in (0,1),$$

then the infimum on the right-hand side is attained in K if and only if K belongs to the set of λ-quantiles for μ_T, i.e., if and only if

$$\mu_T([0,K)) \le \lambda \le \mu_T([0,K]).$$

In particular, it is attained in

$$K^* := \inf\big\{\, K \mid \mu_T([0,K]) \ge \lambda \,\big\}.$$

Proof. The convex function C_T has left- and right-hand derivatives

$$(C_T)'_-(K) = -\mu_T\big([K,\infty)\big) \quad \text{and} \quad (C_T)'_+(K) = -\mu_T\big((K,\infty)\big)\,;$$

see also Proposition A.4. Thus, the function $g(K) := C_T(K)/(B-K)$ has a minimum in K if and only if its left- and right-hand derivatives satisfy

$$g'_-(K) \le 0 \quad \text{and} \quad g'_+(K) \ge 0.$$

By computing g'_- and g'_+, one sees that these two conditions are equivalent to the requirement that K is a λ-quantile for μ_T; see Lemma A.15. □

Lemma 7.29. *There exists a martingale measure $\hat{P} \in \mathcal{P}_{\Phi_T}$ such that*

$$\hat{P}[\,\tau_B \le T\,] = \min_{0 \le K < B} \frac{C_T(K)}{B-K}.$$

Moreover, $\hat{P}$ can be taken such that

$$\tau_B = T-1 \quad \textit{and} \quad X_{\tau_B} = B \quad \hat{P}\textit{-a.s. on } \{\tau_B \le T\}$$

and

$$\{\, X_T > K^* \,\} \subset \{\, \tau_B \le T \,\} \subset \{\, X_T \ge K^* \,\} \quad \textit{modulo } \hat{P}\textit{-nullsets,}$$

where K^* is as in Lemma 7.28.

Proof. Let λ be as in Lemma 7.28, and let

$$q(t) := q_\mu^-(t) = \inf\left\{\, K \mid \mu_T([0,K]) \ge t \,\right\}$$

be the lower quantile function for μ; see A.3. We take an auxiliary probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P})$ supporting a random variable U which is uniformly distributed on $(0,1)$. By Lemma A.19, $\widetilde{X}_T := q(U)$ has distribution μ_T under $\widetilde{P}$. Let γ be such that

$$X_0 = \gamma\lambda + B(1-\lambda).$$

Since $B > X_0$ we have $0 \le \gamma < X_0$. We define $\widetilde{X}_{T-1}$ by

$$\widetilde{X}_{T-1} := \gamma\, \mathrm{I}_{\{U \le \lambda\}} + B\, \mathrm{I}_{\{U > \lambda\}},$$

and we let $\widetilde{X}_t := X_0$ for $0 \le t \le T-2$.

We now prove that $\widetilde{X}$ is a martingale with respect to its natural filtration $\widetilde{\mathcal{F}}_t := \sigma(\widetilde{X}_0, \dots, \widetilde{X}_t)$. To this end, note first that $\widetilde{\mathcal{F}}_{T-2} = \{\emptyset, \widetilde{\Omega}\}$, and hence

$$\widetilde{E}[\,\widetilde{X}_{T-1} \mid \widetilde{\mathcal{F}}_{T-2}\,] = \widetilde{E}[\,\widetilde{X}_{T-1}\,] = X_0 = \widetilde{X}_{T-2}.$$

Furthermore, since $K^* = q(\lambda)$,

$$\begin{aligned}
\widetilde{E}[\,\widetilde{X}_T;\ \widetilde{X}_{T-1} = B\,] &= \widetilde{E}[\,\widetilde{X}_T;\ U > \lambda\,] \\
&= \widetilde{E}[\,(\widetilde{X}_T - K^*)^+\,] + K^*\,\widetilde{P}[\,U > \lambda\,] \\
&= C_T(K^*) + K^*\,(1-\lambda) \\
&= (1-\lambda)(B-K^*) + K^*\,(1-\lambda) \\
&= B \cdot \widetilde{P}[\,\widetilde{X}_{T-1} = B\,].
\end{aligned}$$

Hence,

$$\begin{aligned}
\widetilde{E}[\,\widetilde{X}_T;\ \widetilde{X}_{T-1} = \gamma\,] &= \widetilde{E}[\,\widetilde{X}_T\,] - \widetilde{E}[\,\widetilde{X}_T;\ \widetilde{X}_{T-1} = B\,] \\
&= X_0 - B \cdot \widetilde{P}[\,\widetilde{X}_{T-1} = B\,] \\
&= \gamma\,\widetilde{P}[\,\widetilde{X}_{T-1} = \gamma\,],
\end{aligned}$$

It follows that
$$\widetilde{E}[\,\widetilde{X}_T \mid \widetilde{\mathcal{F}}_{T-1}\,] = \widetilde{E}[\,\widetilde{X}_T \mid \widetilde{X}_{T-1}\,] = \widetilde{X}_{T-1},$$
and so $\widetilde{X}$ is indeed a martingale.

As the next step, we note that
$$\begin{aligned}\{\,\widetilde{X}_T \geq B\,\} &\subset \{\,\widetilde{X}_T > K^*\,\}\\ &\subset \{\,U > \lambda\,\} = \{\,\widetilde{X}_{T-1} = B\,\}\\ &\subset \{\,U \geq \lambda\,\} \subset \{\,\widetilde{X}_T \geq K^*\,\},\end{aligned}$$
where we have used the fact that $K^* = q(\lambda)$. Hence, if we denote by $\widetilde{\tau}_B := \inf\{\,t \geq 0 \mid \widetilde{X}_t \geq B\,\}$ the first time at which $\widetilde{X}$ hits the barrier B, then
$$\{\,\widetilde{X}_T > K^*\,\} \subset \big\{\,\widetilde{\tau}_B \leq T\,\big\} = \{\,\widetilde{X}_{T-1} = B\,\} \subset \{\,\widetilde{X}_T \geq K^*\,\}.$$
Hence, $\{\widetilde{\tau}_B \leq T\} = \{\widetilde{\tau}_B = T-1\} = \{\widetilde{X}_{T-1} = B\}$,
$$\widetilde{P}[\,\widetilde{\tau}_B \leq T\,] = \widetilde{P}[\,\widetilde{X}_{T-1} = B\,] = 1 - \lambda = \min_{0 \leq K < B} \frac{C_T(K)}{B-K},$$
and the distribution $\hat{P}$ of $\widetilde{X}$ under $\widetilde{P}$ is as desired. □

Remark 7.30. The inequality
$$H^{\text{dig}} \leq \frac{(X_T - K)^+}{B-K} + \frac{X_{\tau_B} - X_T}{B-K}\, \mathrm{I}_{\{\tau_B \leq T\}}$$
appearing in the proof of Theorem 7.27 can be interpreted in terms of a suitable *superhedging strategy* for the claim H^{dig} by using call options and forward contracts: At time $t = 0$, we buy $(B-K)^{-1}$ call options with strike K, and at the first time when the price process passes the barrier B, we sell forward $(B-K)^{-1}$ shares of the asset. This strategy will be optimal if the strike price K is such that it realizes the minimum on the right-hand side of (7.12). By virtue of Lemma 7.28, such an optimal strike price can be identified as the *Value at Risk* at level $1 - \lambda$ of a short position $-X_T$ in the asset. ◇

Let us now derive bounds on the arbitrage-free prices of *barrier call options*. More precisely, we will consider an *up-and-in call option*
$$H^{\text{call}}_{\text{u\&i}} := \begin{cases} (X_T - K)^+ & \text{if } \max\limits_{0 \leq t \leq T} X_t \geq B,\\ 0 & \text{otherwise,}\end{cases}$$
and the corresponding *up-and-out call*
$$H^{\text{call}}_{\text{u\&o}} := \begin{cases} (X_T - K)^+ & \text{if } \max\limits_{0 \leq t \leq T} X_t < B,\\ 0 & \text{otherwise.}\end{cases}$$

If the barrier B is below the strike price K, then the up-and-in call is identical to a "plain vanilla call" $(X_T - K)^+$, and the payoff of the up-and-out call is zero. Thus, we assume from now on that

$$K < B.$$

Recall that K^* denotes the minimizer of the function $c \mapsto C_T(c)/(B-c)$ as constructed in Lemma 7.28.

Theorem 7.31. *For an up-and-in call option,*

$$\max_{P^* \in \mathcal{P}_{\Phi_T}} E^*[\, H^{\text{call}}_{\text{u\&i}}\,] = \begin{cases} C_T(K) & \text{if } K^* \le K, \\ \dfrac{B-K}{B-K^*}\, C_T(K^*) & \text{if } K^* > K. \end{cases}$$

Proof. For any c with $K \le c < B$,

$$H^{\text{call}}_{\text{u\&i}} \le \frac{B-K}{B-c}\,(X_T - c)^+ + \frac{c-K}{B-c}\,(X_{\tau_B} - X_T)\,\mathrm{I}_{\{\tau_B \le T\}}.$$

Indeed, on $\{X_T \le K\}$ or on $\{\tau_B > T\}$ the payoff of $H^{\text{call}}_{\text{u\&i}}$ is zero, and the right-hand side is non-negative. On $\{X_T \ge c,\ \tau_B \le T\}$, both sides are equal, and on $\{c > X_T > K,\ \tau_B \le T\}$ we may have a strict inequality. The expectation of the right-hand side under a martingale measure $P^* \in \mathcal{P}_{\Phi_T}$ is equal to

$$\frac{B-K}{B-c}\,C_T(c) + \frac{c-K}{B-c}\,E^*[\,(X_{\tau_B} - X_T)\,\mathrm{I}_{\{\tau_B \le T\}}\,] = \frac{B-K}{B-c}\,C_T(c),$$

due to the stopping theorem. The minimum of this upper bound over all $c \in [K, B)$ is attained in $c = K \vee K^*$, which shows $\le$ in the assertion.

Finally, let $\hat{P}$ be the martingale measure constructed in Lemma 7.29. If $K^* \le K$ then

$$(X_T - K)^+\,\mathrm{I}_{\{\tau_B \le T\}} = (X_T - K)^+ \quad \hat{P}\text{-a.s.},$$

and so $\hat{E}[\,H^{\text{call}}_{\text{u\&i}}\,] = C_T(K)$. If $K^* > K$ then $\hat{P}$-a.s.

$$\frac{B-K}{B-K^*}\,(X_T - K^*)^+ + \frac{K^*-K}{B-K^*}\,(B - X_T)\,\mathrm{I}_{\{\tau_B \le T\}} = H^{\text{call}}_{\text{u\&i}}.$$

Taking expectations with respect to $\hat{P}$ concludes the proof. □

Remark 7.32. The inequality

$$H^{\text{call}}_{\text{u\&i}} \le \frac{B-K}{B-c}\,(X_T - c)^+ + \frac{c-K}{B-c}\,(X_{\tau_B} - X_T)\,\mathrm{I}_{\{\tau_B \le T\}}$$

appearing in the preceding proof can be interpreted as a *superhedging strategy* for the up-and-in call with liquid derivatives: At time $t = 0$, we purchase $(B-K)/(B-c)$ call options with strike c, and at the first time when the stock price passes the barrier B, we sell forward $(c-K)/(B-c)$ shares of the asset. This strategy will be optimal for $c = K^* \vee K$. ◇

We now turn to the analysis of the up-and-out call option

$$H^{\text{call}}_{\text{u\&o}} = (X_T - K)^+ \, \mathrm{I}_{\{\tau_B > T\}}.$$

Theorem 7.33. *For an up-and-out call,*

$$\max_{P^* \in \mathcal{P}_{\Phi_T}} E^*\big[\, H^{\text{call}}_{\text{u\&o}} \,\big] = C_T(K) - C_T(B) - (B-K)\,\mu_T\big([B,\infty)\big).$$

Proof. Clearly,

$$\begin{aligned} H^{\text{call}}_{\text{u\&o}} &\le (X_T - K)^+ \, \mathrm{I}_{\{X_T < B\}} \\ &= (X_T - K)^+ - (X_T - B)^+ - (B-K)\,\mathrm{I}_{\{X_T \ge B\}}. \end{aligned} \tag{7.13}$$

Taking expectations yields $\le$ in the assertion.

Now consider the measure $\widetilde{P}$ on $(\Omega, \mathcal{F})$ defined as

$$\widetilde{P} := \delta_{X_0} \otimes \cdots \otimes \delta_{X_0} \otimes \mu_T,$$

i.e., under $\widetilde{P}$ we have $X_t = X_0$ $\widetilde{P}$-a.s. for $t < T$, and the law of X_T is μ_T. Clearly $\widetilde{P} \in \mathcal{P}_{\Phi_T}$, and (7.13) is $\widetilde{P}$-a.s. an identity. □

Using the identity

$$(X_T - K)^+ = H^{\text{call}}_{\text{u\&o}} + H^{\text{call}}_{\text{u\&i}},$$

we get the following lower bounds as an immediate corollary.

Corollary 7.34. *We have*

$$\min_{P^* \in \mathcal{P}_{\Phi_T}} E^*\big[\, H^{\text{call}}_{\text{u\&i}} \,\big] = C_T(B) + (B-K)\,\mu_T\big([B,\infty)\big),$$

and

$$\min_{P^* \in \mathcal{P}_{\Phi_T}} E^*\big[\, H^{\text{call}}_{\text{u\&o}} \,\big] = \begin{cases} 0 & \text{if } K^* \le K, \\ C_T(K) - \dfrac{B-K}{B-K^*}\, C_T(K^*) & \text{if } K^* > K. \end{cases}$$

Chapter 8
Efficient hedging

In an incomplete financial market model, a contingent claim typically will not admit a perfect hedge. Superhedging provides a method for staying on the safe side, but the required cost is usually too high both from a theoretical and from a practical point of view. It is thus natural to relax the requirements.

As a first preliminary step, we consider strategies of *quantile hedging* which stay on the safe side with high probability. In other words, we maximize the probability for staying on the safe side under a given cost constraint. The main idea consists in reducing the construction of such strategies for a given claim H to a problem of superhedging for a modified claim $\widetilde{H}$, which is the solution to a static optimization problem of Neyman–Pearson type. Typically, $\widetilde{H}$ will have the form of a knock-out option, that is, $\widetilde{H} = H \cdot I_A$. At this stage, we only focus on the probability that a shortfall occurs; we do not take into account the *size* of the shortfall if it does occur.

In Section 8.2 we take a more comprehensive view of the downside risk. Our discussion of risk measures in Section 4.8 suggests to quantify the downside risk in terms of an acceptance set for suitably hedged positions. If acceptability is defined in terms of shortfall risk as in Section 4.9, we are led to the problem of constructing *efficient* strategies which minimize the shortfall risk under a given cost constraint. As in the case of quantile hedging, this problem can be decomposed into a static optimization problem and the construction of a superhedging strategy for a modified payoff profile $\widetilde{H}$.

8.1 Quantile hedging

Let H be a discounted European claim in an arbitrage-free market model such that

$$\pi_{\sup}(H) = \sup_{P^* \in \mathcal{P}} E^*[\, H \,] < \infty.$$

We saw in Corollary 7.15 that there exists a self-financing trading strategy whose value process $V^{\uparrow}$ satisfies

$$V_T^{\uparrow} \geq H \quad P\text{-a.s.}$$

By using such a superhedging strategy, the seller of H can cover almost any possible obligation which may arise from the sale of H and thus eliminate completely the corresponding risk. The smallest amount for which such a superhedging strategy is available is given by $\pi_{\sup}(H)$. This cost will often be too high from a practical point of view, as illustrated by Example 7.21. Furthermore, if H is not attainable then $\pi_{\sup}(H)$, viewed as a price for H, is too high from a theoretical point of view since it would permit arbitrage. Even if H is attainable, a complete elimination of risk by using a replicating strategy for H would consume the entire proceeds from the sale of H, and any opportunity of making a profit would be lost along with the risk.

Let us therefore suppose that the seller is unwilling to put up the initial amount of capital required by a superhedge and is ready to accept some risk. What is the optimal partial hedge which can be achieved with a given smaller amount of capital? In order to make this question precise, we need a criterion expressing the seller's attitude towards risk. Several of such criteria will be studied in the following sections. In this section, our aim is to construct a strategy which *maximizes the probability of a successful hedge* given a constraint on the initial cost.

More precisely, let us fix an initial amount

$$v < \pi_{\sup}(H).$$

We are looking for a self-financing trading strategy whose value process maximizes the probability

$$P[\, V_T \geq H \,]$$

among all those strategies whose initial investment V_0 is bounded by v and which respect the bounds $V_t \geq 0$ for $t = 0, \dots, T$. In view of Theorem 5.26, the second restriction amounts to admissibility in the following sense:

Definition 8.1. A self-financing trading strategy is called an *admissible strategy* if its value process satisfies $V_T \geq 0$.

The problem of *quantile hedging* consists in constructing an admissible strategy $\overline{\xi}^*$ such that its value process V^* satisfies

$$P[\, V_T^* \geq H \,] = \max P[\, V_T \geq H \,] \tag{8.1}$$

where the maximum is taken over all value processes V of admissible strategies subject to the constraint

$$V_0 \leq v. \tag{8.2}$$

Note that this problem would not be well posed if considered without the constraint of admissibility.

Let us emphasize that the idea of quantile hedging corresponds to a *Value at Risk* criterion, and that it invites the same criticism: Only the *probability* of a shortfall is taken into account, not the *size* of the loss if a shortfall occurs. This exclusive focus on the shortfall probability may be reasonable in cases where a loss is to be avoided by any means. But for most applications, other optimality criteria as considered in the next section will usually be more appropriate from an economic point of view. In view of the mathematical techniques, however, some key ideas already appear quite clearly in our present context.

Let us first consider the particularly transparent situation of a complete market model before passing to the general incomplete case. The set

$$\{ V_T \geq H \}$$

will be called the *success set* associated with the value process V of an admissible strategy. As a first step, we reduce our problem to the construction of a success set of maximal probability.

Proposition 8.2. *Let P^* denote the unique equivalent martingale measure in a complete market model, and assume that $A^* \in \mathcal{F}_T$ maximizes the probability $P[\, A\,]$ among all sets $A \in \mathcal{F}_T$ satisfying the constraint*

$$E^*[\, H \cdot I_A \,] \leq v. \tag{8.3}$$

Then the replicating strategy $\overline{\xi^*}$ of the knock-out option

$$H^* := H \cdot I_{A^*}$$

solves the optimization problem defined by (8.1) *and* (8.2)*, and A^* coincides up to P-null sets with the success set of $\overline{\xi^*}$.*

Proof. As a first step, let V be the value process of any admissible strategy such that $V_0 \leq v$. We denote by $A := \{V_T \geq H\}$ the corresponding success set. Admissibility yields that $V_T \geq H \cdot I_A$. Moreover, the results of Section 5.3 imply that V is a P^*-martingale. Hence, we obtain that

$$E^*[\, H \cdot I_A \,] \leq E^*[\, V_T \,] = V_0 \leq v.$$

Therefore, A fulfills the constraint (8.3) and it follows that

$$P[\, A \,] \leq P[\, A^* \,].$$

As a second step, we consider the trading strategy $\overline{\xi^*}$ and its value process V^*. Clearly, $\overline{\xi^*}$ is admissible, and its success set satisfies

$$\{ V_T^* \geq H \} = \{ H \cdot I_{A^*} \geq H \} \supseteq A^*.$$

On the other hand, the first part of the proof yields that

$$P[\, V_T^* \geq H \,] \leq P[\, A^* \,].$$

It follows that the two sets A^* and $\{ V_T^* \geq H \}$ coincide up to P-null sets. In particular, $\overline{\xi^*}$ is an optimal strategy. □

Our next goal is the construction of the optimal success set A^*, whose existence was assumed in Proposition 8.2. This problem is solved by using the Neyman–Pearson lemma. To this end, we introduce the measure Q^* given by

$$\frac{dQ^*}{dP^*} := \frac{H}{E^*[\, H \,]}. \tag{8.4}$$

The constraint (8.3) can be written as

$$Q^*[\,A\,] \le \alpha := \frac{v}{E^*[\,H\,]}. \tag{8.5}$$

Thus, an optimal success set must maximize the probability $P[\,A\,]$ under the constraint $Q^*[\,A\,] \le \alpha$. We denote by dP/dQ^* the generalized density of P with respect to Q^* in the sense of the Lebesgue decomposition as constructed in Theorem A.13. Thus, we may define the level

$$c^* := \inf\left\{ c \ge 0 \;\middle|\; Q^*\!\left[\frac{dP}{dQ^*} > c \cdot E^*[\,H\,] \right] \le \alpha \right\}, \tag{8.6}$$

and the set

$$A^* := \left\{ \frac{dP}{dQ^*} > c^* \cdot E^*[\,H\,] \right\} = \left\{ \frac{dP}{dP^*} > c^* \cdot H \right\}. \tag{8.7}$$

Proposition 8.3. *If the set A^* in* (8.7) *satisfies*

$$Q^*[\,A^*\,] = \alpha,$$

then A^* maximizes the probability $P[\,A\,]$ over all $A \in \mathcal{F}_T$ satisfying the constraint

$$E^*[\,H \cdot I_A\,] \le v.$$

Proof. The condition $E^*[\,H \cdot I_A\,] \le v$ is equivalent to $Q^*[\,A\,] \le \alpha = Q^*[\,A^*\,]$. Thus, the particular form of the set A^* in (8.7) and the Neyman–Pearson lemma in the form of Proposition A.28 imply that $P[\,A\,] \le P[\,A^*\,]$. □

By combining the two Propositions 8.2 and 8.3, we obtain the following result.

Corollary 8.4. *Denote by P^* the unique equivalent martingale measure in a complete market model, and assume that the set A^* of* (8.7) *satisfies*

$$Q^*[\,A^*\,] = \alpha.$$

Then the optimal strategy solving (8.1) *and* (8.2) *is given by the replicating strategy of the knock-out option $H^* = H \cdot I_{A^*}$.*

Our solution to the optimization problem (8.1) and (8.2) still relies on the assumption that the set A^* of (8.7) satisfies $Q^*[\,A^*\,] = \alpha$. This condition is clearly satisfied if

$$P\left[\frac{dP}{dP^*} = c^* \cdot H \right] = 0.$$

However, it may not in general be possible to find any set A whose Q^*-probability is exactly α. In such a situation, the Neyman–Pearson theory suggests replacing the indicator function I_{A^*} of the "critical region" A^* by a *randomized test*, i.e., by an $\mathcal{F}_T$-measurable $[0, 1]$-valued function ψ. Let $\mathcal{R}$ denote the class of all randomized tests, and consider the following optimization problem:

$$E[\,\psi^*\,] = \max\Big\{ E[\,\psi\,] \mid \psi \in \mathcal{R} \text{ and } E_{Q^*}[\,\psi\,] \le \alpha \Big\},$$

where Q^* is the measure defined in (8.4) and $\alpha = v/E^*[\,H\,]$ as in (8.5). The generalized Neyman–Pearson lemma in the form of Theorem A.30 states that the solution is given by

$$\psi^* = I_{\left\{\frac{dP}{dP^*} > c^* \cdot H\right\}} + \gamma \cdot I_{\left\{\frac{dP}{dP^*} = c^* \cdot H\right\}}, \tag{8.8}$$

where c^* is defined through (8.6) and γ is chosen such that $E_{Q^*}[\,\psi^*\,] = \alpha$, i.e.,

$$\gamma = \frac{\alpha - Q^*\big[\,\frac{dP}{dP^*} > c^* \cdot H\,\big]}{Q^*\big[\,\frac{dP}{dP^*} = c^* \cdot H\,\big]} \quad \text{in case } P\left[\,\frac{dP}{dP^*} = c^* \cdot H\,\right] \ne 0.$$

Definition 8.5. Let V be the value process of an admissible strategy $\overline{\xi}$. The *success ratio* of $\overline{\xi}$ is defined as the randomized test

$$\psi_V = I_{\{V_T \ge H\}} + \frac{V_T}{H} \cdot I_{\{V_T < H\}}.$$

Note that the set $\{\,\psi_V = 1\,\}$ coincides with the success set $\{\,V_T \ge H\,\}$ of V. In the extended version of our original problem, we are now looking for a strategy which maximizes the expected success ratio $E[\,\psi_V\,]$ under the measure P under the cost constraint $V_0 \le v$:

Theorem 8.6. *Suppose that P^* is the unique equivalent martingale measure in a complete market model. Let ψ^* be given by* (8.8)*, and denote by $\overline{\xi^*}$ a replicating strategy for the discounted claim $H^* = H \cdot \psi^*$. Then the success ratio ψ_{V^*} of $\overline{\xi^*}$ maximizes the expected success ratio $E[\,\psi_V\,]$ among all admissible strategies with initial investment $V_0 \le v$. Moreover, the optimal success ratio ψ_{V^*} is P-a.s. equal to ψ^*.*

We do not prove this theorem here, as it is a special case of Theorem 8.7 below and its proof is similar to the one of Corollary 8.4, once the optimal randomized test ψ^* has been determined by the generalized Neyman–Pearson lemma. Note that the condition

$$P\left[\,\frac{dP}{dP^*} = c^* \cdot H\,\right] = 0$$

implies that $\psi^* = I_{A^*}$ with A^* as in (8.7), so in this case the strategy $\overline{\xi^*}$ reduces to the one described in Corollary 8.4.

Now we turn to the general case of an arbitrage-free but possibly incomplete market model, i.e., we no longer assume that the set $\mathcal{P}$ of equivalent martingale measures consists of a single element, but we assume only that

$$\mathcal{P} \neq \emptyset.$$

In this setting, our aim is to find an admissible strategy whose success ratio ψ_{V^*} satisfies

$$E[\,\psi_{V^*}\,] = \max E[\,\psi_V\,], \tag{8.9}$$

where the maximum on the right-hand side is taken over all admissible strategies whose initial investment satisfies the constraint

$$V_0 \leq v. \tag{8.10}$$

Theorem 8.7. *There exists a randomized test ψ^* such that*

$$\sup_{P^* \in \mathcal{P}} E^*[\,H \cdot \psi^*\,] = v, \tag{8.11}$$

and which maximizes $E[\,\psi\,]$ among all $\psi \in \mathcal{R}$ subject to the constraints

$$E^*[\,H \cdot \psi\,] \leq v \quad \textit{for all } P^* \in \mathcal{P}. \tag{8.12}$$

Moreover, the superhedging strategy for the modified claim

$$H^* = H \cdot \psi^*$$

with initial investment $\pi_{\sup}(H^*)$ solves the problem (8.9) *and* (8.10).

Proof. Denote by $\mathcal{R}_0$ the set of all $\psi \in \mathcal{R}$ which satisfy the constraints (8.12), and take a sequence $\psi_n \in \mathcal{R}_0$ such that

$$E[\,\psi_n\,] \longrightarrow \sup_{\psi \in \mathcal{R}_0} E[\,\psi\,] \quad \text{as } n \uparrow \infty.$$

Lemma 1.69 yields a sequence of convex combinations $\widetilde{\psi}_n \in \operatorname{conv}\{\psi_n, \psi_{n+1}, \dots\}$ converging P-a.s. to a function $\widetilde{\psi} \in \mathcal{R}$. Clearly, $\widetilde{\psi}_n \in \mathcal{R}_0$ for each n. Hence, Fatou's lemma yields that

$$E^*[\,H\,\widetilde{\psi}\,] \leq \liminf_{n\uparrow\infty} E^*[\,H\,\widetilde{\psi}_n\,] \leq v \quad \text{for all } P^* \in \mathcal{P},$$

and it follows that $\widetilde{\psi} \in \mathcal{R}_0$. Moreover,

$$E[\,\widetilde{\psi}\,] = \lim_{n\uparrow\infty} E[\,\widetilde{\psi}_n\,] = \lim_{n\uparrow\infty} E[\,\psi_n\,] = \sup_{\psi \in \mathcal{R}_0} E[\,\psi\,],$$

so $\psi^* := \widetilde{\psi}$ is the desired maximizer.

We must also show that (8.11) holds. To this end, note first that $P[\,\psi^* = 1\,] = 1$ is impossible due to our assumption $v < \pi_{\text{sup}}(H)$. Hence, if $\sup_{P^*\in\mathcal{P}} E^*[\,H \cdot \psi^*\,] < v$, then we can find some $\varepsilon > 0$ such that $\psi_\varepsilon := \varepsilon + (1-\varepsilon)\psi^* \in \mathcal{R}_0$, and the expectation $E[\,\psi_\varepsilon\,]$ must be strictly larger than $E[\,\psi^*\,]$. This, however, contradicts the maximality of $E[\,\psi^*\,]$.

Now let $\overline{\xi}$ be any admissible strategy whose value process V satisfies $V_0 \le v$. If ψ_V denotes the corresponding success ratio, then

$$H \cdot \psi_V = H \wedge V_T \le V_T.$$

The $\mathcal{P}$-martingale property of V yields that for all $P^* \in \mathcal{P}$,

$$E^*[\,H \cdot \psi_V\,] \le E^*[\,V_T\,] = V_0 \le v. \tag{8.13}$$

Therefore, ψ_V is contained in $\mathcal{R}_0$ and it follows that

$$E[\,\psi_V\,] \le E[\,\psi^*\,]. \tag{8.14}$$

Consider the superhedging strategy $\overline{\xi^*}$ of $H^* = H \cdot \psi^*$ and denote by V^* its value process. Clearly, $\overline{\xi^*}$ is an admissible strategy. Moreover,

$$V_0^* = \pi_{\text{sup}}(H^*) = \sup_{P^*\in\mathcal{P}} E^*[\,H \cdot \psi^*\,] = v.$$

Thus, (8.14) yields that ψ_{V^*} satisfies

$$E[\,\psi_{V^*}\,] \le E[\,\psi^*\,]. \tag{8.15}$$

On the other hand, V_T^* dominates H^*, so

$$H \cdot \psi_{V^*} = H \wedge V_T^* \ge H \wedge H^* = H \cdot \psi^*.$$

Therefore, ψ_{V^*} dominates ψ^* on the set $\{\,H > 0\,\}$. Moreover, any success ratio is equal to one on $\{\,H = 0\,\}$, and we obtain that $\psi_{V^*} \ge \psi^*$ P-almost surely. According to (8.15), this can only happen if the two randomized tests ψ_{V^*} and ψ^* coincide P-almost everywhere. This proves that $\overline{\xi^*}$ solves the hedging problem (8.9) and (8.10). □

8.2 Hedging with minimal shortfall risk

Our starting point in this section is the same as in the previous one: At time T, an investor must pay the discounted random amount $H \ge 0$. A complete elimination of the corresponding risk would involve the cost

$$\pi_{\text{sup}}(H) = \sup_{P^*\in\mathcal{P}} E^*[\,H\,]$$

of superhedging H, but the investor is only willing to put up a smaller amount

$$v \in \big(0, \pi_{\sup}(H)\big).$$

This means that the investor is ready to take some risk: Any "partial" hedging strategy whose value process V satisfies the capital constraint $V_0 \le v$ will generate a non-trivial *shortfall*

$$(H - V_T)^+.$$

In the previous section, we constructed trading strategies which minimize the shortfall probability

$$P[\, V_T < H \,]$$

among the class of trading strategies whose initial investment is bounded by v, and which are admissible in the sense of Definition 8.1, i.e., their terminal value V_T is non-negative. In this section, we assess the shortfall in terms of a *loss function*, i.e., an increasing function $\ell : \mathbb{R} \to \mathbb{R}$ which is not identically constant. We assume furthermore that

$$\ell(x) = 0 \text{ for } x \le 0 \quad \text{and} \quad E[\, \ell(H) \,] < \infty.$$

A particular role will be played by *convex* loss functions, which correspond to *risk aversion* in view of the shortfall; compare the discussion in Section 4.9.

Definition 8.8. Given a loss function ℓ satisfying the above assumptions, the *shortfall risk* of an admissible strategy with value process V is defined as the expectation

$$E\big[\, \ell(H - V_T) \,\big] = E\big[\, \ell\big((H - V_T)^+ \big) \,\big]$$

of the shortfall weighted by the loss function ℓ.

Our aim is to minimize the shortfall risk among all admissible strategies satisfying the capital constraint $V_0 \le v$. Alternatively, we could minimize the cost under a given bound on the shortfall risk. In other words, the problem consists in constructing strategies which are *efficient* with respect to the trade-off between cost and shortfall risk. This generalizes our discussion of quantile hedging in the previous Section 8.1, which corresponds to a minimization of the shortfall risk with respect to the non-convex loss function

$$\ell(x) = I_{(0,\infty)}(x).$$

Remark 8.9. Recall our discussion of risk measures in Chapter 4. From this point of view, it is natural to quantify the downside risk in terms of an acceptance set $\mathcal{A}$ for hedged positions. As in Section 4.8, we denote by $\tilde{\mathcal{A}}$ the class of all positions X such that there exists an admissible strategy $\overline{\xi}$ with value process V such that

$$V_0 = 0 \quad \text{and} \quad X + V_T \ge A \quad P\text{-a.s.}$$

for some $A \in \mathcal{A}$. Thus, the downside risk of the position $-H$ takes the form

$$\overline{\rho}(-H) = \inf\{ m \in \mathbb{R} \mid m - H \in \bar{\mathcal{A}} \}.$$

Suppose that the acceptance set $\mathcal{A}$ is defined in terms of shortfall risk, i.e.,

$$\mathcal{A} := \{ X \in L^\infty \mid E[\, \ell(X^-) \,] \le x_0 \},$$

where ℓ is a convex loss function and x_0 is a given threshold. Then $\overline{\rho}(-H)$ is the smallest amount m such that there exists an admissible strategy $\overline{\xi}$ whose value process V satisfies $V_0 = m$ and

$$E\big[\, \ell\big((H - V_T)^+\big) \,\big] \le x_0.$$

For a given m, we are thus led to the problem of finding a strategy $\overline{\xi}$ which minimizes the shortfall risk under the cost constraint $V_0 \le m$. In this way, the problem of quantifying the downside risk of a contingent claim is reduced to the construction of efficient hedging strategies as discussed in this section. ◇

As in the preceding section, the construction of the optimal hedging strategy is carried out in two steps. The first one is to solve the "static" problem of minimizing

$$E[\, \ell(H - Y) \,]$$

among all $\mathcal{F}_T$-measurable random variables $Y \ge 0$ which satisfy the constraints

$$\sup_{P^* \in \mathcal{P}} E^*[\, Y \,] \le v.$$

If Y^* solves this problem, then so does $\widetilde{Y} := H \wedge Y^*$. Hence, we may assume that $0 \le Y^* \le H$ or, equivalently, that $Y^* = H\,\psi^*$ for some *randomized test* ψ^*, which belongs to the set $\mathcal{R}$ of all $\mathcal{F}_T$-measurable random variables with values in $[0, 1]$. Thus, the static problem can be reformulated as follows: Find a randomized test $\psi^* \in \mathcal{R}$ which minimizes the "shortfall risk"

$$E\big[\, \ell\big(H(1 - \psi) \big)\big] \tag{8.16}$$

among all $\psi \in \mathcal{R}$ subject to the constraints

$$E^*[\, H\,\psi \,] \le v \quad \text{for all } P^* \in \mathcal{P}. \tag{8.17}$$

The next step is to fit the terminal value V_T of an admissible strategy to the optimal profile $H\,\psi^*$. It turns out that this step can be carried out without any further assumptions on our loss function ℓ. Thus, we assume at this point that the optimal ψ^* of step one is granted, and we construct the corresponding optimal strategy.

Theorem 8.10. *Given a randomized test ψ^* which minimizes* (8.16) *subject to* (8.17)*, a superhedging strategy $\overline{\xi^*}$ for the modified discounted claim*

$$H^* := H\,\psi^*$$

with initial investment $\pi_{\sup}(H^*)$ has minimal shortfall risk among all admissible strategies $\overline{\xi}$ which satisfy the capital constraint $\overline{\xi}_1 \cdot \overline{X}_0 \le v$.

Proof. The proof extends the last argument in the proof of Theorem 8.7. As a first step, we take any admissible strategy $\overline{\xi}$ such that the corresponding value process V satisfies the capital constraint $V_0 \le v$. Denote by

$$\psi_V = I_{\{V_T \ge H\}} + \frac{V_T}{H} \cdot I_{\{V_T < H\}}$$

the corresponding success ratio. It follows as in (8.13) that ψ_V satisfies the constraints

$$E^*[\, H\, \psi_V \big] \le v \quad \text{for all } P^* \in \mathcal{P}.$$

Thus, the optimality of ψ^* implies the following lower bound on the shortfall risk of $\overline{\xi}$:

$$E\big[\, \ell\big(H - V_T\big)\big] = E\big[\, \ell\big(\, H(1 - \psi_V)\,\big)\big] \ge E\big[\, \ell\big(\, H(1 - \psi^*)\,\big)\big].$$

In the second step, we consider the admissible strategy $\overline{\xi^*}$ and its value process V^*. On the one hand,

$$V_0^* = \pi_{\text{sup}}(H^*) = \sup_{P^* \in \mathcal{P}} E^*[\, H\, \psi^*] \le v,$$

so $\overline{\xi^*}$ satisfies the capital constraint. Hence, the first part of the proof yields

$$E\big[\, \ell\big(H - V_T^*\big)\big] = E\big[\, \ell\big(\, H(1 - \psi_{V^*})\,\big)\big] \ge E\big[\, \ell\big(\, H(1 - \psi^*)\,\big)\big]. \tag{8.18}$$

On the other hand, $V_T^* \ge H^* = H\, \psi^*$, and therefore

$$\psi_{V^*} \ge \psi^* \quad P\text{-a.s.}$$

Hence, the inequality in (8.18) is in fact an equality, and the assertion follows. □

Let us now return to the static problem defined by (8.16) and (8.17). We start by considering the special case of risk aversion in view of the shortfall.

Proposition 8.11. *If the loss function ℓ is convex, then there exists a randomized test $\psi^* \in \mathcal{R}$ which minimizes the shortfall risk*

$$E\big[\, \ell\big(\, H(1 - \psi)\,\big)\big]$$

among all $\psi \in \mathcal{R}$ subject to the constraints

$$E^*[\, H\, \psi\,] \le v \quad \textit{for all } P^* \in \mathcal{P}. \tag{8.19}$$

If ℓ is strictly convex on $[0, \infty)$, then ψ^* is uniquely determined on $\{H > 0\}$.

Proof. The proof is similar to the one of Proposition 3.36. Let $\mathcal{R}_0$ denote the set of all randomized tests which satisfy the constraints (8.19). Take $\psi_n \in \mathcal{R}_0$ such that $E[\ell(\, H(1 - \psi_n)\,)]$ converges to the infimum of the shortfall risk, and use Lemma 1.69

to select convex combinations $\widetilde{\psi}_n \in \operatorname{conv}\{\psi_n, \psi_{n+1}, \dots\}$ which converge P-a.s. to some $\widetilde{\psi} \in \mathcal{R}$. Since ℓ is continuous and increasing, Fatou's lemma implies that

$$E\big[\,\ell\big(H(1-\widetilde{\psi})\big)\big] \le \liminf_{n\uparrow\infty} E\big[\,\ell\big(H(1-\widetilde{\psi}_n)\big)\big] = \inf_{\psi\in\mathcal{R}_0} E\big[\,\ell\big(H(1-\psi)\big)\big],$$

where we have used the convexity of ℓ to conclude that $E[\ell(H(1-\widetilde{\psi}_n))]$ tends to the same limit as $E[\ell(H(1-\psi_n))]$.

Fatou's lemma also yields that for all $P^* \in \mathcal{P}$

$$E^*[\,H\,\widetilde{\psi}\,] \le \liminf_{n\uparrow\infty} E^*[\,H\,\widetilde{\psi}_n\,] \le v.$$

Hence $\widetilde{\psi} \in \mathcal{R}_0$, and we conclude that $\psi^* := \widetilde{\psi}$ is the desired minimizer. The uniqueness part is obvious. □

Remark 8.12. The proof shows that the analogous existence result holds if we use a robust version of the shortfall risk defined as

$$\sup_{Q\in\mathcal{Q}} E_Q\big[\,\ell\big(H(1-\psi)\big)\big],$$

where $\mathcal{Q}$ is a class of equivalent probability measures; see also Remark 3.37. ◇

Combining Proposition 8.11 and Theorem 8.10 yields existence and uniqueness of an optimal hedging strategy under risk aversion in a general arbitrage-free market model.

Corollary 8.13. *Assume that the loss function ℓ is strictly convex on $[0,\infty)$. Then there exists an admissible strategy which is optimal in the sense that it minimizes the shortfall risk among all admissible strategies $\overline{\xi}$ subject to the capital constraint $\overline{\xi}_1 \cdot \overline{X}_0 \le v$. Moreover, any optimal strategy requires the exact initial investment v, and its success ratio is P-a.s. equal to*

$$\psi^* \cdot \mathrm{I}_{\{H>0\}} + \mathrm{I}_{\{H=0\}},$$

where ψ^* denotes the solution of the static problem constructed in Proposition 8.11.

Proof. The existence of an optimal strategy follows by combining Proposition 8.11 and Theorem 8.10. Strict convexity of ℓ implies that ψ^* is P-a.s. unique on $\{H>0\}$. Since ℓ is strictly increasing on $[0,\infty)$, ψ^* and the success ratio ψ_{V^*} of any optimal strategy $\overline{\xi}^*$ must coincide P-a.s. on $\{H>0\}$. On $\{H=0\}$, the success ratio ψ_{V^*} is equal to 1 by definition.

Since ℓ is strictly increasing on $[0,\infty)$, we must have that

$$\sup_{P^*\in\mathcal{P}} E^*[\,H\,\psi^*\,] = v,$$

for otherwise we could find some $\varepsilon > 0$ such that $\psi_\varepsilon := \varepsilon + (1-\varepsilon)\psi^*$ would also satisfy the constraints (8.17). Since we have assumed that $v < \pi_{\sup}(H)$, the constraints (8.17) imply that $\psi^* \not\equiv 1$ and hence that

$$E\big[\,\ell\big(H(1-\psi_\varepsilon)\big)\big] < E\big[\,\ell\big(H(1-\psi^*)\big)\big].$$

This, however, contradicts the optimality of ψ^*.

Since the value process V^* of an optimal strategy is a $\mathcal{P}$-martingale, and since

$$V_T^* \geq H\,\psi_{V^*} = H\,\psi^*,$$

we conclude from the above that

$$v \geq V_0^* = \sup_{P^* \in \mathcal{P}} E^*[\,V_T^*\,] \geq \sup_{P^* \in \mathcal{P}} E^*[\,H\,\psi^*\,] = v.$$

Thus, V_0^* is equal to v. □

Beyond the general existence statement of Proposition 8.11, it is possible to obtain an explicit formula for the optimal solution of the static problem if the market model is complete. Recall that we assume that the loss function $\ell(x)$ vanishes for $x \leq 0$. In addition, we will also assume that

$$\ell \text{ is strictly convex and continuously differentiable on } (0,\infty).$$

Then the derivative ℓ' of ℓ is strictly increasing on $(0,\infty)$. Let J denote the inverse function of ℓ' defined on the range of ℓ', i.e., on the interval (a,b) where $a := \lim_{x\downarrow 0}\ell'(x)$ and $b := \lim_{x\uparrow\infty}\ell'(x)$. We extend J to a function $J^+ : [0,\infty] \to [0,\infty]$ by setting

$$J^+(y) := \begin{cases} +\infty & \text{for } y \geq b, \\ 0 & \text{for } y \leq a. \end{cases}$$

From now on, we assume also that

$$\mathcal{P} = \{P^*\},$$

i.e., P^* is the unique equivalent martingale measure in a complete market model. Its density will be denoted by

$$\varphi^* := \frac{dP^*}{dP}.$$

Theorem 8.14. *Under the above assumptions, the solution of the static optimization problem of Proposition* 8.11 *is given by*

$$\psi^* = 1 - \frac{J^+(c\,\varphi^*)}{H} \wedge 1 \quad P\text{-a.s. on } \{\,H > 0\,\},$$

where the constant c is determined by the condition $E^*[\,H\,\psi^*\,] = v$.

Proof. The problem is of the same type as those considered in Section 3.3. It can in fact be reduced to Corollary 3.43 by considering the random utility function

$$u(x,\omega) := -\ell(H(\omega)-x), \quad 0 \le x \le H(\omega).$$

Just note that the shortfall risk $E[\,\ell(H-Y)\,]$ coincides with the negative expected utility $-E[u(Y,\cdot)\,]$ for any profile Y such that $0 \le Y \le H$. Moreover, since our market model is complete, it has a finite structure by Theorem 5.38, and so all integrability conditions are automatically satisfied. Thus, Corollary 3.43 states that the optimal profile $H^* := Y^*$ which maximizes the expected utility $E[\,u(Y,\cdot)\,]$ under the constraints $0 \le Y \le H$ and $E^*[\,Y\,] \le v$ is given by

$$H^*(\omega) = I^+(c\,\varphi^*(\omega),\omega) \wedge H(\omega) = \big(H(\omega) - J^+(c\,\varphi^*(\omega))\big)^+.$$

Dividing by H yields the formula for the optimal randomized test ψ^*. □

Corollary 8.15. *In the situation of Theorem* 8.14*, suppose that the objective probability measure P is equal to the martingale measure P^*. Then the modified discounted claim takes the simple form*

$$H^* = H\,\psi^* = \big(H - J^+(c^*)\big)^+.$$

Example 8.16. Consider the discounted payoff H of a European call option $(S^i_T - K)^+$ with strike K under the assumption that the numéraire S^0 is a riskless bond, i.e., that $S^0_t = (1+r)^t$ for a certain constant $r \ge 0$. If the assumptions of Corollary 8.15 hold, then the modified profile H^* is the discounted value of the European call option struck at $\widetilde{K} := K + J^+(c^*)\cdot(1+r)^T$, i.e.,

$$H^* = \frac{(S^i_T - \widetilde{K})^+}{(1+r)^T}. \qquad \diamond$$

Example 8.17. Consider an exponential loss function $\ell(x) = (e^{\alpha x} - 1)^+$ for some $\alpha > 0$. In this case,

$$J^+(y) = \left(\frac{1}{\alpha}\log\frac{y}{\alpha}\right)^+, \quad y \ge 0,$$

and the optimal profile is given by

$$H^* = H - \left(\frac{1}{\alpha}\log\frac{c\varphi^*}{\alpha}\right)^+ \wedge H. \qquad \diamond$$

Example 8.18. If ℓ is the particular loss function

$$\ell(x) = \frac{x^p}{p}, \quad x \ge 0,$$

for some $p > 1$, then the problem is to minimize a lower partial moment of the difference $V_T - H$. Theorem 8.14 implies that it is optimal to hedge the modified claim

$$H\,\psi_p^* = H - (c_p \cdot \varphi^*)^{1/(p-1)} \wedge H \tag{8.20}$$

where the constant c_p is determined by $E^*[\,H\,\psi_p^*] = v$. ◇

Let us now consider the limit $p \uparrow \infty$ in (8.20), corresponding to ever increasing risk aversion with respect to large losses.

Proposition 8.19. *Let us consider the loss functions*

$$\ell_p(x) = \frac{x^p}{p}, \quad x \geq 0,$$

for $p > 1$. As $p \uparrow \infty$, the modified claims $H\,\psi_p^*$ of (8.20) *converge P-a.s. and in $L^1(P^*)$ to the discounted claim*

$$(H - c_\infty)^+$$

where the constant c_∞ is determined by

$$E^*[\,(H - c_\infty)^+\,] = v. \tag{8.21}$$

Proof. Let $\gamma(p)$ be shorthand for $1/(p-1)$ and note that

$$(\varphi^*)^{\gamma(p)} \longrightarrow 1 \quad P\text{-a.s. as } p \uparrow \infty.$$

Hence, if (p_n) is a sequence for which $c_{p_n}^{\gamma(p_n)}$ converges to some $\widetilde{c} \in [0, \infty]$, then

$$\lim_{n\uparrow\infty} H\,\psi_{p_n}^* = H - \widetilde{c} \wedge H = (H - \widetilde{c})^+.$$

Hence,

$$E^*[\,H\,\psi_{p_n}^*\,] \longrightarrow E^*[\,(H - \widetilde{c})^+\,].$$

Since each term on the left-hand side equals v, we must have

$$E^*[\,(H - \widetilde{c})^+\,] = v,$$

which determines $\widetilde{c}$ uniquely as the constant c_∞ of (8.21). □

Example 8.20. If the discounted claim H in Proposition 8.19 is the discounted payoff of a call option with strike K, and the numéraire is a riskless bond as in Example 8.16, then the limiting profile $\lim_{p\uparrow\infty} H\,\psi_p^*$ is equal to the discounted call with the higher strike price $K + c_\infty \cdot S_T^0$. ◇

In the remainder of this section, we consider loss functions which are not convex but which correspond to risk neutrality and to risk-seeking preferences. Let us first consider the risk-neutral case.

Example 8.21. In the case of risk neutrality, the loss function is given by

$$\ell(x) = x \quad \text{for } x \geq 0.$$

Thus, the task is to minimize the *expected shortfall*

$$E[\,(H - V_T)^+\,]$$

under the capital constraint $V_0 \leq v$. Let P^* be the unique equivalent martingale measure in a complete market model. Then the static problem corresponding to Proposition 8.11 is to maximize the expectation

$$E[\,H\,\psi\,]$$

under the constraint that $\psi \in \mathcal{R}$ satisfies

$$E^*[\,H\,\psi\,] \leq v.$$

For simplicity, we assume $P[\,H > 0\,] = 1$. Then we can define two equivalent measures Q and Q^* by

$$\frac{dQ}{dP} = \frac{H}{E[\,H\,]} \quad \text{and} \quad \frac{dQ^*}{dP^*} = \frac{H}{E^*[\,H\,]}.$$

The problem then becomes the hypothesis testing problem of maximizing $E_Q[\,\psi\,]$ under the side condition

$$E_{Q^*}[\,\psi\,] \leq \alpha := \frac{v}{E^*[\,H\,]}.$$

Since the density dQ/dQ^* is proportional to the inverse of the density $\varphi^* = dP^*/dP$, Theorem A.30 implies that the optimal test takes the form

$$\psi_1^* = I_{\{\varphi^* < c_1\}} + \gamma \cdot I_{\{\varphi^* = c_1\}}$$

where the constant c_1 is given by

$$c_1 = \sup\left\{\, c \in \mathbb{R} \mid E^*[\,H; \varphi^* < c\,] \leq v \,\right\},$$

and where the constant γ is chosen such that $E^*[\,H\,\psi_1^*\,] = v$. ◇

Assume now that the shortfall risk is assessed by an investor who, instead of being risk-averse, is in fact inclined to take risk. In our context, this corresponds to a loss function which is concave on $[0, \infty)$ rather than convex. It is not difficult to generalize Theorem 8.14 so that it covers this situation. Here we limit ourselves to the following explicit case study.

Example 8.22. Consider the loss function

$$\ell(x) = \frac{x^q}{q}, \quad x \geq 0,$$

for some $q \in (0, 1)$. In order to solve our static optimization problem, one could apply the results and techniques of Section 3.3. Here we will use an approach based on the Neyman–Pearson lemma. Note first that for $\psi \in \mathcal{R}$

$$\ell\big(H(1-\psi)\big) = (1-\psi)^q \cdot \ell(H) \geq \ell(H) - \psi \cdot \ell(H).$$

Hence, we get a lower bound on the "shortfall risk" of ψ:

$$E\big[\,\ell\big(H(1-\psi)\big)\,\big] \geq E\big[\ell(H)\,\big] - E\big[\,\psi \cdot \ell(H)\,\big]. \tag{8.22}$$

The problem of finding a minimizer of the right-hand side is equivalent to maximizing the expectation $E_Q[\,\psi\,]$ under the constraint that $E_{Q^*}[\,\psi\,] \leq v/E^*[\,H\,]$ for the measures Q and Q^* defined via

$$\frac{dQ}{dP} = \frac{H^q}{E[\,H^q\,]} \quad \text{and} \quad \frac{dQ^*}{dP^*} = \frac{H}{E^*[\,H\,]},$$

if we assume again $P[\,H > 0\,] = 1$. As in Example 8.21, we then conclude that the optimal test must be of the form

$$I_{\{1 > c_q^* \varphi^* H^{1-q}\}} + \gamma \cdot I_{\{1 = c_q^* \varphi^* H^{1-q}\}}. \tag{8.23}$$

for certain constants c_q^* and γ. Under the simplifying assumption that

$$P[\,1 = c_q^* \varphi^* H^{1-q}\,] = 0, \tag{8.24}$$

the formula (8.23) reduces to

$$\psi_q^* = \begin{cases} 1 & \text{on } \{\,1 > c_q^* \varphi^* H^{1-q}\,\}, \\ 0 & \text{otherwise.} \end{cases} \tag{8.25}$$

By taking $\psi = \psi_q^*$ we obtain an identity in (8.22), and so ψ_q^* must be a minimizer for $E[\,\ell(H(1-\psi))\,]$ under the constraint that $E^*[\,H\,\psi\,] \leq v$. ◇

In our last result of this section, we recover the knock-out option

$$H \cdot I_{\{1 > c_0^* \cdot H\,\varphi^*\}},$$

which was obtained as the solution to the problem of quantile hedging by taking the limit $q \downarrow 0$ in (8.25). Intuitively, decreasing q corresponds to an increasing appetite for risk in view of the shortfall.

Proposition 8.23. *Let us assume for simplicity that* (8.24) *holds for all* $q \in (0, 1)$, *that* $P[\, H > 0\,] = 1$, *and that there exists a unique constant* c_0^* *such that*

$$E^*\big[\, H \cdot I_{\{1 > c_0^* \cdot H\, \varphi^*\}}\,\big] = v. \tag{8.26}$$

Then the solutions ψ_q^* *of* (8.25) *converge* P*-a.s. to the solution*

$$\psi_0^* = I_{\{1 > c_0^* \cdot H\, \varphi^*\}}$$

of the corresponding problem of quantile hedging as constructed in Proposition 8.3.

Proof. Take any sequence $q_n \downarrow 0$ such that $(c_{q_n}^*)^{1/(1-q_n)}$ converges to some $\widetilde{c} \in [0, \infty]$. Then

$$\lim_{n \uparrow \infty} \psi_{q_n}^* = I_{\{1 > \widetilde{c} \cdot H\, \varphi^*\}}.$$

Hence,

$$E^*[\, H\, \psi_{q_n}^*\,] \longrightarrow E^*\big[\, H \cdot I_{\{1 > \widetilde{c} \cdot H\, \varphi^*\}}\,\big].$$

Since we assumed (8.24) for all $q \in (0, 1)$, the left-hand terms are all equal to v, and it follows from (8.26) that $\widetilde{c} = c_0^*$. This establishes the desired convergence. □

Chapter 9

Hedging under constraints

So far, we have focussed on *frictionless* market models, where asset transactions can be carried out with no limitation. In this chapter, we study the impact of market imperfections generated by *convex trading constraints*. Thus, we develop the theory of dynamic hedging under the condition that only trading strategies from a given class $\mathcal{S}$ may be used. In Section 9.1 we characterize those market models for which $\mathcal{S}$ does not contain arbitrage opportunities. Then we take a direct approach to the superhedging duality for American options. To this end, we first derive a uniform Doob decomposition under constraints in Section 9.2. The appropriate upper Snell envelopes are analyzed in Section 9.3. In Section 9.4 we derive a superhedging duality under constraints, and we explain its role in the analysis of convex risk measures in a financial market model.

9.1 Absence of arbitrage opportunities

In practice, it may be reasonable to restrict the class of trading strategies which are admissible for hedging purposes. As discussed in Section 4.8, there may be upper bounds on the capital invested into risky assets, or upper and lower bounds on the number of shares of an asset. Here we model such portfolio constraints by a set $\mathcal{S}$ of d-dimensional predictable processes, viewed as admissible investment strategies into risky assets. Throughout this chapter, we will assume that $\mathcal{S}$ satisfies the following conditions:

(a) $0 \in \mathcal{S}$.

(b) $\mathcal{S}$ is *predictably convex*: If $\xi, \eta \in \mathcal{S}$ and h is a predictable process with $0 \leq h \leq 1$, then the process

$$h_t\, \xi_t + (1 - h_t)\, \eta_t, \quad t = 1, \dots, T,$$

belongs to $\mathcal{S}$.

(c) For each $t \in \{1, \dots, T\}$, the set

$$\mathcal{S}_t := \{\, \xi_t \mid \xi \in \mathcal{S}\,\}$$

is closed in $L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$.

(d) For all t, $\xi_t \in \mathcal{S}_t$ implies $\xi_t^{\perp} \in \mathcal{S}_t$.

In order to explain condition (d), let us recall from Lemma 1.65 that each $\xi_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$ can be uniquely decomposed as

$$\xi_t = \eta_t + \xi_t^\perp, \quad \text{where } \eta_t \in N_t \text{ and } \xi_t^\perp \in N_t^\perp,$$

and where

$$N_t = \big\{ \eta_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d) \mid \eta_t \cdot (X_t - X_{t-1}) = 0 \ P\text{-a.s.} \big\},$$

$$N_t^\perp = \big\{ \xi_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d) \mid \xi_t \cdot \eta_t = 0 \ P\text{-a.s. for all } \eta_t \in N_t \big\}.$$

Remark 9.1. Under condition (d), we may replace $\xi_t \cdot (X_t - X_{t-1})$ by $\xi_t^\perp \cdot (X_t - X_{t-1})$, and $\xi_t^\perp \cdot (X_t - X_{t-1}) = 0$ P-a.s. implies $\xi_t^\perp = 0$. Note that condition (d) holds if the price increments satisfy the following *non-redundance condition*: For all $t \in \{1, \dots, T\}$ and $\xi_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$,

$$\xi_t \cdot (X_t - X_{t-1}) = 0 \ P\text{-a.s.} \quad \Longrightarrow \quad \xi_t = 0 \ P\text{-a.s.} \tag{9.1}$$

◇

Example 9.2. For each t let C_t be a closed convex subset of $\mathbb{R}^d$ such that $0 \in C_t$. Take $\mathcal{S}$ as the class of all d-dimensional predictable processes ξ such that $\xi_t \in C_t$ P-a.s. for all t. If the non-redundance condition (9.1) holds, then $\mathcal{S}$ satisfies conditions (a) through (d). This case includes short sales constraints and restrictions on the size of a long position. ◇

Example 9.3. Let a, b be two constants such that $-\infty \le a < 0 < b \le \infty$, and take $\mathcal{S}$ as the set of all d-dimensional predictable processes such that

$$a \le \xi_t \cdot X_{t-1} \le b \quad P\text{-a.s. for } t = 1, \dots, T.$$

This class $\mathcal{S}$ corresponds to constraints on the capital invested into risky assets. If we assume that the non-redundance condition (9.1) holds, then $\mathcal{S}$ satisfies conditions (a) through (d). More generally, instead of the two constants a and b, one can take dynamic margins defined via two predictable processes (a_t) and (b_t). ◇

Let $\overline{\mathcal{S}}$ denote the set of all self-financing trading strategies $\overline{\xi} = (\xi^0, \xi)$ which arise from an investment strategy $\xi \in \mathcal{S}$, i.e.,

$$\overline{\mathcal{S}} = \big\{ \overline{\xi} = (\xi^0, \xi) \mid \overline{\xi} \text{ is self-financing and } \xi \in \mathcal{S} \big\}.$$

In this section, our goal is to characterize the absence of arbitrage opportunities in $\overline{\mathcal{S}}$. The existence of an equivalent martingale measure $P^* \in \mathcal{P}$ is clearly sufficient. A condition which is both necessary and sufficient will involve a larger class $\mathcal{P}_{\mathcal{S}} \supset \mathcal{P}$. In order to introduce this condition, we need some preparation.

Definition 9.4. An adapted stochastic process Z on $(\Omega, \mathcal{F}, (\mathcal{F}_t), Q)$ is called a *local Q-martingale* if there exists a sequence of stopping times $(\tau_n)_{n\in\mathbb{N}} \subset \mathcal{T}$ such that $\tau_n \nearrow T$ Q-a.s., and such that the stopped processes Z^{τ_n} are Q-martingales. The sequence $(\tau_n)_{n\in\mathbb{N}}$ is called a *localizing sequence* for Z. In the same way, we define *local supermartingales* and *local submartingales*.

Remark 9.5. If Q is a martingale measure for the discounted price process X, then the value process V of each self-financing trading strategy $\overline{\xi} = (\xi^0, \xi)$ is a local Q-martingale. To prove this, one can take the sequence

$$\tau_n := \inf \Big\{ t \geq 0 \mid |\xi_{t+1}| > n \Big\} \wedge T$$

as a localizing sequence. With this choice, $|\xi_t| \leq n$ on $\{\tau_n \geq t\}$, and the increments

$$V_t^{\tau_n} - V_{t-1}^{\tau_n} = \mathrm{I}_{\{\tau_n \geq t\}}\, \xi_t \cdot (X_t - X_{t-1}), \quad t = 1, \dots, T,$$

of the stopped process V^{τ_n} are Q-integrable and satisfy

$$E_Q[\, V_t^{\tau_n} - V_{t-1}^{\tau_n} \mid \mathcal{F}_{t-1}\,] = \mathrm{I}_{\{\tau_n \geq t\}}\, \xi_t \cdot E_Q[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,] = 0. \qquad \diamondsuit$$

The following proposition is a generalization of an argument which we have already used in the proof of Theorem 5.26. Throughout this chapter, we will assume that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = \mathcal{F}$.

Proposition 9.6. *A local Q-supermartingale Z whose negative part Z_t^- is integrable for each $t \in \{1, \dots, T\}$ is a Q-supermartingale.*

Proof. Let (τ_n) be a localizing sequence. Then

$$Z_t^{\tau_n} \geq -\sum_{s=0}^{T} Z_s^- \in \mathcal{L}^1(Q).$$

In view of $\lim_n Z_t^{\tau_n} = Z_t$, Fatou's lemma for conditional expectations implies that Q-a.s.

$$E_Q[\, Z_t \mid \mathcal{F}_{t-1}\,] \leq \liminf_{n\uparrow\infty} E_Q[\, Z_t^{\tau_n} \mid \mathcal{F}_{t-1}\,] \leq \liminf_{n\uparrow\infty} Z_{t-1}^{\tau_n} = Z_{t-1}.$$

We get in particular that $E_Q[\, Z_t\,] \leq Z_0 < \infty$. Thus $Z_t \in \mathcal{L}^1(Q)$, and the assertion follows. □

Definition 9.7. By $\mathcal{P}_{\mathcal{S}}$ we denote the class of all probability measures $\widetilde{P} \approx P$ such that

$$X_t \in \mathcal{L}^1(\widetilde{P}) \quad \text{for all } t, \tag{9.2}$$

and such that the value process of any trading strategy in $\overline{\mathcal{S}}$ is a local $\widetilde{P}$-supermartingale.

Remark 9.8. If $\overline{\mathcal{S}}$ contains all self-financing trading strategies $\overline{\xi} = (\xi^0, \xi)$ with bounded ξ, then $\mathcal{P}_{\mathcal{S}}$ coincides with the class $\mathcal{P}$ of all equivalent martingale measures. To prove this, let $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$, and note that the value process V of any such $\overline{\xi}$ is a $\widetilde{P}$-supermartingale by (9.2) and by Proposition 9.6. The same applies to the strategy $-\overline{\xi}$, so V is in fact a $\widetilde{P}$-martingale, and Theorem 5.15 shows that $\widetilde{P}$ is a martingale measure for X. ◇

The following result extends the "fundamental theorem of asset pricing" to our present setting; see Theorem 5.17.

Theorem 9.9. *There are no arbitrage opportunities in $\overline{\mathcal{S}}$ if and only if $\mathcal{P}_{\mathcal{S}}$ is non-empty. In this case, there exists a measure $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$ which has a bounded density $d\widetilde{P}/dP$.*

We will first show that the condition $\mathcal{P}_{\mathcal{S}} \neq \emptyset$ implies the absence of arbitrage opportunities in $\overline{\mathcal{S}}$:

Proof of sufficiency. Suppose $\widetilde{P}$ is a measure in $\mathcal{P}_{\mathcal{S}}$, and V is the value process of a trading strategy in $\overline{\mathcal{S}}$ such that $V_T \geq 0$ P-almost surely. Combining Lemma 9.10 below with Proposition 9.6 shows that V is a $\widetilde{P}$-supermartingale. Hence $V_0 \geq \widetilde{E}[\, V_T\,]$, so V cannot be the value process of an arbitrage opportunity. □

Lemma 9.10. *Suppose that $\mathcal{P}_{\mathcal{S}} \neq \emptyset$ and that V is the value process of a trading strategy in $\overline{\mathcal{S}}$ such that $V_T \geq 0$ P-almost surely. Then $V_t \geq 0$ P-a.s. for all t.*

Proof. The assertion will be proved by backward induction on t. We have $V_T \geq 0$ by assumption, so let us assume that $V_t \geq 0$ P-a.s. for some t. For $\overline{\xi} = (\xi^0, \xi) \in \overline{\mathcal{S}}$ with value process V, we let $\xi_s^{(c)} := \xi_s \, \mathrm{I}_{\{|\xi_s| \leq c\}}$ for $c > 0$ and for all s. Then the value process $V^{(c)}$ of $\xi^{(c)}$ is a $\widetilde{P}$-supermartingale for any fixed $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$. Furthermore,

$$\begin{aligned} V_{t-1} \, \mathrm{I}_{\{|\xi_t| \leq c\}} &= V_t \, \mathrm{I}_{\{|\xi_t| \leq c\}} - \xi_t^{(c)} \cdot (X_t - X_{t-1}) \\ &\geq -\xi_t^{(c)} \cdot (X_t - X_{t-1}) \\ &= V_{t-1}^{(c)} - V_t^{(c)}. \end{aligned}$$

The last term on the right belongs to $\mathcal{L}^1(\widetilde{P})$, so we may take the conditional expectation $\widetilde{E}[\,\cdot \mid \mathcal{F}_{t-1}\,]$ on both sides of the inequality. We get

$$V_{t-1} \, \mathrm{I}_{\{|\xi_t| \leq c\}} \geq \widetilde{E}[\, V_{t-1}^{(c)} - V_t^{(c)} \mid \mathcal{F}_{t-1}\,] \geq 0 \quad \widetilde{P}\text{-a.s.}$$

By letting $c \uparrow \infty$, we obtain $V_{t-1} \geq 0$. □

Let us now prepare for the proof that the condition $\mathcal{P}_{\mathcal{S}} \neq \emptyset$ is necessary. First we show that the absence of arbitrage opportunities in $\overline{\mathcal{S}}$ is equivalent to the absence of arbitrage opportunities in each of the embedded one-period models, i.e., to the non-existence of $\xi_t \in \mathcal{S}_t$ such that $\xi_t \cdot (X_t - X_{t-1})$ amounts to a non-trivial positive gain. This observation will allow us to apply the techniques of Section 1.6. Let us denote

$$\mathcal{S}^\infty := \{\, \xi \in \mathcal{S} \mid \xi \text{ is bounded}\,\}.$$

Similarly, we define

$$\mathcal{S}_t^\infty := \{\, \xi_t \mid \xi \in \mathcal{S}^\infty \,\} = \mathcal{S}_t \cap L^\infty(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d).$$

Lemma 9.11. *The following conditions are equivalent:*

(a) *There exists an arbitrage opportunity in* $\overline{\mathcal{S}}$.

(b) *There exist* $t \in \{1, \dots, T\}$ *and* $\xi_t \in \mathcal{S}_t$ *such that*

$$\xi_t \cdot (X_t - X_{t-1}) \ge 0 \text{ } P\text{-a.s.}, \quad \text{and} \quad P[\, \xi_t \cdot (X_t - X_{t-1}) > 0\,] > 0. \tag{9.3}$$

(c) *There exist* $t \in \{1, \dots, T\}$ *and* $\xi_t \in \mathcal{S}_t^\infty$ *which satisfies* (9.3).

Proof. The proof is essentially the same as the one of Proposition 5.11. □

In order to apply the results of Section 1.6, we introduce the convex sets

$$\mathcal{K}_t^{\mathcal{S}} := \big\{\, \xi_t \cdot (X_t - X_{t-1}) \mid \xi_t \in \mathcal{S}_t \big\},$$

for $t \in \{1, \dots, T\}$. Lemma 9.11 shows that $\overline{\mathcal{S}}$ contains no arbitrage opportunities if and only if the condition

$$\mathcal{K}_t^{\mathcal{S}} \cap L_+^0 = \{0\} \tag{9.4}$$

holds for all $t \in \{1, \dots, T\}$.

Lemma 9.12. *Condition* (9.4) *implies that* $\mathcal{K}_t^{\mathcal{S}} - L_+^0(\Omega, \mathcal{F}_t, P)$ *is a closed convex subset of* $L^0(\Omega, \mathcal{F}_t, P)$.

Proof. The proof is essentially the same as the one of Lemma 1.67. Only the following additional observation is required: If (ξ^n) is sequence in $\mathcal{S}_t$, and if α and σ are two $\mathcal{F}_{t-1}$-measurable random variables such that $0 \le \alpha \le 1$ and σ is integer-valued, then $\zeta := \alpha \xi^\sigma \in \mathcal{S}_t$. Indeed, predictable convexity of $\mathcal{S}$ implies that

$$\alpha \sum_{k=1}^n \mathrm{I}_{\{\sigma = k\}} \xi^k \in \mathcal{S}_t$$

for each n, and the closedness of $\mathcal{S}_t$ in $L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$ yields

$$\zeta = \alpha \sum_{k=1}^\infty \mathrm{I}_{\{\sigma = k\}} \xi^k \in \mathcal{S}_t.$$

□

From now on, we will assume that

$$E[\,|X_s|\,] < \infty \quad \text{for all } s. \tag{9.5}$$

For the purpose of proving Theorem 9.9, this can be assumed without loss of generality: If (9.5) does not hold, then we replace P by an equivalent measure P' which has a bounded density dP'/dP and for which the price process X is integrable. For instance, we can take

$$dP' = c\,\exp\Big[-\sum_{s=1}^{T}|X_s|\Big]\,dP,$$

where c denotes the normalizing constant. If there exist a measure $\widetilde{P} \approx P'$ such that each value process for a strategy in $\overline{\mathcal{S}}$ is a local $\widetilde{P}$-supermartingale and such that the density $d\widetilde{P}/dP'$ is bounded, then $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$, and the density $d\widetilde{P}/dP$ is bounded as well.

Lemma 9.13. *If $\overline{\mathcal{S}}$ contains no arbitrage opportunities, then for each $t \in \{1,\dots,T\}$ there exists some $Z_t \in L^\infty(\Omega,\mathcal{F}_t,P)$ such that $Z_t > 0$ P-a.s. and such that*

$$E[\,Z_t\;\xi_t\cdot(X_t - X_{t-1})\,] \le 0 \quad \textit{for all } \xi \in \mathcal{S}^\infty. \tag{9.6}$$

Proof. Consider the positive cone

$$\overline{\mathcal{R}} := \{\,\lambda\,\overline{\xi} \mid \overline{\xi} \in \overline{\mathcal{S}},\ \lambda \ge 0\,\}$$

generated by $\overline{\mathcal{S}}$. Accordingly, we define the cones $\mathcal{R}$, $\mathcal{R}_t$, and $\mathcal{K}_t^{\mathcal{R}}$. Clearly, $\overline{\mathcal{R}}$ does not contain arbitrage opportunities if and only if $\overline{\mathcal{S}}$ is arbitrage-free. Hence, for each t

$$\mathcal{K}_t^{\mathcal{R}} \cap L^0_+(\Omega,\mathcal{F}_t,P) = \{0\} \tag{9.7}$$

by Lemma 9.11. Thus, Lemma 9.12 implies that each

$$\mathcal{C}_t^{\mathcal{R}} := \big(\mathcal{K}_t^{\mathcal{R}} - L^0_+(\Omega,\mathcal{F}_T,P)\big) \cap L^1$$

is a closed convex cone in L^1 which contains $-L^0_+(\Omega,\mathcal{F}_t,P)$. Furthermore, it follows from (9.7) that $\mathcal{C}_t^{\mathcal{R}} \cap L^0_+ = \{0\}$, so $\mathcal{C}_t^{\mathcal{R}}$ satisfies the assumptions of the Kreps–Yan theorem, which is stated in Theorem 1.61. We conclude that there exist $Z_t \in L^\infty(\Omega,\mathcal{F}_t,P)$ such that $P[\,Z_t > 0\,] = 1$, and such that $E[\,Z_t\,W\,] \le 0$ for each $W \in \mathcal{C}_t^{\mathcal{R}}$. As $\xi_t\cdot(X_t - X_{t-1}) \in \mathcal{C}_t^{\mathcal{R}}$ for each $\xi \in \mathcal{S}^\infty$, Z_t has property (9.6). □

Now we can complete the proof of Theorem 9.9 by showing that the absence of arbitrage opportunities in $\overline{\mathcal{S}}$ implies the existence of a measure $\widetilde{P}$ that belongs to the class $\mathcal{P}_{\mathcal{S}}$ and has a bounded density $d\widetilde{P}/dP$.

Proof of necessity in Theorem 9.9. Suppose that $\overline{\mathcal{S}}$ does not contain arbitrage opportunities. We are going to construct the desired measure $\widetilde{P}$ via backward recursion. First we consider the case $t = T$. Take a bounded random variable $Z_T > 0$ as constructed in Lemma 9.13, and define a probability measure $\widetilde{P}_T$ by

$$\frac{d\widetilde{P}_T}{dP} = \frac{Z_T}{E[\, Z_T \,]}.$$

Clearly, $\widetilde{P}_T$ is equivalent to P, and $X_t \in \mathcal{L}^1(\widetilde{P}_T)$ for all t. We claim that

$$\widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \le 0 \quad \text{for all } \xi \in \mathcal{S}^\infty. \tag{9.8}$$

To prove this claim, consider the family

$$\Phi := \big\{ \, \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \,|\, \mathcal{F}_{T-1} \,] \mid \xi \in \mathcal{S}^\infty \big\}.$$

For $\xi, \widetilde{\xi} \in \mathcal{S}^\infty$, let

$$A := \big\{ \, \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] > \widetilde{E}_T[\, \widetilde{\xi}_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \big\},$$

and define ξ' by $\xi'_t = 0$ for $t < T$ and

$$\xi'_T := \xi_T \mathrm{I}_A + \widetilde{\xi}_T \mathrm{I}_{A^c}.$$

The predictable convexity of $\mathcal{S}$ implies that $\xi' \in \mathcal{S}^\infty$. Furthermore, we have

$$\begin{aligned} &\widetilde{E}_T[\, \xi'_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \\ &\qquad = \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \vee \widetilde{E}_T[\, \widetilde{\xi}_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,]. \end{aligned}$$

Hence, the family Φ is directed upwards in the sense of Theorem A.32. By virtue of that theorem, $\operatorname{ess\,sup} \Phi$ is the increasing limit of a sequence in Φ. By monotone convergence, we get

$$\begin{aligned} &\widetilde{E}_T\Big[\operatorname*{ess\,sup}_{\xi \in \mathcal{S}^\infty} \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \Big] \\ &\qquad = \sup_{\xi \in \mathcal{S}^\infty} \widetilde{E}_T\Big[\, \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] \Big] \\ &\qquad = \frac{1}{E[\, Z_T \,]} \sup_{\xi \in \mathcal{S}^\infty} E[\, \xi_T \cdot (X_T - X_{T-1}) \, Z_T \,] \\ &\qquad \le 0, \end{aligned} \tag{9.9}$$

where we have used (9.6) in the last step. Since $\mathcal{S}$ contains 0, it follows that

$$\operatorname*{ess\,sup}_{\xi \in \mathcal{S}^\infty} \widetilde{E}_T[\, \xi_T \cdot (X_T - X_{T-1}) \mid \mathcal{F}_{T-1} \,] = 0 \quad \widetilde{P}_T\text{-a.s.},$$

which yields our claim (9.8).

Now we apply the previous argument inductively: Suppose we already have a probability measure $\widetilde{P}_{t+1} \approx P$ with a bounded density $d\widetilde{P}_{t+1}/dP$ such that

$$\widetilde{E}_{t+1}[\,|X_s|\,] < \infty \quad \text{for all } s,$$

and such that

$$\widetilde{E}_{t+1}[\,\xi_k \cdot (X_k - X_{k-1}) \mid \mathcal{F}_{k-1}\,] \leq 0 \quad P\text{-a.s. for } k \geq t+1 \text{ and } \xi \in \mathcal{S}^\infty. \tag{9.10}$$

Then we may apply Lemma 9.13 with P replaced by $\widetilde{P}_{t+1}$, and we get some strictly positive $\widetilde{Z}_t \in L^\infty(\Omega, \mathcal{F}_t, \widetilde{P}_{t+1})$ satisfying (9.6) with $\widetilde{P}_{t+1}$ in place of P. We now proceed as in the first step by defining a probability measure $\widetilde{P}_t \approx \widetilde{P}_{t+1} \approx P$ as

$$\frac{d\widetilde{P}_t}{d\widetilde{P}_{t+1}} = \frac{\widetilde{Z}_t}{\widetilde{E}_{t+1}[\,\widetilde{Z}_t\,]}.$$

Then $\widetilde{P}_t$ has bounded densities with respect to both $\widetilde{P}_{t+1}$ and P. In particular, $\widetilde{E}_t[\,|X_s|\,] < \infty$ for all s. Moreover, the $\mathcal{F}_t$-measurability of $d\widetilde{P}_t/d\widetilde{P}_{t+1}$ implies that (9.10) is satisfied for $\widetilde{P}_t$ replacing $\widetilde{P}_{t+1}$. Repeating the arguments that led to (9.8) yields

$$\widetilde{E}_t[\,\xi_t \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1}\,] \leq 0 \quad \text{for all } \xi \in \mathcal{S}^\infty.$$

After T steps, we arrive at the desired measure $\widetilde{P} := \widetilde{P}_1 \in \mathcal{P}_{\mathcal{S}}$. □

9.2 Uniform Doob decomposition

The goal of this section is to characterize those non-negative adapted processes U which can be decomposed as

$$U_t = U_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) - B_t, \tag{9.11}$$

where the predictable d-dimensional process ξ belongs to $\mathcal{S}$, and where B is an adapted and increasing process such that $B_0 = 0$. In the unconstrained case where $\mathcal{S}$ consists of all strategies, we have seen in Section 7.2 that such a decomposition exists if and only if U is a supermartingale under each equivalent martingale measure $P^* \in \mathcal{P}$. In our present context, a first guess might be that the role of $\mathcal{P}$ is now played by $\mathcal{P}_{\mathcal{S}}$. Since each value process of a strategy in $\overline{\mathcal{S}}$ is a local $\widetilde{P}$-supermartingale for each $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$, any process U which has a decomposition (9.11) is also a local $\widetilde{P}$-supermartingale for $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$. Thus, one might suspect that the latter property would also be sufficient for the existence of a decomposition (9.11). This, however, is not the case, as is illustrated by the following simple example.

Example 9.14. Consider a one-period market model with the riskless bond $S_0^0 \equiv S_1^0 \equiv 1$ and with one risky asset S^1. We assume that $S_0^1 \equiv 1$ and that S_1^1 takes the values $S_1^1(\omega^-) = \frac{1}{2}$ and $S_1^1(\omega^+) = \frac{3}{2}$ on $\Omega := \{\omega^-, \omega^+\}$. We choose any measure P on Ω which assigns positive mass to both ω^+ and ω^-. If we let $\mathcal{S} = [0, 1]$, then a measure $\widetilde{P}$ belongs to $\mathcal{P}_\mathcal{S}$ if and only if $\widetilde{P}[\{\omega^+\}] \in \big(0, \frac{1}{2}\big]$. Thus, for any positive initial value U_0, the process defined by $U_1(\omega^-) := 0$ and $U_1(\omega^+) := 2U_0$ is a $\mathcal{P}_\mathcal{S}$-supermartingale. If U can be decomposed according to (9.11), then we must be able to write

$$2U_0 = U_1(\omega^+) = U_0 + \xi \cdot \big(S_1^1(\omega^+) - S_0^1(\omega^+)\big) - B_1(\omega^+)$$

for some $B_1(\omega^+) \geq 0$. This requirement is equivalent to $U_0 \leq \xi/2$. Hence the decomposition (9.11) fails for $U_0 > 1/2$. ◇

The reason for the failure of the decomposition (9.11) for certain $\mathcal{P}_\mathcal{S}$-supermartingales is that $\mathcal{P}_\mathcal{S}$ does not reflect the full structure of $\mathcal{S}$; the definition of $\mathcal{P}_\mathcal{S}$ depends only on the cone

$$\{\lambda\xi \mid \lambda > 0,\ \xi \in \mathcal{S}\}$$

generated by $\mathcal{S}$. In the approach we are going to present here, the structure of $\mathcal{S}$ will be reflected by a stochastic process which we associate to any measure $Q \ll P$:

Definition 9.15. For a measure $Q \ll P$, the *upper variation process* for $\mathcal{S}$ is the increasing process A^Q defined by

$$A_0^Q := 0 \quad \text{and} \quad A_{t+1}^Q - A_t^Q := \operatorname*{ess\,sup}_{\xi \in \mathcal{S}} \big[\xi_{t+1} \cdot \big(E_Q[\, X_{t+1} \mid \mathcal{F}_t\,] - X_t\big)\big]$$

for $t = 0, \dots, T-1$. By $\mathcal{Q}_\mathcal{S}$ we denote the set of all $Q \approx P$ such that

$$E_Q[\, A_T^Q\,] < \infty$$

and such that

$$E_Q[\, |X_{t+1} - X_t| \mid \mathcal{F}_t\,] < \infty \quad P\text{-a.s. for all } t.$$

Clearly, the upper variation process of any measure $Q \approx P$ satisfies

$$A_{t+1}^Q - A_t^Q = \operatorname*{ess\,sup}_{\xi \in \mathcal{S}^\infty} \big[\xi_{t+1} \cdot \big(E_Q[\, X_{t+1} \mid \mathcal{F}_t\,] - X_t\big)\big],$$

where $\mathcal{S}^\infty$ are the bounded processes in $\mathcal{S}$. Hence for $Q \in \mathcal{Q}_\mathcal{S}$ and $\xi \in \mathcal{S}^\infty$, the condition $E_Q[\, |X_{t+1} - X_t| \mid \mathcal{F}_t\,] < \infty$ guarantees that

$$\xi_{t+1} \cdot \big(E_Q[\, X_{t+1} \mid \mathcal{F}_t\,] - X_t\big) = E_Q[\, \xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t\,],$$

and it follows that

$$A_{t+1}^Q - A_t^Q = \operatorname*{ess\,sup}_{\xi \in \mathcal{S}^\infty} E_Q[\, \xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t\,] \quad \text{for } Q \in \mathcal{Q}_\mathcal{S}. \tag{9.12}$$

In particular, we have

$$A_T^{\widetilde{P}} = 0 \quad \text{for } \widetilde{P} \in \mathcal{P}_{\mathcal{S}}, \tag{9.13}$$

which implies the inclusion

$$\mathcal{P} \subset \mathcal{P}_{\mathcal{S}} \subset \mathcal{Q}_{\mathcal{S}}. \tag{9.14}$$

Proposition 9.16. *If $Q \in \mathcal{Q}_{\mathcal{S}}$, and V is the value process of a trading strategy in $\overline{\mathcal{S}}$, then $V - A^Q$ is a local Q-supermartingale.*

Proof. Let V be the value process of $\overline{\xi} = (\xi^0, \xi) \in \overline{\mathcal{S}}$. Denote by $\tau_n(\omega)$ the first time t at which

$$|\xi_{t+1}(\omega)| > n \quad \text{or} \quad E_Q\big[\, |X_{t+1} - X_t| \mid \mathcal{F}_t \,\big](\omega) > n.$$

If such a t does not exist, let $\tau_n(\omega) := T$. Then τ_n is a stopping time. Since

$$|V_{t+1}^{\tau_n} - V_t^{\tau_n}| \le \mathrm{I}_{\{\tau_n \ge t+1\}} |\xi_{t+1}| \cdot |X_{t+1} - X_t| \in \mathcal{L}^1(Q),$$

$V_t^{\tau_n}$ belongs to $\mathcal{L}^1(Q)$, and

$$\begin{aligned} E_Q[\, V_{t+1}^{\tau_n} - V_t^{\tau_n} \mid \mathcal{F}_t \,] &= \mathrm{I}_{\{\tau_n \ge t+1\}} \xi_{t+1} \cdot \big(E_Q[\, X_{t+1} \mid \mathcal{F}_t \,] - X_t \big) \\ &\le (A^Q)_{t+1}^{\tau_n} - (A^Q)_t^{\tau_n}. \end{aligned}$$

This proves that $V^{\tau_n} - (A^Q)^{\tau_n}$ is a Q-supermartingale. □

Let us identify the class $\mathcal{Q}_{\mathcal{S}}$ in some special cases.

Remark 9.17. If $\mathcal{S}^\infty$ consists of all bounded predictable processes ξ with non-negative components, then $\mathcal{Q}_{\mathcal{S}} = \mathcal{P}_{\mathcal{S}}$. To prove this, take $Q \in \mathcal{Q}_{\mathcal{S}}$, and note first that $A^Q \equiv 0$, due to (9.12) and the fact that $\mathcal{S}$ is a cone. Thus, value processes of strategies in $\overline{\mathcal{S}}$ are local Q-supermartingales by Proposition 9.16. By taking $\xi \in \mathcal{S}$ such that $\xi_t^i \equiv 1$ and $\xi_t^j \equiv 0$ for $j \neq i$, we get that X^i is a local Q-supermartingale, and Proposition 9.6 implies that X^i is a Q-supermartingale. In particular, X_t^i is Q-integrable, and we conclude $Q \in \mathcal{P}_{\mathcal{S}}$. ◇

Remark 9.18. If $\mathcal{S}^\infty$ consists of all bounded predictable processes ξ, then $\mathcal{Q}_{\mathcal{S}} = \mathcal{P}$. This follows by combining Remarks 9.8 and 9.17. ◇

Example 9.19. Suppose our market model contains just one risky asset, and $\mathcal{S}$ consists of all predictable processes ξ such that

$$a_t \le \xi_t \le b_t \quad P\text{-a.s. for all } t,$$

where a and b are two given predictable processes with

$$-\infty < a_t \le 0 \le b_t < \infty \quad P\text{-a.s.}$$

If we assume in addition that $E[\,|X_{t+1}-X_t|\mid \mathcal{F}_t\,] > 0$ P-a.s., then the non-redundance condition (9.1) holds, and $\mathcal{S}$ satisfies the assumptions (a) through (d) stated at the beginning of this chapter. If $Q \approx P$ is any probability measure such that

$$E_Q\big[\,|X_{t+1} - X_t| \mid \mathcal{F}_t\,\big] < \infty \quad P\text{-a.s. for all } t, \tag{9.15}$$

then

$$\begin{aligned}
&\operatorname*{ess\,sup}_{\xi \in \mathcal{S}^\infty} E_Q[\,\xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t\,] \\
&\qquad = b_{t+1}\big(E_Q[\,X_{t+1} - X_t \mid \mathcal{F}_t\,]\big)^+ - a_{t+1}\big(E_Q[\,X_{t+1} - X_t \mid \mathcal{F}_t\,]\big)^- \\
&\qquad < \infty \quad P\text{-a.s.}
\end{aligned}$$

Hence, $\mathcal{Q}_{\mathcal{S}}$ consists of all measures $Q \approx P$ with (9.15). ◇

We now state the uniform Doob decomposition under constraints, which is the main result of this section.

Theorem 9.20. *Suppose that $\mathcal{P}_{\mathcal{S}}$ is non-empty. Then for any adapted process U with $U_T \geq 0$ P-a.s., the following conditions are equivalent.*

(a) *$U - A^Q$ is a Q-supermartingale for every $Q \in \mathcal{Q}_{\mathcal{S}}$.*

(b) *There exists $\xi \in \mathcal{S}$ and an adapted increasing process B such that $B_0 = 0$ and*

$$U_t = U_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) - B_t \quad P\text{-a.s. for all } t.$$

Proof. (b)⇒(a): Fix $Q \in \mathcal{Q}_{\mathcal{S}}$. According to Proposition 9.16, the process

$$M_t^Q := U_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}) - A_t^Q$$

is a local Q-supermartingale. Since $M_T^Q + A_T^Q \geq U_T \geq 0$ P-a.s., we get from Lemma 9.10 that the negative part of M_t^Q is bounded below by $-A_t^Q \in \mathcal{L}^1(Q)$. So M^Q is a Q-supermartingale by Proposition 9.6. Since

$$M_T^Q \geq M_T^Q - B_T = U_T - A_T^Q \geq -A_T^Q \quad P\text{-a.s.},$$

and since B is increasing, each B_t belongs to $\mathcal{L}^1(Q)$, and $M - B = U - A^Q$ is a Q-supermartingale.

(a)⇒(b): We must show that for any given $t \in \{1, \dots, T\}$ there exist some $\xi \in \mathcal{S}$ and a non-negative random variable R_t playing the role of $B_t - B_{t-1}$ such that $U_t - U_{t-1} = \xi_t \cdot (X_t - X_{t-1}) - R_t$, i.e.,

$$U_t - U_{t-1} \in \mathcal{K}_t^{\mathcal{S}} - L_+^0(\Omega, \mathcal{F}_t, P),$$

where

$$\mathcal{K}_t^{\mathcal{S}} = \{\, \xi_t \cdot (X_t - X_{t-1}) \mid \xi \in \mathcal{S} \,\}.$$

The formulation of this problem does not change if we switch from P to any equivalent probability measure, so we can assume without loss of generality that $P \in \mathcal{P}_{\mathcal{S}}$. In this case $A^P \equiv 0$, and U is a P-supermartingale. In particular, $U_s \in L^1(P)$ for all s.

We assume by way of contradiction that

$$U_t - U_{t-1} \notin \mathcal{C}_t^{\mathcal{S}} = \big(\mathcal{K}_t^{\mathcal{S}} - L_+^0(\Omega, \mathcal{F}_t, P)\big) \cap L^1(P). \tag{9.16}$$

Recall that we have proved in Lemma 9.12 that $\mathcal{C}_t^{\mathcal{S}}$ is a closed convex subset of $L^1(\Omega, \mathcal{F}_t, P)$. The Hahn–Banach separation theorem, Theorem A.56, now implies the existence of a random variable $Z \in L^\infty(\Omega, \mathcal{F}_t, P)$ such that

$$\alpha := \sup_{W \in \mathcal{C}_t^{\mathcal{S}}} E[\, Z\, W\,] < E[\, Z\,(U_t - U_{t-1})\,] =: \beta < \infty. \tag{9.17}$$

Note that the function $-\lambda \mathrm{I}_{\{Z<0\}}$ belongs to $\mathcal{C}_t^{\mathcal{S}}$ for all $\lambda \geq 0$. Thus

$$0 \leq (-\lambda)\, E\big[\, Z\, \mathrm{I}_{\{Z<0\}}\,\big] \leq \alpha$$

for every $\lambda \geq 0$, and it follows that $Z \geq 0$ P-almost surely.

In fact, we can always modify Z such that it is bounded below by some $\varepsilon > 0$ and still satisfies (9.17). To see this, note first that every $W \in \mathcal{C}_t^{\mathcal{S}}$ is dominated by some $\xi_t \cdot (X_t - X_{t-1}) \in \mathcal{K}_t^{\mathcal{S}}$ with integrable negative part. Therefore

$$E[\, W\,] \leq E[\, \xi_t \cdot (X_t - X_{t-1})\,] \leq \liminf_{c \uparrow \infty} E\big[\, \xi_t \cdot (X_t - X_{t-1}) \mathrm{I}_{\{|\xi_t| \leq c\}}\,\big] \leq 0,$$

where we have used our assumption that $P \in \mathcal{P}_{\mathcal{S}}$. If we let $Z^\varepsilon := \varepsilon\, 1 + (1-\varepsilon) Z$, then Z^ε still satisfies $E[\, Z^\varepsilon\, W\,] \leq 0$ for all $W \in \mathcal{C}_t^{\mathcal{S}}$, and for ε small enough, the expectation $E[\, Z^\varepsilon\, (U_t - U_{t-1})\,]$ is still larger than α. So Z^ε also satisfies (9.17). Therefore, we may assume from now on that our Z with (9.17) is bounded below by some constant $\varepsilon > 0$.

For the next step, let $Z_{t-1} := E[\, Z \mid \mathcal{F}_{t-1}\,]$ and

$$\frac{dQ}{dP} := \frac{Z}{Z_{t-1}}.$$

Since this density is bounded and since $P \in \mathcal{P}_{\mathcal{S}}$, we get

$$E_Q\big[\, |X_s - X_{s-1}| \mid \mathcal{F}_{s-1}\,\big] < \infty \quad P\text{-a.s. for all } s. \tag{9.18}$$

Moreover, it is not difficult to check that

$$E_Q[\, \xi_s \cdot (X_s - X_{s-1}) \mid \mathcal{F}_{s-1}\,] = E[\, \xi_s \cdot (X_s - X_{s-1}) \mid \mathcal{F}_{s-1}\,] \leq 0 \quad \text{for } s \neq t; \tag{9.19}$$

see the proof of Theorem 7.5 for details. We now consider the case $s = t$. As we have seen in the proof of Theorem 9.9, the family

$$\left\{ E_Q[\, \xi_t \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] \mid \xi \in \mathcal{S} \right\}$$

is directed upwards. Therefore, we may conclude as in (9.9) that

$$\begin{aligned} E_Q\Big[\, Z_{t-1} \operatorname*{ess\,sup}_{\xi \in \mathcal{S}} E_Q[\, \xi_t \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] \Big] \\ &= \sup_{\xi \in \mathcal{S}} E[\, \xi_t \cdot (X_t - X_{t-1})\, Z \,] \\ &\le \alpha. \end{aligned} \tag{9.20}$$

Since $Z_{t-1} \ge \varepsilon$, (9.20) implies that

$$E_Q\big[\, A_t^Q - A_{t-1}^Q \,\big] = E_Q\Big[\operatorname*{ess\,sup}_{\xi \in \mathcal{S}} E_Q[\, \xi_t \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] \Big] \le \frac{\alpha}{\varepsilon}.$$

By (9.19) we may conclude that $E_Q[\, A_T^Q \,] \le \alpha/\varepsilon$, and so (9.18) yields $Q \in \mathcal{Q}_{\mathcal{S}}$.

As a final step, we show that $U - A^Q$ cannot be a Q-supermartingale, thus leading our assumption (9.16) to a contradiction with our hypothesis (a). To this end, we use again (9.20):

$$\begin{aligned} E_Q\big[\, Z_{t-1}\, E_Q[\, U_t - U_{t-1} \mid \mathcal{F}_{t-1} \,] \big] \\ &= E_Q[\, Z_{t-1}\, (U_t - U_{t-1}) \,] \\ &= E[\, Z\, (U_t - U_{t-1}) \,] = \beta \\ &> \alpha \\ &\ge E_Q\Big[\, Z_{t-1} \operatorname*{ess\,sup}_{\xi \in \mathcal{S}} E_Q[\, \xi_t \cdot (X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,] \Big] \\ &= E_Q\big[Z_{t-1}\, (A_t^Q - A_{t-1}^Q) \big]. \end{aligned}$$

Thus, we cannot have

$$E_Q[\, U_t - U_{t-1} \mid \mathcal{F}_{t-1} \,] \le A_t^Q - A_{t-1}^Q \quad P\text{-a.s.},$$

so $U - A^Q$ cannot be a Q-supermartingale, in contradiction to our hypothesis (a). □

9.3 Upper Snell envelopes

From now on, we assume that $\overline{\mathcal{S}}$ does not contain arbitrage opportunities, which is equivalent to the condition $\mathcal{P}_{\mathcal{S}} \neq \emptyset$. Let H be a discounted American claim. Our goal is to construct a superhedging strategy for H that belongs to our class $\overline{\mathcal{S}}$ of admissible strategies. The uniform Doob decomposition suggests that we should find an adapted process $U \geq H$ such that $U - A^Q$ is a Q-supermartingale for each $Q \in \mathcal{Q}_{\mathcal{S}}$. If we consider only one such Q, then the minimal process U which satisfies these requirements is given by $\widetilde{U}^Q + A^Q$, where

$$\widetilde{U}_t^Q := \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E_Q[\, H_\tau - A_\tau^Q \mid \mathcal{F}_t \,] \quad t = 0, \dots, T, \tag{9.21}$$

is the Snell envelope of $H - A^Q$ with respect to Q. Thus, one may guess that

$$\operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} (\widetilde{U}_t^Q + A_t^Q), \qquad t = 0, \dots, T,$$

is the minimal process U which dominates H and for which $U - A^Q$ is a Q-supermartingale for each $Q \in \mathcal{Q}_{\mathcal{S}}$. Let us assume that

$$\sup_{Q \in \mathcal{Q}_{\mathcal{S}}} \widetilde{U}_0^Q = \sup_{Q \in \mathcal{Q}_{\mathcal{S}}} \sup_{\tau \in \mathcal{T}} E_Q[\, H_\tau - A_\tau^Q \,] < \infty.$$

Note that this condition holds if H is bounded.

Definition 9.21. The process

$$\begin{aligned} \widetilde{U}_t^{\uparrow} &:= \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(A_t^Q + \widetilde{U}_t^Q \big) \\ &= \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(A_t^Q + \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E_Q[\, H_\tau - A_\tau^Q \mid \mathcal{F}_t \,] \big), \qquad t = 0, \dots, T, \end{aligned}$$

will be called the *upper $\mathcal{Q}_{\mathcal{S}}$-Snell envelope* of H.

The main result of this section confirms our guess that $\widetilde{U}^{\uparrow}$ is the process we are looking for.

Theorem 9.22. *The upper $\mathcal{Q}_{\mathcal{S}}$-Snell envelope of H is the smallest process $U \geq H$ such that $U - A^Q$ is a Q-supermartingale for each $Q \in \mathcal{Q}_{\mathcal{S}}$.*

For a European claim, we have the following additional result.

Proposition 9.23. *For a discounted European claim H^E with*

$$\sup_{Q \in \mathcal{Q}_{\mathcal{S}}} E_Q[\, H^E - A_T^Q \,] < \infty,$$

the upper $\mathcal{Q}_{\mathcal{S}}$-Snell envelope takes the form

$$\widetilde{U}_t^{\uparrow} = \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\, H^E - A_T^Q \mid \mathcal{F}_t \,] + A_t^Q \big), \quad t = 0, \dots, T.$$

Proposition 9.23 will follow from Lemma 9.28 below. The next result provides a scheme for the recursive calculation of $\widetilde{U}^{\uparrow}$. It will be used in the proof of Theorem 9.22.

Proposition 9.24. *For fixed $Q_0 \in \mathcal{Q}_{\mathcal{S}}$ let $\mathcal{Q}_t(Q_0)$ denote the set of all $Q \in \mathcal{Q}_{\mathcal{S}}$ which coincide with Q_0 on $\mathcal{F}_t$. Then $\widetilde{U}^{\uparrow}$ satisfies the following recursion formula:*

$$\widetilde{U}_t^{\uparrow} - A_t^{Q_0} = (H_t - A_t^{Q_0}) \vee \underset{Q \in \mathcal{Q}_t(Q_0)}{\operatorname{ess\,sup}} \; E_Q[\, \widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^{Q} \mid \mathcal{F}_t \,], \quad t = 0, \dots, T-1.$$

The proofs of this proposition and of Theorem 9.22 will be given at the end of this section. Let us recall the following concepts from Section 6.4. The *pasting* of two probability measures $Q_1 \approx Q_2$ in a stopping time $\tau \in \mathcal{T} = \{\, \sigma \mid \sigma \text{ is a stopping time} \le T \,\}$ is the probability measure

$$\widetilde{Q}[\, A \,] = E_{Q_1}\big[\, Q_2[\, A \mid \mathcal{F}_\tau \,]\big], \quad A \in \mathcal{F}.$$

It was shown in Lemma 6.43 that, for all stopping times σ and $\mathcal{F}_T$-measurable $Y \ge 0$,

$$E_{\widetilde{Q}}[\, Y \mid \mathcal{F}_\sigma \,] = E_{Q_1}\big[\, E_{Q_2}[\, Y \mid \mathcal{F}_{\tau \vee \sigma} \,] \mid \mathcal{F}_\sigma \,\big]. \tag{9.22}$$

Recall also that a set $\mathcal{Q}$ of equivalent probability measures on $(\Omega, \mathcal{F})$ is called *stable* if for any pair $Q_1, Q_2 \in \mathcal{Q}$ and all $\tau \in \mathcal{T}$ the corresponding pasting also belongs to $\mathcal{Q}$. A technical inconvenience arises from the fact that our set $\mathcal{Q}_{\mathcal{S}}$ may *not* be stable. We must introduce a further condition on τ which guarantees that the pasting of $Q_1, Q_2 \in \mathcal{Q}_{\mathcal{S}}$ in τ also belongs to $\mathcal{Q}_{\mathcal{S}}$.

Lemma 9.25. *For $\tau \in \mathcal{T}$, the pasting $\widetilde{Q}$ of $Q_1, Q_2 \in \mathcal{Q}_{\mathcal{S}}$ in τ satisfies*

$$E_{\widetilde{Q}}[\, |X_{t+1} - X_t| \mid \mathcal{F}_t \,] < \infty \quad P\text{-a.s.},$$

and its upper variation process is given by

$$A_t^{\widetilde{Q}} = A_{t \wedge \tau}^{Q_1} + (A_t^{Q_2} - A_\tau^{Q_2}) \cdot \mathrm{I}_{\{\tau < t\}}.$$

Moreover, we have $\widetilde{Q} \in \mathcal{Q}_{\mathcal{S}}$ under the condition that there exists $\varepsilon > 0$ such that

$$\frac{dQ_2}{dQ_1}\bigg|_{\mathcal{F}_\tau} \ge \varepsilon \text{ a.s. on } \{\tau < T\}. \tag{9.23}$$

Proof. The identity (9.22) yields

$$E_{\widetilde{Q}}[\, |X_{t+1} - X_t| \mid \mathcal{F}_t \,] = E_{Q_1}[\, |X_{t+1} - X_t| \mid \mathcal{F}_t \,]\mathrm{I}_{\{\tau > t\}} + E_{Q_2}[\, |X_{t+1} - X_t| \mid \mathcal{F}_t \,]\mathrm{I}_{\{\tau \le t\}},$$

and each of the two conditional expectations is finite almost surely.

Now we will compute the upper variation process $A^{\widetilde{Q}}$ of $\widetilde{Q}$. As above, (9.22) yields

$$\begin{aligned} &E_{\widetilde{Q}}[\, \xi_t \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t \,] \\ &\qquad = E_{Q_1}[\, \xi_t \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t \,]\mathrm{I}_{\{\tau > t\}} + E_{Q_2}[\, \xi_t \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t \,]\mathrm{I}_{\{\tau \le t\}}. \end{aligned}$$

Taking the essential supremum over $\xi_t \in \mathcal{S}_t^\infty$ gives

$$A_{t+1}^{\widetilde{Q}} - A_t^{\widetilde{Q}} = (A_{t+1}^{Q_1} - A_t^{Q_1})\mathrm{I}_{\{\tau > t\}} + (A_{t+1}^{Q_2} - A_t^{Q_2})\mathrm{I}_{\{\tau \le t\}},$$

and from this our formula for $A_t^{\widetilde{Q}}$ follows.

In a final step, we show that $\widetilde{Q}$ belongs to $\mathcal{Q}_{\mathcal{S}}$ under condition (9.23). We must show that $E_{\widetilde{Q}}[\, A_T^{\widetilde{Q}} \,] < \infty$. Let Z_t denote the density of Q_2 with respect to Q_1 on $\mathcal{F}_t$. Then, by our formula for $A_T^{\widetilde{Q}}$,

$$\begin{aligned} E_{\widetilde{Q}}[\, A_T^{\widetilde{Q}} \,] &= E_{\widetilde{Q}}[\, A_\tau^{Q_1} + A_T^{Q_2} - A_\tau^{Q_2} \,] \\ &= E_{Q_1}[\, A_\tau^{Q_1} + E_{Q_2}[\, A_T^{Q_2} - A_\tau^{Q_2} \mid \mathcal{F}_\tau \,]\,] \\ &\le E_{Q_1}[\, A_T^{Q_1} \,] + E_{Q_1}\Big[\frac{1}{Z_\tau} E_{Q_1}[\, (A_T^{Q_2} - A_\tau^{Q_2})\, Z_T \mid \mathcal{F}_\tau \,] \Big] \\ &\le E_{Q_1}[\, A_T^{Q_1} \,] + \frac{1}{\varepsilon} E_{Q_2}[\, A_T^{Q_2} \,], \end{aligned}$$

which is finite for $Q_1, Q_2 \in \mathcal{Q}_{\mathcal{S}}$. □

Lemma 9.26. *Suppose we are given $Q_1, Q_2 \in \mathcal{Q}_{\mathcal{S}}$, a stopping time $\tau \in \mathcal{T}$, and a set $B \in \mathcal{F}_\tau$ such that $dQ_2/dQ_1|_{\mathcal{F}_\tau} \ge \varepsilon$ a.s. on B. Let $\widetilde{Q}$ be the pasting of Q_1 and Q_2 in the stopping time*

$$\sigma := \tau \mathrm{I}_B + T \mathrm{I}_{B^c}.$$

Then $\widetilde{Q} \in \mathcal{Q}_{\mathcal{S}}$, and the Snell envelopes associated with these three measures by (9.21) *are related as follows:*

$$\widetilde{U}_\tau^{\widetilde{Q}} + A_\tau^{\widetilde{Q}} = (\widetilde{U}_\tau^{Q_1} + A_\tau^{Q_1}) \cdot \mathrm{I}_{B^c} + (\widetilde{U}_\tau^{Q_2} + A_\tau^{Q_2}) \cdot \mathrm{I}_B \quad P\text{-a.s.} \tag{9.24}$$

Proof. We have $dQ_2/dQ_1|_{\mathcal{F}_\sigma} \ge \varepsilon$, hence $\widetilde{Q} \in \mathcal{Q}_{\mathcal{S}}$ follows from Lemma 9.25. Let ρ be a stopping time in the set $\mathcal{T}_\tau$ of all stopping times $\ge \tau$. The formula for $A^{\widetilde{Q}}$ in Lemma 9.25 yields

$$A_\rho^{\widetilde{Q}} = A_\tau^{Q_1} + (A_\rho^{Q_1} - A_\tau^{Q_1})\mathrm{I}_{B^c} + (A_\rho^{Q_2} - A_\tau^{Q_2})\mathrm{I}_B.$$

Moreover, (9.22) implies that

$$E_{\widetilde{Q}}[\, Y \mid \mathcal{F}_\tau \,] = E_{Q_1}[\, Y \mid \mathcal{F}_\tau \,]\mathrm{I}_{B^c} + E_{Q_2}[\, Y \mid \mathcal{F}_\tau \,]\mathrm{I}_B$$

for all random variables Y such that all conditional expectations make sense. Hence,

$$\begin{aligned} &E_{\widetilde{Q}}[\, H_\rho - A_\rho^{\widetilde{Q}} \mid \mathcal{F}_\tau \,] + A_\tau^{\widetilde{Q}} \\ &\quad = \big(E_{Q_1}[\, H_\rho - A_\rho^{Q_1} \mid \mathcal{F}_\tau \,] + A_\tau^{Q_1}\big)\, \mathrm{I}_{B^c} + \big(E_{Q_2}[\, H_\rho - A_\rho^{Q_2} \mid \mathcal{F}_\tau \,] + A_\tau^{Q_2}\big)\, \mathrm{I}_B. \end{aligned}$$

Whenever $\rho_1,\ \rho_2$ are stopping times in $\mathcal{T}_\tau$, then $\rho := \rho_1\, \mathrm{I}_{B^c} + \rho_2\, \mathrm{I}_B$ is also a stopping time in $\mathcal{T}_\tau$. Conversely, every $\rho \in \mathcal{T}_\tau$ can be written in that way for stopping times ρ_1 and ρ_2. Thus, taking the essential supremum over all ρ_1 and ρ_2 and applying Proposition 6.38 yields (9.24). □

In fact, we have $A_\tau^{\widetilde{Q}} = A_\tau^{Q_1}$ in (9.24), as we will have $A_\tau^{Q_k} = A_\tau^{Q_0}$ in the following lemma.

Lemma 9.27. *For any $Q_0 \in \mathcal{Q}_{\mathcal{S}}$, $\tau \in \mathcal{T}$, and $\delta > 0$, there exist a set $\Lambda_\delta \in \mathcal{F}_\tau$ such that $Q_0[\,\Lambda_\delta\,] \geq 1-\delta$ and measures $Q_k \in \mathcal{Q}_{\mathcal{S}}$ such that $Q_k = Q_0$ on $\mathcal{F}_\tau$ and*

$$\widetilde{U}_\tau^{Q_k} + A_\tau^{Q_k} \nearrow \operatorname*{ess\,sup}_{Q\in\mathcal{Q}_{\mathcal{S}}}(\widetilde{U}_\tau^{Q} + A_\tau^{Q}) = \widetilde{U}_\tau^{\uparrow} \quad P\text{-a.s. on } \Lambda_\delta.$$

Proof. By Theorem A.32 and its proof, there exists a sequence $(Q_n^0) \subset \mathcal{Q}_{\mathcal{S}}$ such that

$$\lim_{k\uparrow\infty} \max_{n\leq k} \big(\widetilde{U}_\tau^{Q_n^0} + A_\tau^{Q_n^0}\big) = \widetilde{U}_\tau^{\uparrow} \quad P\text{-a.s.}$$

We will recursively define measures $Q_k \in \mathcal{Q}_{\mathcal{S}}$ and sets $\Lambda_\delta^k \in \mathcal{F}_\tau$ such that $\Lambda_\delta^k \subset \Lambda_\delta^{k-1}$, $Q_0[\,\Lambda_\delta^k\,] \geq 1-(1-2^{-k})\delta$, and

$$\widetilde{U}_\tau^{Q_k} + A_\tau^{Q_k} = \max_{n\leq k} \big(\widetilde{U}_\tau^{Q_n^0} + A_\tau^{Q_n^0}\big) \quad P\text{-a.s. on } \Lambda_\delta^k.$$

By letting $\Lambda_\delta := \bigcap_k \Lambda_\delta^k$, this will imply the first part of the assertion. We start this recursion in $k=0$ by taking Q_0 and $\Lambda_\delta^0 := \Omega$.

For Q_k given, the equivalence of Q_k and Q_{k+1}^0 implies that there exists some $\varepsilon > 0$ such that the set

$$D := \left\{ \frac{dQ_{k+1}^0}{dQ_k}\bigg|_{\mathcal{F}_\tau} \geq \varepsilon \right\} \in \mathcal{F}_\tau$$

satisfies $Q_0[\,D\,] \geq 1-2^{-(k+1)}\,\delta$. Thus, $\Lambda_\delta^{k+1} := \Lambda_\delta^k \cap D$ satisfies $Q_0[\,\Lambda_\delta^{k+1}\,] \geq 1-(1-2^{-(k+1)})\delta$. We now define a set

$$B := \left\{ \widetilde{U}_\tau^{Q_{k+1}^0} + A_\tau^{Q_{k+1}^0} > \widetilde{U}_\tau^{Q_k} + A_\tau^{Q_k} \right\} \cap D,$$

and consider the pasting Q_{k+1} of Q_k and Q_{k+1}^0 in the stopping time $\sigma := \tau \mathrm{I}_B + T\mathrm{I}_{B^c}$. By Lemma 9.26, $Q_{k+1} \in \mathcal{Q}_{\mathcal{S}}$ and

$$\begin{aligned}
\widetilde{U}_\tau^{Q_{k+1}} + A_\tau^{Q_{k+1}} &= \big(\widetilde{U}_\tau^{Q_k} + A_\tau^{Q_k}\big)\cdot \mathrm{I}_{B^c} + \big(\widetilde{U}_\tau^{Q_{k+1}^0} + A_\tau^{Q_{k+1}^0}\big)\cdot \mathrm{I}_B && P\text{-a.s.}\\
&= \big(\widetilde{U}_\tau^{Q_k} + A_\tau^{Q_k}\big) \vee \big(\widetilde{U}_\tau^{Q_{k+1}^0} + A_\tau^{Q_{k+1}^0}\big) && P\text{-a.s. on } D\\
&= \max_{n\leq k+1} \big(\widetilde{U}_\tau^{Q_n^0} + A_\tau^{Q_n^0}\big) && P\text{-a.s. on } \Lambda_\delta^{k+1}.
\end{aligned}$$

□

Now we can proceed to proving the main results in this section.

Proof of Proposition 9.24. For $Q_0 \in \mathcal{Q}_{\mathcal{S}}$ and $t \in \{t, \dots, T\}$, $\mathcal{Q}_t(Q_0)$ denotes the set of all $Q \in \mathcal{Q}_{\mathcal{S}}$ which coincide with Q_0 on $\mathcal{F}_t$. By Lemma 9.27 and by the definition of $\widetilde{U}^Q$ as the Snell envelope of $H - A^Q$,

$$\begin{aligned}\widetilde{U}_t^{\uparrow} - A_t^{Q_0} &= \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_t(Q_0)} \widetilde{U}_t^Q \\ &= \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_t(Q_0)} \big((H_t - A_t^Q) \vee E_Q[\,\widetilde{U}_{t+1}^Q \mid \mathcal{F}_t\,]\big) \\ &= (H_t - A_t^{Q_0}) \vee \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_t(Q_0)} E_Q[\,\widetilde{U}_{t+1}^Q \mid \mathcal{F}_t\,].\end{aligned}$$

Since $\widetilde{U}_{t+1}^Q \le \widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^Q$, we get

$$\widetilde{U}_t^{\uparrow} - A_t^{Q_0} \le (H_t - A_t^{Q_0}) \vee \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_t(Q_0)} E_Q[\,\widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^Q \mid \mathcal{F}_t\,]. \tag{9.25}$$

For the proof of the converse inequality, let us fix an arbitrary $Q \in \mathcal{Q}_t(Q_0)$. For any $\delta > 0$, Lemma 9.27 yields a set $\Lambda_\delta \in \mathcal{F}_{t+1}$ with measure $Q[\,\Lambda_\delta\,] \ge 1 - \delta$ and $Q_k \in \mathcal{Q}_{t+1}(Q)$ such that $\widetilde{U}_{t+1}^{Q_k} \nearrow \widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^Q$ P-a.s. on Λ_δ. Since Q_k coincides with Q on $\mathcal{F}_{t+1}$, we have P-a.s. on Λ_δ

$$\begin{aligned}E_Q[\,\widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^Q \mid \mathcal{F}_t\,] = \lim_{k \uparrow \infty} E_Q[\,\widetilde{U}_{t+1}^{Q_k} \mid \mathcal{F}_t\,] &= \lim_{k \uparrow \infty} E_{Q_k}[\,\widetilde{U}_{t+1}^{Q_k} \mid \mathcal{F}_t\,] \\ &\le \limsup_{k \uparrow \infty} \widetilde{U}_t^{Q_k} \le \operatorname*{ess\,sup}_{\widetilde{Q} \in \mathcal{Q}_{t+1}(Q)} \widetilde{U}_t^{\widetilde{Q}} \\ &\le \operatorname*{ess\,sup}_{\widetilde{Q} \in \mathcal{Q}_t(Q)} \widetilde{U}_t^{\widetilde{Q}} = \widetilde{U}_t^{\uparrow} - A_t^Q \\ &= \widetilde{U}_t^{\uparrow} - A_t^{Q_0}\end{aligned}$$

By taking $\delta \downarrow 0$ and by recalling $\widetilde{U}_t^{\uparrow} \ge H_t$, we arrive at the converse of the inequality (9.25). □

Proof of Theorem 9.22. Since $Q_0 \in \mathcal{Q}_{\mathcal{S}}$ is obviously contained in $\mathcal{Q}_t(Q_0)$, the recursion formula of Proposition 9.24 yields

$$\widetilde{U}_t^{\uparrow} - A_t^{Q_0} \ge (H_t - A_t^{Q_0}) \vee E_{Q_0}[\,\widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^{Q_0} \mid \mathcal{F}_t\,] \ge E_{Q_0}[\,\widetilde{U}_{t+1}^{\uparrow} - A_{t+1}^{Q_0} \mid \mathcal{F}_t\,],$$

i.e., $\widetilde{U}_t^{\uparrow} - A^{Q_0}$ is indeed a Q_0-supermartingale for each $Q_0 \in \mathcal{Q}_{\mathcal{S}}$. We also know that $\widetilde{U}^{\uparrow}$ dominates H.

Let U be any process which dominates H and for which $U - A^Q$ is a Q-supermartingale for each $Q \in \mathcal{Q}_{\mathcal{S}}$. For fixed Q, the Q-supermartingale $U - A^Q$

dominates $H - A^Q$ and hence also $\widetilde{U}^Q$, since $\widetilde{U}^Q$ is the smallest Q-supermartingale dominating $H - A^Q$ by Proposition 6.11. It follows that

$$U_t \geq \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(\widetilde{U}_t^Q + A_t^Q\big) = \widetilde{U}_t^{\uparrow} \quad P\text{-a.s. for all } t. \qquad \square$$

Proposition 9.23 we will be implied by taking $\tau^* \equiv T$ in the following lemma.

Lemma 9.28. *Let H be a discounted American claim whose payoff is zero if it is not exercised at a given stopping time $\tau^* \in \mathcal{T}$, i.e., $H_t(\omega) = 0$ if $t \neq \tau^*(\omega)$. Then its upper $\mathcal{Q}_{\mathcal{S}}$-Snell envelope is given by*

$$\widetilde{U}_t^{\uparrow} = \mathrm{I}_{\{\tau^* \geq t\}} \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\, H_{\tau^*} - A_{\tau^*}^Q \mid \mathcal{F}_t\,] + A_t^Q\big), \quad t = 0, \dots, T.$$

Proof. By definition,

$$\widetilde{U}_t^{\uparrow} = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\, H_\tau - A_\tau^Q \mid \mathcal{F}_t\,] + A_t^Q\big).$$

Since each process A^Q is increasing, it is clearly optimal to take $\tau = t$ on $\{\tau^* \leq t\}$. Hence,

$$\widetilde{U}_t^{\uparrow} = \begin{cases} 0 & \text{on } \{\tau^* < t\}, \\ H_t & \text{on } \{\tau^* = t\}. \end{cases}$$

So we have to show that choosing $\tau \equiv \tau^*$ is optimal on $\{\tau^* > t\}$. If $\sigma \in \mathcal{T}_t$ is a stopping time with $P[\,\sigma > \tau^*\,] > 0$, then $\tau := \sigma \wedge \tau^*$ is as least as good as σ, since each process A^Q is increasing. So it remains to exclude the case that there exists a stopping time $\sigma \in \mathcal{T}_t$ with $\sigma \leq \tau^*$ on $\{\tau^* > t\}$ and $P[\,\sigma < \tau^*\,] > 0$, such that σ yields a strictly better result than τ^*. In this case, there exists some $Q_1 \in \mathcal{Q}_{\mathcal{S}}$ such that

$$E_{Q_1}[\, H_\sigma - A_\sigma^{Q_1} \mid \mathcal{F}_t\,] + A_t^{Q_1} > \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\, H_{\tau^*} - A_{\tau^*}^Q \mid \mathcal{F}_t\,] + A_t^Q\big) \tag{9.26}$$

with strictly positive probability on $\{\tau^* > t\}$. Take any $\widetilde{P} \in \mathcal{P}_{\mathcal{S}}$ and $\varepsilon > 0$, and define

$$B_\varepsilon := \left\{ \frac{d\widetilde{P}}{dQ_1}\bigg|_{\mathcal{F}_\sigma} \geq \varepsilon \right\}.$$

Now let Q^ε be the pasting Q_1 and $\widetilde{P}$ in the stopping time $\sigma \mathrm{I}_{B_\varepsilon} + T\mathrm{I}_{B_\varepsilon^c}$. According to Lemmas 9.25 and 9.26, $Q^\varepsilon \in \mathcal{Q}_{\mathcal{S}}$, and its upper variation process satisfies $A_t^{Q^\varepsilon} = A_t^{Q_1}$ and $A_\sigma^{Q^\varepsilon} = A_\sigma^{Q_1}$ as well as

$$A_{\tau^*}^{Q^\varepsilon} = A_{\tau^*}^{Q_1} \mathrm{I}_{B_\varepsilon^c} + A_\sigma^{Q_1} \mathrm{I}_{B_\varepsilon} \quad P\text{-a.s. on } \{\tau^* > t\}.$$

By using our assumption that $H_\sigma \le H_{\tau^*}$, we get

$$E_{Q_1}[\, H_\sigma - A_\sigma^{Q_1} \mid \mathcal{F}_t\,] + A_t^{Q_1} \le E_{Q^\varepsilon}[\, H_{\tau^*} - A_{\tau^*}^{Q^\varepsilon} \mid \mathcal{F}_t\,] + A_t^{Q^\varepsilon} \quad P\text{-a.s. on } B_\varepsilon.$$

By letting $\varepsilon \downarrow 0$, the P-measure of B_ε becomes arbitrarily close to 1, and we arrive at a contradiction to (9.26). □

9.4 Superhedging and risk measures

Let H be a discounted American claim such that

$$\widetilde{U}_0^\uparrow = \sup_{Q\in\mathcal{Q}_\mathcal{S}} \widetilde{U}_0^Q = \sup_{Q\in\mathcal{Q}_\mathcal{S}} \sup_{\tau\in\mathcal{T}} E_Q[\, H_\tau - A_\tau^Q\,] < \infty.$$

Our aim in this section is to construct superhedging strategies for H which belong to our set $\overline{\mathcal{S}}$ of admissible trading strategies. Recall that a superhedging strategy for H is any self-financing trading strategy whose value process dominates H. If applied with $t = 0$, the following theorem shows that $\widetilde{U}_0^\uparrow$ is the minimal amount for which a superhedging strategy is available.

Denote by $\widetilde{\mathcal{U}}_t^\uparrow(H)$ the set of all $\mathcal{F}_t$-measurable random variables $U_t \ge 0$ for which there exists some $\eta \in \mathcal{S}$ such that

$$U_t + \sum_{k=t+1}^{u} \eta_k \cdot (X_k - X_{k-1}) \ge H_u \quad \text{for all } u \ge t \;\; P\text{-a.s.} \tag{9.27}$$

Theorem 9.29. *The upper $\mathcal{Q}_\mathcal{S}$-Snell envelope $\widetilde{U}_t^\uparrow$ of H is the minimal element of $\widetilde{\mathcal{U}}_t^\uparrow(H)$. More precisely,*

(a) $\widetilde{U}_t^\uparrow \in \widetilde{\mathcal{U}}_t^\uparrow(H)$,

(b) $\widetilde{U}_t^\uparrow = \operatorname{ess\,inf}\ \widetilde{\mathcal{U}}_t^\uparrow(H)$.

Proof. The uniform Doob decomposition in Section 9.2 combined with Theorem 9.22 yields an increasing adapted process B and some $\xi \in \mathcal{S}$ such that

$$\widetilde{U}_u^\uparrow = \widetilde{U}_t^\uparrow + \sum_{k=t+1}^{u} \xi_k \cdot (X_k - X_{k-1}) + B_t - B_u \quad P\text{-a.s. for } u \ge t.$$

So the fact that $\widetilde{U}^\uparrow$ dominates H proves (a).

As to part (b), we first get $\widetilde{U}_t^\uparrow \ge \operatorname{ess\,inf}\ \widetilde{\mathcal{U}}_t^\uparrow(H)$ from (a). For the proof of the converse inequality, take $U_t \in \widetilde{\mathcal{U}}_t^\uparrow(H)$ and choose a predictable process $\eta \in \mathcal{S}$ for which (9.27) holds. We must show that the set $B := \{ \widetilde{U}_t^{\uparrow} \le U_t \}$ satisfies $P[\,B\,] = 1$. Let

$$\hat{U}_t := \widetilde{U}_t^{\uparrow} \wedge U_t = \widetilde{U}_t^{\uparrow} \cdot I_B + U_t \cdot I_{B^c}.$$

Then $\hat{U}_t \le \widetilde{U}_t^{\uparrow}$, and our claim will follow if we can show that $U_t^{\uparrow} \le \hat{U}_t$. Let ξ denote the predictable process obtained from the uniform Doob decomposition of the $\mathcal{P}$-supermartingale $\widetilde{U}^{\uparrow}$, and define

$$\hat{\xi}_s := \begin{cases} \xi_s & \text{if } s \le t, \\ \xi_s \cdot I_B + \eta_s \cdot I_{B^c} & \text{if } s > t. \end{cases}$$

With this choice, $\hat{\xi} \in \mathcal{S}$ by predictable convexity, and $\hat{U}_t$ satisfies (9.27), i.e., $\hat{U}_t \in \widetilde{\mathcal{U}}_t^{\uparrow}(H)$. Let

$$\hat{V}_s := \widetilde{U}_0^{\uparrow} + \sum_{k=1}^{s} \hat{\xi}_k \cdot (X_k - X_{k-1}).$$

Then $\hat{V}_s \ge H_s \ge 0$ for all s, and so $\hat{V} - A^Q$ is a Q-supermartingale for each $Q \in \mathcal{Q}_{\mathcal{S}}$ by Propositions 9.16 and 9.6. Hence,

$$\begin{aligned} \widetilde{U}_t^{\uparrow} &= \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_t} E_Q[\,H_\tau - A_\tau^Q + A_t^Q \mid \mathcal{F}_t\,] \\ &\le \operatorname*{ess\,sup}_{Q,\,\tau} E_Q\Big[\,\hat{U}_t + \sum_{k=t+1}^{\tau} \hat{\xi}_k \cdot (X_k - X_{k-1}) - A_\tau^Q + A_t^Q \;\Big|\; \mathcal{F}_t\,\Big] \\ &\le \hat{U}_t. \end{aligned}$$

This proves $\widetilde{U}_t^{\uparrow} \le \operatorname{ess\,inf} \widetilde{\mathcal{U}}_t^{\uparrow}(H)$. □

For European claims, the upper $\mathcal{Q}_{\mathcal{S}}$-Snell envelope takes the form

$$\widetilde{U}_t^{\uparrow} = \operatorname*{ess\,sup}_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\,H^E - A_T^Q \mid \mathcal{F}_t\,] + A_t^Q\big), \qquad t = 0, \dots, T.$$

By taking $t = 0$, it follows that

$$\widetilde{U}_0^{\uparrow} = \sup_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\,H^E\,] - E_Q[\,A_T^Q\,]\big) \tag{9.28}$$

is the smallest initial investment which suffices for superhedging the claim H^E. In fact, the formula above can be regarded as a special case of the representation theorem for *convex measures of risk* in our financial market model. This will be explained next. Let us take $L^\infty := L^\infty(\Omega, \mathcal{F}, P)$ as the space of all financial positions. A position $Y \in L^\infty$ will be regarded as *acceptable* if it can be hedged with a strategy in $\overline{\mathcal{S}}$ at no additional cost. Thus, we introduce the *acceptance set*

$$\mathcal{A}^{\mathcal{S}} := \Big\{\, Y \in L^\infty \;\Big|\; \exists\, \xi \in \mathcal{S} :\; Y + \sum_{t=1}^{T} \xi_t \cdot (X_t - X_{t-1}) \ge 0 \ P\text{-a.s.} \Big\}.$$

Due to the convexity of $\mathcal{S}$, this set $\mathcal{A}^{\mathcal{S}}$ is convex, and under the mild condition

$$\inf\{ m \in \mathbb{R} \mid m \in \mathcal{A}^{\mathcal{S}} \} > -\infty, \tag{9.29}$$

$\mathcal{A}^{\mathcal{S}}$ induces a convex measure of risk $\rho^{\mathcal{S}} := \rho_{\mathcal{A}^{\mathcal{S}}}$ via

$$\rho^{\mathcal{S}}(Y) := \inf\{ m \in \mathbb{R} \mid m + Y \in \mathcal{A}^{\mathcal{S}} \};$$

see Section 4.1. Note that condition (9.29) holds in particular if $\mathcal{S}$ does not contain arbitrage opportunities. In this case, we have in fact

$$\rho^{\mathcal{S}}(0) = \inf\{ m \in \mathbb{R} \mid m \in \mathcal{A}^{\mathcal{S}} \} = 0,$$

i.e., $\rho^{\mathcal{S}}$ is *normalized*. The main results of this chapter can be restated in terms of $\rho^{\mathcal{S}}$:

Corollary 9.30. *The following conditions are equivalent:*

(a) $\rho^{\mathcal{S}}$ *is sensitive.*

(b) $\overline{\mathcal{S}}$ *contains no arbitrage opportunities.*

(c) $\mathcal{P}_{\mathcal{S}} \neq \emptyset$.

If these equivalent conditions hold, then

$$\rho^{\mathcal{S}}(Y) = \sup_{Q \in \mathcal{Q}_{\mathcal{S}}} \big(E_Q[\, -Y\,] - E_Q[\, A_T^Q \,]\big), \quad Y \in L^\infty. \tag{9.30}$$

In other words, $\rho^{\mathcal{S}}$ can be represented in terms of the penalty function

$$\alpha(Q) = \begin{cases} E_Q[\, A_T^Q \,] & \text{if } Q \in \mathcal{Q}_{\mathcal{S}}, \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. That (a) implies (b) is obvious. The equivalence between (b) and (c) was shown in Theorem 9.9. Since both sides of (9.30) are cash invariant, it suffices to prove (9.30) for $Y \leq 0$. But then the representation for $\rho^{\mathcal{S}}$ is just a special case of the superhedging duality (9.28). Finally, (9.30) and (c) imply that $\rho^{\mathcal{S}}(X) \geq \sup_{\widetilde{P} \in \mathcal{P}_{\mathcal{S}}} \widetilde{E}[\, -X\,]$, and the sensitivity of $\rho^{\mathcal{S}}$ follows. □

Chapter 10
Minimizing the hedging error

In this chapter, we present an alternative approach to the problem of hedging in an incomplete market model. Instead of controlling the downside risk, we simply aim at minimizing the quadratic hedging error. We begin with a local version of the minimization problem, which may be viewed as a sequential regression procedure. Its solution involves an orthogonal decomposition of a given contingent claim; this extends a classical decomposition theorem for martingales known as the Kunita–Watanabe decomposition. Often, the value process generated by a locally risk-minimizing strategy can be described as the martingale of conditional expectations of the given contingent claim for a special choice of an equivalent martingale measure. Such “minimal” martingale measures will be studied in Section 10.2. In Section 10.3, we investigate the connection between local risk minimization and the problem of variance-optimal hedging where one tries to minimize the global quadratic hedging error. The local and the global versions coincide if the underlying measure is itself a martingale measure.

10.1 Local quadratic risk

In this section, we no longer restrict our discussion to strategies which are self-financing. Instead, we admit the possibility that the value of a position is readjusted at the end of each period by an additional investment in the numéraire asset. This means that, in addition to the initial investment at time $t = 0$, we allow for a cash flow throughout the trading periods up to the final time T. In particular, it will now be possible to replicate any given European claim, simply by matching the difference between the payoff of the claim and the value generated by the preceding strategy with a final transfer at time T.

Definition 10.1. A *generalized trading strategy* is a pair of two stochastic process (ξ^0, ξ) such that $\xi^0 = (\xi_t^0)_{t=0,\dots,T}$ is adapted, and such that $\xi = (\xi_t)_{t=1,\dots,T}$ is a d-dimensional predictable process. The (discounted) *value process* V of (ξ^0, ξ) is defined as

$$V_0 := \xi_0^0 \quad \text{and} \quad V_t := \xi_t^0 + \xi_t \cdot X_t \quad \text{for } t \geq 1.$$

For such a generalized trading strategy (ξ^0, ξ), the gains and losses accumulated up to time t by investing into the risky assets are given by the sum

$$\sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}).$$

The value process V takes the form

$$V_0 = \xi_1^0 + \xi_1 \cdot X_0 \quad \text{and} \quad V_t = V_0 + \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}), \quad t = 1, \dots, T,$$

if and only if $\overline{\xi} = (\xi_t^0, \xi_t)_{t=1,\dots,T}$ is a *self-financing* trading strategy with initial investment $V_0 = \xi_0^0 = \xi_1^0 + \xi_1 \cdot X_0$. In this case, $(\xi_t^0)_{t=1,\dots,T}$ is a predictable process. In general, however, the difference

$$V_t - \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1})$$

is now non-trivial, and it can be interpreted as the cumulative cost up to time t. This motivates the following definition.

Definition 10.2. The *gains process* G of a generalized trading strategy (ξ^0, ξ) is given by

$$G_0 := 0 \quad \text{and} \quad G_t := \sum_{k=1}^{t} \xi_k \cdot (X_k - X_{k-1}), \quad t = 1, \dots, T.$$

The *cost process* C of (ξ^0, ξ) is defined by the difference

$$C_t := V_t - G_t, \quad t = 0, \dots, T,$$

of the value process V and the gains process G.

In this and in the following sections, we will measure the risk of a strategy in terms of quadratic criteria for the hedging error, based on the "objective" measure P. Our aim will be to minimize such criteria within the class of those generalized strategies (ξ^0, ξ) which replicate a given discounted European claim H in the sense that their value process satisfies

$$V_T = H \quad P\text{-a.s.}$$

The claim H will be fixed for the remainder of this section. As usual we assume that the σ-field $\mathcal{F}_0$ is trivial, i.e., $\mathcal{F}_0 = \{\emptyset, \Omega\}$. In contrast to the previous sections of Part II, however, our approach does *not* exclude a priori the existence of arbitrage opportunities, even though the interesting cases will be those in which there exist equivalent martingale measures. Since our approach is based on L^2-techniques, another set of hypotheses is needed:

Assumption 10.3. *Throughout this section, we assume that the discounted claim H and the discounted price process X of the risky assets are both square-integrable with respect to the objective measure P:*

(a) $H \in \mathcal{L}^2(\Omega, \mathcal{F}_T, P) =: \mathcal{L}^2(P)$.

(b) $X_t \in \mathcal{L}^2(\Omega, \mathcal{F}_t, P; \mathbb{R}^d)$ *for all* t.

In addition to these assumptions, the quadratic optimality criteria we have in mind require the following integrability conditions for strategies.

Definition 10.4. An $\mathcal{L}^2$*-admissible strategy* for H is a generalized trading strategy (ξ^0, ξ) whose value process V satisfies

$$V_T = H \ P\text{-a.s.} \quad \text{and} \quad V_t \in \mathcal{L}^2(P) \text{ for each } t,$$

and whose gains process G is such that

$$G_t \in \mathcal{L}^2(P) \quad \text{for each } t.$$

We can now introduce the local version of a quadratic criterion for the hedging error of an $\mathcal{L}^2$-admissible strategy.

Definition 10.5. The *local risk process* of an $\mathcal{L}^2$-admissible strategy (ξ^0, ξ) is the process

$$R_t^{\text{loc}}(\xi^0, \xi) := E[\,(C_{t+1} - C_t)^2 \mid \mathcal{F}_t\,], \qquad t = 0, \dots, T-1.$$

An $\mathcal{L}^2$-admissible strategy $(\hat{\xi}^0, \hat{\xi})$ is called a *locally risk-minimizing strategy* if, for all t,

$$R_t^{\text{loc}}(\hat{\xi}^0, \hat{\xi}) \leq R_t^{\text{loc}}(\xi^0, \xi) \quad P\text{-a.s.}$$

for each $\mathcal{L}^2$-admissible strategy (ξ^0, ξ) whose value process satisfies $V_{t+1} = \hat{\xi}^0_{t+1} + \hat{\xi}_{t+1} \cdot X_{t+1} = \hat{V}_{t+1}$.

Remark 10.6. The reason for fixing the value $V_{t+1} = \hat{V}_{t+1}$ in the preceding definition becomes clear when we try to construct a locally risk-minimizing strategy $(\hat{\xi}^0, \hat{\xi})$ backwards in time. At time T, we want to construct $\hat{\xi}^0_{T-1}$, $\hat{\xi}^0_T$, $\hat{\xi}_{T-1}$, $\hat{\xi}_T$ as a minimizer for the local risk $R^{\text{loc}}_{T-1}(\xi^0, \xi)$. Since the terminal value of every $\mathcal{L}^2$-admissible strategy must be equal to H, this minimization requires the side condition $\xi^0_T + \xi_T \cdot X_T = H = \hat{V}_T$. As we will see in the proof of Theorem 10.9 below, minimality of $R^{\text{loc}}_{T-1}(\hat{\xi}^0, \hat{\xi})$ completely determines $\hat{\xi}^0_T$ and $\hat{\xi}_T$ and $\hat{V}_{T-1}$, but one is still free to choose $\hat{\xi}^0_{T-1}$ and $\hat{\xi}_{T-1}$ among all ξ^0_{T-1}, ξ_{T-1} with $\xi^0_{T-1} + \xi_{T-1} \cdot X_{T-1} = \hat{V}_{T-1}$. In the next step, it is therefore natural to minimize $R^{\text{loc}}_{T-2}(\xi^0, \xi)$ under the condition that V_{T-1} is equal to the value $\hat{V}_{T-1}$ obtained from the preceding step. Moreover, the problem will now be of the same type as the previous one. ◇

Although locally risk-minimizing strategies are generally not self-financing, it will turn out that they are "self-financing on average" in the following sense:

Definition 10.7. An $\mathcal{L}^2$-admissible strategy is called *mean self-financing* if its cost process C is a P-martingale, i.e., if

$$E[\, C_{t+1} - C_t \mid \mathcal{F}_t\,] = 0 \quad P\text{-a.s. for all } t.$$

In order to formulate conditions for the existence of a locally risk-minimizing strategy, let us first introduce some notation. The *conditional covariance* of two random variables W and Z with respect to P is defined as

$$\operatorname{cov}(W,\ Z \mid \mathcal{F}_t) := E[\, W Z \mid \mathcal{F}_t\,] - E[\, W \mid \mathcal{F}_t\,]\, E[\, Z \mid \mathcal{F}_t\,]$$

provided that the conditional expectations and their difference make sense. Similarly, we define the *conditional variance* of W under P:

$$\begin{aligned}\operatorname{var}(\, W \mid \mathcal{F}_t\,) &:= E[\, W^2 \mid \mathcal{F}_t\,] - E[\, W \mid \mathcal{F}_t\,]^2 \\ &= \operatorname{cov}(W,\ W \mid \mathcal{F}_t\,).\end{aligned}$$

Definition 10.8. Two adapted processes U and Y are called *strongly orthogonal* with respect to P if the conditional covariances

$$\operatorname{cov}(\, U_{t+1} - U_t,\ Y_{t+1} - Y_t \mid \mathcal{F}_t\,), \quad t = 0, \dots, T-1,$$

are well-defined and vanish P-almost surely.

When we consider the strong orthogonality of two processes U and Y in the sequel, then usually one of them will be a P-martingale. In this case, their conditional covariance reduces to

$$\operatorname{cov}(\, U_{t+1} - U_t,\ Y_{t+1} - Y_t \mid \mathcal{F}_t\,) = E\big[\, (U_{t+1} - U_t)(Y_{t+1} - Y_t) \mid \mathcal{F}_t\,\big].$$

After these preparations, we are now ready to state our first result, namely the following characterization of locally risk-minimizing strategies.

Theorem 10.9. *An $\mathcal{L}^2$-admissible strategy is locally risk-minimizing if and only if it is mean self-financing and its cost process is strongly orthogonal to X.*

Proof. The local risk process of any $\mathcal{L}^2$-admissible strategy (ξ^0, ξ) can be expressed as a sum of two non-negative terms:

$$R_t^{\mathrm{loc}}(\xi^0, \xi) = \operatorname{var}(\, C_{t+1} - C_t \mid \mathcal{F}_t\,) + E[\, C_{t+1} - C_t \mid \mathcal{F}_t\,]^2.$$

Since the conditional variance does not change if we add $\mathcal{F}_t$-measurable random variables to its argument, the first term on the right-hand side takes the form

$$\operatorname{var}(\, C_{t+1} - C_t \mid \mathcal{F}_t\,) = \operatorname{var}\big(\, V_{t+1} - \xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t\,\big). \tag{10.1}$$

The second term satisfies

$$E[\, C_{t+1} - C_t \mid \mathcal{F}_t\,]^2 = \big(E[\, V_{t+1} \mid \mathcal{F}_t\,] - \xi_{t+1} \cdot E[\, X_{t+1} - X_t \mid \mathcal{F}_t\,] - V_t\big)^2. \tag{10.2}$$

In a second step, we fix t and V_{t+1}, and we consider ξ_{t+1} and V_t as parameters. Our purpose is to derive necessary conditions for the minimality of $R_t^{\mathrm{loc}}(\xi^0, \xi)$ with respect to variations of ξ_{t+1} and V_t. To this end, note first that it is possible to change the parameters ξ_t^0 and ξ_t in such a way that V_t takes any given value, that the modified strategy is still an $\mathcal{L}^2$-admissible strategy for H, and that the values of ξ_{t+1} and V_{t+1} remain unchanged. In particular, the value in (10.1) is not affected by such a modification, and so it is necessary for the optimality of $R_t^{\mathrm{loc}}(\xi^0, \xi)$ that V_t minimizes (10.2). This is the case if and only if

$$V_t = E[\, V_{t+1} \mid \mathcal{F}_t \,] - \xi_{t+1} \cdot E[\, X_{t+1} - X_t \mid \mathcal{F}_t \,]. \tag{10.3}$$

The value of (10.1) is independent of V_t and a quadratic form in terms of the $\mathcal{F}_t$-measurable random vector ξ_{t+1}. Thus, (10.1) is minimal if and only if ξ_{t+1} solves the linear equation

$$0 = \operatorname{cov}\big(V_{t+1} - \xi_{t+1} \cdot (\, X_{t+1} - X_t),\ X_{t+1} - X_t \mid \mathcal{F}_t \big). \tag{10.4}$$

Note that (10.3) is equivalent to

$$E[\, C_{t+1} - C_t \mid \mathcal{F}_t \,] = E\big[\, V_{t+1} - \xi_{t+1} \cdot (X_{t+1} - X_t) \mid \mathcal{F}_t \,\big] - V_t = 0.$$

Moreover, given (10.3), the condition (10.4) holds if and only if

$$E\big[\, (C_{t+1} - C_t)(X_{t+1} - X_t) \mid \mathcal{F}_t \,\big] = 0$$

where we have used the fact that the conditional covariance in (10.4) is not changed by subtracting the $\mathcal{F}_t$-measurable random variable V_t from the first argument. Backward induction on t concludes the proof. □

The previous proof provides a recipe for a recursive construction of a locally risk-minimizing strategy: If V_{t+1} is already given, minimize

$$E\big[\, (C_{t+1} - C_t)^2 \mid \mathcal{F}_t \,\big] = E\Big[\, \big(V_{t+1} - \big(V_t + \xi_{t+1} \cdot (X_{t+1} - X_t) \big) \big)^2 \mid \mathcal{F}_t \,\Big]$$

with respect to V_t and ξ_{t+1}. This is just a conditional version of the standard problem of determining the *linear regression* of V_{t+1} on the increment $X_{t+1} - X_t$. Let us now consider the case

$$d = 1,$$

where our market model contains just one risky asset. Then the following recursive scheme yields formally an explicit solution:

$$\begin{aligned} \hat{V}_T &:= H, \\ \hat{\xi}_{t+1} &:= \frac{\operatorname{cov}(\, \hat{V}_{t+1},\ X_{t+1} - X_t \mid \mathcal{F}_t \,)}{\sigma_{t+1}^2} \cdot \mathrm{I}_{\{\sigma_{t+1} \neq 0\}}, \\ \hat{V}_t &:= E[\, \hat{V}_{t+1} \mid \mathcal{F}_t \,] - \hat{\xi}_{t+1} \cdot E[\, X_{t+1} - X_t \mid \mathcal{F}_t \,]. \end{aligned} \tag{10.5}$$

Here σ_{t+1}^2 is a shorthand notation for the conditional variance

$$\sigma_{t+1}^2 := \operatorname{var}(\, X_{t+1} - X_t \mid \mathcal{F}_t \,).$$

Defining $\hat{\xi}_t^0 := \hat{V}_t - \hat{\xi}_t \cdot X_t$, we obtain a generalized trading strategy $(\hat{\xi}^0, \hat{\xi})$ whose terminal portfolio value $\hat{V}_T$ coincides with H. However, an extra condition is needed to conclude that this strategy is indeed $\mathcal{L}^2$-admissible.

Proposition 10.10. *Consider a market model with a single risky asset and assume that there exists a constant C such that*

$$\big(E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,]\big)^2 \le C \cdot \sigma_t^2 \quad P\text{-a.s. for all } t. \tag{10.6}$$

Then the recursion (10.5) *defines a locally risk-minimizing strategy $(\hat{\xi}^0, \hat{\xi})$. Moreover, any other locally risk-minimizing strategy coincides with $(\hat{\xi}^0, \hat{\xi})$ up to modifications of $\hat{\xi}_t$ on the set $\{\sigma_t^2 = 0\}$.*

Proof. We have to show that $(\hat{\xi}^0, \hat{\xi})$ is $\mathcal{L}^2$-admissible. To this end, observe that the recursion (10.5) and the condition (10.6) imply that

$$\begin{aligned}
&E\big[\, \big(\hat{\xi}_t \cdot (X_t - X_{t-1})\big)^2 \,\big] \\
&\quad = E\bigg[\, \frac{\operatorname{cov}(\, \hat{V}_t,\ X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,)^2}{\sigma_t^4} \cdot E[\, (X_t - X_{t-1})^2 \mid \mathcal{F}_{t-1} \,]\, \mathrm{I}_{\{\sigma_t^2 \neq 0\}} \,\bigg] \\
&\quad \le (1 + C) \cdot E\bigg[\, \frac{\operatorname{cov}(\, \hat{V}_t,\ X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,)^2}{\sigma_t^2} \,\bigg] \\
&\quad \le (1 + C) \cdot E\big[\, \operatorname{var}(\, \hat{V}_t \mid \mathcal{F}_{t-1}) \,\big].
\end{aligned}$$

The last expectation is finite if $\hat{V}_t$ is square-integrable. In this case, $\hat{\xi}_t \cdot (X_t - X_{t-1}) \in \mathcal{L}^2(P)$ and in turn $\hat{V}_{t-1} \in \mathcal{L}^2(P)$. Hence, $\mathcal{L}^2$-admissibility of $(\hat{\xi}^0, \hat{\xi})$ follows by backward induction. The claim that $(\hat{\xi}^0, \hat{\xi})$ is locally risk-minimizing as well as the uniqueness assertion follow immediately from the construction. □

Remark 10.11. The predictable process

$$\sum_{s=1}^{t} \frac{\big(E[\, X_s - X_{s-1} \mid \mathcal{F}_{s-1} \,]\big)^2}{\operatorname{var}(\, X_s - X_{s-1} \mid \mathcal{F}_{s-1} \,)}, \quad t = 1, \dots, T,$$

is called the *mean-variance trade-off process* of X, and condition (10.6) is known as the assumption of *bounded mean-variance trade-off*. Intuitively, it states that the forecast $E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1} \,]$ of the price increment $X_t - X_{t-1}$ is of the same order as the corresponding standard deviation σ_t. ◇

Remark 10.12. The assumption of bounded mean-variance trade-off is equivalent to the existence of some $\delta < 1$ such that

$$\big(E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]\big)^2 \le \delta \cdot E[\,(X_t - X_{t-1})^2 \mid \mathcal{F}_{t-1}\,] \qquad P\text{-a.s. for all } t. \quad (10.7)$$

Indeed, with $\alpha_t = E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]$, the assumption of bounded mean-variance trade-off is equivalent to

$$\alpha_t^2 \le C\big(E[\,(X_t - X_{t-1})^2 \mid \mathcal{F}_{t-1}\,] - \alpha_t^2\big),$$

which is seen to be equivalent to (10.7) by choosing $\delta = C/(1+C)$. ◇

Example 10.13. Let us consider a market model consisting of a single risky asset S^1 and a riskless bond

$$S_t^0 = (1+r)^t, \quad t = 0, \dots, T,$$

with constant return $r > -1$. We assume that $S_0^1 = 1$, and that the returns

$$R_t := \frac{S_t^1 - S_{t-1}^1}{S_{t-1}^1}, \quad t = 1, \dots, T,$$

of the risky asset are independent and identically distributed random variables in $\mathcal{L}^2(P)$. Under these assumptions, the discounted price process X, defined by

$$X_t = \prod_{s=1}^{t} \frac{1+R_s}{1+r}, \quad t = 0, \dots, T,$$

is square-integrable. Denoting by $\widetilde{m}$ the mean of R_t and by $\widetilde{\sigma}^2$ its variance, we get

$$E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,] = X_{t-1} \cdot \frac{\widetilde{m} - r}{1+r},$$

$$\operatorname{var}(\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,) = X_{t-1}^2 \cdot \frac{\widetilde{\sigma}^2}{(1+r)^2}.$$

Thus, the condition of bounded mean-variance trade-off holds without any further assumptions, and a locally risk-minimizing strategy exists. Moreover, P is a martingale measure if and only if $\widetilde{m} = r$. ◇

Let us return to our general market model with an arbitrary number of risky assets

$$X = (X^1, \dots, X^d).$$

The following result characterizes the existence of locally risk-minimizing strategies in terms of a decomposition of the claim H.

Corollary 10.14. *There exists a locally risk-minimizing strategy if and only if H admits a decomposition*

$$H = c + \sum_{t=1}^{T} \xi_t \cdot (X_t - X_{t-1}) + L_T \quad P\text{-a.s.}, \tag{10.8}$$

where c is a constant, ξ is a d-dimensional predictable process such that

$$\xi_t \cdot (X_t - X_{t-1}) \in \mathcal{L}^2(P) \quad \text{for all } t,$$

and where L is a square integrable P-martingale which is strongly orthogonal to X and satisfies $L_0 = 0$. In this case, the locally risk-minimizing strategy $(\hat{\xi}^0, \hat{\xi})$ is given by $\hat{\xi} = \xi$ and by the adapted process $\hat{\xi}^0$ defined via $\hat{\xi}^0_0 = c$ and

$$\hat{\xi}^0_t = c + \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}) + L_t - \xi_t \cdot X_t, \quad t = 1, \dots, T.$$

Moreover, the decomposition (10.8) *is unique in the sense that the constant c and the martingale L are uniquely determined.*

Proof. If $(\hat{\xi}^0, \hat{\xi})$ is a given locally risk-minimizing strategy with cost process $\hat{C}$, then $L_t := \hat{C}_t - \hat{C}_0$ is a square-integrable P-martingale which is strongly orthogonal to X by Theorem 10.9. Hence, we obtain a decomposition (10.8). Conversely, if such a decomposition exists, then the strategy $(\hat{\xi}^0, \hat{\xi})$ has the cost process $\hat{C} = c + L$, and Theorem 10.9 implies that $(\hat{\xi}^0, \hat{\xi})$ is locally risk-minimizing.

To show that L is uniquely determined, suppose that there exists another decomposition of H in terms of $\widetilde{c}$, $\widetilde{\xi}$, and $\widetilde{L}$. Then

$$N_t := c - \widetilde{c} + L_t - \widetilde{L}_t = \sum_{s=1}^{t} (\widetilde{\xi}_s - \xi_s) \cdot (X_s - X_{s-1})$$

is a square-integrable P-martingale which is strongly orthogonal to X and which can be represented as a "stochastic integral" with respect to X. Strong orthogonality means that

$$0 = E\big[\, (\widetilde{\xi}_t - \xi_t) \cdot (X_t - X_{t-1})(X_t - X_{t-1}) \mid \mathcal{F}_{t-1} \,\big].$$

Multiplying this identity with $\widetilde{\xi}_t - \xi_t$ gives

$$0 = E\big[\, \big((\widetilde{\xi}_t - \xi_t) \cdot (X_t - X_{t-1})\big)^2 \mid \mathcal{F}_{t-1} \,\big],$$

and so $N_t - N_{t-1} = (\widetilde{\xi}_t - \xi_t) \cdot (X_t - X_{t-1}) = 0$ P-almost surely. In view of $\widetilde{L}_0 = L_0 = 0$, we thus get $\widetilde{L} = L$ and in turn $\widetilde{c} = c$. □

A decomposition of the form (10.8) will be called the *orthogonal decomposition* of the contingent claim H with respect to the process X. If X is itself a P-martingale, then the orthogonal decomposition reduces to the *Kunita–Watanabe decomposition*, which we will explain next. To this end, we will need some preparation.

Lemma 10.15. *For two square-integrable martingales M and N, the following two conditions are equivalent:*

(a) *M and N are strongly orthogonal.*

(b) *The product MN is a martingale.*

Proof. The martingale property of M and N gives

$$E[\,(M_{t+1}-M_t)(N_{t+1}-N_t)\mid \mathcal{F}_t\,] = E[\,M_{t+1}N_{t+1}\mid \mathcal{F}_t\,] - M_tN_t,$$

and this expression vanishes if and only if MN is a martingale. □

Let $\mathcal{H}^2$ denote the space of all square-integrable P-martingales. Via the identity $M_t = E[\,M_T \mid \mathcal{F}_t\,]$, each $M \in \mathcal{H}^2$ can be identified with its terminal value $M_T \in \mathcal{L}^2(P)$. With the standard identification of random variables which coincide P-a.s., $\mathcal{H}^2$ becomes a Hilbert space isomorphic to $L^2(P)$, if endowed with the inner product

$$(M, N)_{\mathcal{H}^2} := E[\,M_TN_T\,], \quad M, N \in \mathcal{H}^2.$$

Recall from Definition 6.16 that, for a stopping time τ, the stopped process M^τ is defined as

$$M_t^\tau := M_{\tau\wedge t}, \quad t = 0, \dots, T.$$

Definition 10.16. A subspace $\mathcal{S}$ of $\mathcal{H}^2$ is called *stable* if $M^\tau \in \mathcal{S}$ for each $M \in \mathcal{S}$ and every stopping time τ.

Proposition 10.17. *For a stable subspace $\mathcal{S}$ of $\mathcal{H}^2$ and for $L \in \mathcal{H}^2$ with $L_0 = 0$, the following conditions are equivalent.*

(a) *L is orthogonal to $\mathcal{S}$, i.e.,*

$$(L, M)_{\mathcal{H}^2} = 0 \quad \textit{for all } M \in \mathcal{S}.$$

(b) *L is strongly orthogonal to $\mathcal{S}$, i.e., for each $M \in \mathcal{S}$*

$$E[\,(L_{t+1}-L_t)(M_{t+1}-M_t)\mid \mathcal{F}_t\,] = 0 \quad P\textit{-a.s. for all } t.$$

(c) *The product LM is a martingale for each $M \in \mathcal{S}$.*

Proof. The equivalence of (b) and (c) follows from Lemma 10.15. To prove (a)⇔(c), we will show that LM is a martingale for fixed $M \in \mathcal{S}$ if and only if $(L, M^\tau)_{\mathcal{H}^2} = 0$ for all stopping times $\tau \le T$. By the stopping theorem in the form of Proposition 6.38,

$$(L, M^\tau)_{\mathcal{H}^2} = E[\,L_T M_\tau\,] = E[L_\tau M_\tau].$$

Using the fact that $L_0M_0 = 0$ and applying the stopping theorem in the form of Theorem 6.17, we conclude that $(L, M^\tau)_{\mathcal{H}^2} = 0$ for all stopping times $\tau \le T$ if and only if LM is a martingale. □

After these preparations, we can now state the existence theorem for the discrete-time version of the *Kunita–Watanabe decomposition*.

Theorem 10.18. *If the process X is a square-integrable martingale under P, then every martingale $M \in \mathcal{H}^2$ is of the form*

$$M_t = M_0 + \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}) + L_t$$

where ξ is a d-dimensional predictable process such that $\xi_t \cdot (X_t - X_{t-1}) \in \mathcal{L}^2(P)$ for each t, and where L is a square-integrable P-martingale which is strongly orthogonal to X and satisfies $L_0 = 0$. Moreover, this decomposition is unique in the sense that L is uniquely determined.

Proof. Denote by $\mathcal{X}$ the set of all d-dimensional predictable processes ξ such that $\xi_t \cdot (X_t - X_{t-1}) \in \mathcal{L}^2(P)$ for each t, and denote by

$$G_t(\xi) := \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}), \quad t = 0, \dots, T,$$

the "stochastic integral" of $\xi \in \mathcal{X}$ with respect to X. Since for $\xi \in \mathcal{X}$ the process $G(\xi)$ is a square-integrable P-martingale, the set $\mathcal{G}$ of all those martingales can be regarded as a linear subspace of the Hilbert space $\mathcal{H}^2$. In fact, $\mathcal{G}$ is a closed subspace of $\mathcal{H}^2$. To prove this claim, note that the martingale property of $G(\xi)$ implies that

$$\big(G(\xi), G(\xi)\big)_{\mathcal{H}^2} = E\big[\,\big(G_T(\xi)\big)^2\,\big] = \sum_{t=1}^{T} E\big[\,\big(\xi_t \cdot (X_t - X_{t-1})\big)^2\,\big].$$

Thus, if $\xi^{(n)}$ is such that $G(\xi^{(n)})$ is a Cauchy sequence in $\mathcal{H}^2$, then $\xi_t^{(n)} \cdot (X_t - X_{t-1})$ is a Cauchy sequence in $L^2(P)$ for each t. Since P is a martingale measure, we may apply Lemma 1.68 to conclude that any limit point of $\xi_t^{(n)} \cdot (X_t - X_{t-1})$ is of the form $\xi_t \cdot (X_t - X_{t-1})$ for some $\xi_t \in L^0(\Omega, \mathcal{F}_{t-1}, P; \mathbb{R}^d)$. Hence, $\mathcal{G}$ is closed in $\mathcal{H}^2$. Moreover, $\mathcal{G}$ is stable. Indeed, if $\xi \in \mathcal{X}$ and τ is a stopping time, then $G_{t\wedge\tau}(\xi) = G_t(\widetilde{\xi})$ where

$$\widetilde{\xi}_s := \xi_s \cdot \mathrm{I}_{\{\tau \geq s\}}, \quad s = 1, \dots, T.$$

Furthermore, we have $\widetilde{\xi} \in \mathcal{X}$ since

$$E\big[\big(\widetilde{\xi}_t \cdot (X_t - X_{t-1})\big)^2\big] \leq E\big[\big(\xi_t \cdot (X_t - X_{t-1})\big)^2\big] < \infty.$$

Since $\mathcal{G}$ is closed, the orthogonal projection N of $M - M_0$ onto $\mathcal{G}$ is well-defined by standard Hilbert space techniques. The martingale N belongs to $\mathcal{G}$, and the difference

$L := M - M_0 - N$ is orthogonal to $\mathcal{G}$. By Proposition 10.17, L is strongly orthogonal to $\mathcal{G}$ and hence strongly orthogonal to X. Therefore, $M = M_0 + N + L$ is the desired decomposition of M. The uniqueness of L follows as in the proof of Corollary 10.14. □

Remark 10.19. In dimension $d = 1$, the assumption of bounded mean-variance trade-off (10.6) is clearly satisfied if X is a square-integrable P-martingale. Combining Proposition 10.10 with Corollary 10.14 then yields an alternative proof of Theorem 10.18. Moreover, the recursion (10.5) identifies the predictable process ξ appearing in the Kunita–Watanabe decomposition of a martingale M:

$$\xi_t = \frac{E[\,(M_t - M_{t-1})(X_t - X_{t-1}) \mid \mathcal{F}_{t-1}\,]}{E[\,(X_t - X_{t-1})^2 \mid \mathcal{F}_{t-1}\,]} \cdot \mathrm{I}_{\{E[\,(X_t - X_{t-1})^2 | \mathcal{F}_{t-1}\,] \neq 0\}}. \qquad \diamond$$

10.2 Minimal martingale measures

If P is itself a martingale measure, Theorem 10.18 combined with Corollary 10.14 yields immediately a solution to our original problem of constructing locally risk-minimizing strategies:

Corollary 10.20. *If P is a martingale measure, then there exists a locally risk-minimizing strategy. Moreover, this strategy is unique in the sense that its value process $\hat{V}$ is uniquely determined as*

$$\hat{V}_t = E[\,H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T, \tag{10.9}$$

and that its cost process is given by

$$\hat{C}_t = \hat{V}_0 + L_t, \quad t = 0, \dots, T,$$

where L is the strongly orthogonal P-martingale arising in the Kunita–Watanabe decomposition of $\hat{V}$.

The identity (10.9) allows for a time-consistent interpretation of $\hat{V}_t$ as an arbitrage-free price for H at time t. In the general case in which X is not a martingale under P, one may ask whether there exists an equivalent martingale measure $\hat{P}$ such that the value process $\hat{V}$ of a locally risk-minimizing strategy can be obtained in a similar manner as the martingale

$$\hat{E}[\,H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T. \tag{10.10}$$

Definition 10.21. An equivalent martingale measure $\hat{P} \in \mathcal{P}$ is called a *minimal martingale measure* if

$$E\left[\left(\frac{d\hat{P}}{dP}\right)^2\right] < \infty,$$

and if every P-martingale $M \in \mathcal{H}^2$ which is strongly orthogonal to X is also a $\hat{P}$-martingale.

The following result shows that a minimal martingale measure provides the desired representation (10.10) – *if* such a minimal martingale measure exists.

Theorem 10.22. *If $\hat{P}$ is a minimal martingale measure, and if $\hat{V}$ is the value process of a locally risk-minimizing strategy, then*

$$\hat{V}_t = \hat{E}[\, H \mid \mathcal{F}_t\,], \quad t = 0, \dots, T.$$

Proof. Denote by

$$H = c + \sum_{t=1}^{T} \xi_t \cdot (X_t - X_{t-1}) + L_T$$

an orthogonal decomposition of H as in Corollary 10.14. Then $\hat{V}$ is given by

$$\hat{V}_t = c + \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}) + L_t.$$

The process L is a $\hat{P}$-martingale, because it is a square-integrable P-martingale strongly orthogonal to X. Moreover, $\xi_s \cdot (X_s - X_{s-1}) \in \mathcal{L}^1(\hat{P})$, because both $\xi_s \cdot (X_s - X_{s-1})$ and $d\hat{P}/dP$ are square-integrable with respect to P. It follows that $\hat{V}$ is a $\hat{P}$-martingale. In view of $\hat{V}_T = H$, the assertion follows. □

Our next goal is to derive a characterization of a minimal martingale measure and to use it in order to obtain criteria for its existence. To this end, we have to analyze the effect of an equivalent change of measure on the structure of martingales. The results we will obtain in this direction are of independent interest, and their continuous-time analogues have a wide range of applications in stochastic analysis.

Lemma 10.23. *Let $\widetilde{P}$ be a probability measure equivalent to P. An adapted process $\widetilde{M}$ is a $\widetilde{P}$-martingale if and only if the process*

$$\widetilde{M}_t \cdot E\Big[\, \frac{d\widetilde{P}}{dP} \;\Big|\; \mathcal{F}_t \Big], \quad t = 0, \dots, T,$$

is a P-martingale.

Proof. Let us denote

$$Z_t := E\Big[\, \frac{d\widetilde{P}}{dP} \;\Big|\; \mathcal{F}_t \Big].$$

Observe that $\widetilde{M}_t \in \mathcal{L}^1(\widetilde{P})$ if and only if $\widetilde{M}_t Z_t \in \mathcal{L}^1(P)$. Moreover, the process Z is P-a.s. strictly positive by the equivalence of $\widetilde{P}$ and P. Hence, Proposition A.12 yields that

$$Z_t \cdot \widetilde{E}[\, \widetilde{M}_{t+1} \mid \mathcal{F}_t\,] = E[\, \widetilde{M}_{t+1} Z_{t+1} \mid \mathcal{F}_t\,],$$

and it follows that $\widetilde{E}[\,\widetilde{M}_{t+1} \mid \mathcal{F}_t\,] = \widetilde{M}_t$ if and only if $E[\,\widetilde{M}_{t+1}Z_{t+1} \mid \mathcal{F}_t\,] = \widetilde{M}_t Z_t$. □

The following representation (10.12) of the density process may be viewed as the discrete-time version of the Doléans–Dade stochastic exponential in continuous-time stochastic calculus.

Proposition 10.24. *If $\widetilde{P}$ is a probability measure equivalent to P, then there exists a P-martingale Λ such that*

$$\Lambda_0 = 1 \quad \textit{and} \quad \Lambda_{t+1} - \Lambda_t > -1 \quad P\textit{-a.s. for all } t, \tag{10.11}$$

and such that the martingale

$$Z_t := E\Big[\, \frac{d\widetilde{P}}{dP} \,\Big|\, \mathcal{F}_t \Big], \quad t = 0, \dots, T,$$

can be represented as

$$Z_t = \prod_{s=1}^{t} (1 + \Lambda_s - \Lambda_{s-1}), \quad t = 0, \dots, T. \tag{10.12}$$

Conversely, if Λ is a P-martingale with (10.11) *and such that* (10.12) *defines a P-martingale Z, then*

$$d\widetilde{P} := Z_T \, dP$$

defines a probability measure $\widetilde{P} \approx P$.

Proof. For $\widetilde{P} \approx P$ given, define Λ by $\Lambda_0 = 1$ and

$$\Lambda_{t+1} := \Lambda_t + \frac{Z_{t+1} - Z_t}{Z_t}, \quad t = 0, \dots, T-1.$$

Clearly, (10.12) holds with this choice of Λ. In particular, Λ satisfies (10.11), because the equivalence of P and $\widetilde{P}$ implies that Z_t is P-a.s. strictly positive for all t.

In the next step, we show by induction on t that $\Lambda_t \in \mathcal{L}^1(P)$. For $t = 0$ this holds by definition. Suppose that $\Lambda_t \in \mathcal{L}^1(P)$. Since Z is non-negative, the conditional expectation of Z_{t+1}/Z_t is well-defined and satisfies P-a.s.

$$E\Big[\, \frac{Z_{t+1}}{Z_t} \,\Big|\, \mathcal{F}_t \Big] = \frac{1}{Z_t} \cdot E[\, Z_{t+1} \mid \mathcal{F}_t \,] = 1.$$

It follows that $Z_{t+1}/Z_t \in \mathcal{L}^1(P)$ and in turn that

$$\Lambda_{t+1} = \Lambda_t - 1 + \frac{Z_{t+1}}{Z_t} \in \mathcal{L}^1(P).$$

Now it is easy to derive the martingale property of Λ: Since Z_t is strictly positive, we may divide both sides of the equation $E[\, Z_{t+1} \mid \mathcal{F}_t \,] = Z_t$ by Z_t, and we arrive at $E[\, \Lambda_{t+1} - \Lambda_t \mid \mathcal{F}_t \,] = 0$.

As for the second assertion, it is clear that $E[\, Z_t \,] = Z_0 = 1$ for all t, provided that Λ is a P-martingale such that (10.11) holds and such that (10.12) defines a strictly positive P-martingale Z. □

The following theorem shows how a martingale M is affected by an equivalent change of the underlying probability measure P. Typically, M will no longer be a martingale under the new measure $\widetilde{P}$, and so a non-trivial predictable process $(A_t)_{t=1,\dots,T}$ will appear in the *Doob decomposition*

$$M = \widetilde{M} + A$$

of M under $\widetilde{P}$. Alternatively, $-A$ may be viewed as the predictable process arising in the Doob decomposition of the $\widetilde{P}$-martingale $\widetilde{M}$ under the measure P. The following result, a discrete-time version of the *Girsanov formula*, describes A in terms of the martingale Λ arising in the representation (10.12) of the successive densities.

Theorem 10.25. *Let P and $\widetilde{P}$ be two equivalent probability measures, and let Λ denote the P-martingale arising in the representation* (10.12) *of the successive densities $Z_t := E[\, d\widetilde{P}/dP \mid \mathcal{F}_t \,]$. If $\widetilde{M}$ is a $\widetilde{P}$-martingale such that $\widetilde{M}_t \in \mathcal{L}^1(P)$ for all t, then*

$$M_t := \widetilde{M}_t + \sum_{s=1}^{t} E\big[\, (\Lambda_s - \Lambda_{s-1})(\widetilde{M}_s - \widetilde{M}_{s-1}) \mid \mathcal{F}_{s-1} \,\big]$$

is a P-martingale.

Proof. Note first that

$$(\Lambda_t - \Lambda_{t-1})(\widetilde{M}_t - \widetilde{M}_{t-1}) = \frac{1}{Z_{t-1}}\big(Z_t(\widetilde{M}_t - \widetilde{M}_{t-1})\big) - (\widetilde{M}_t - \widetilde{M}_{t-1})\,. \qquad (10.13)$$

According to Lemma 10.23, $Z_t(\widetilde{M}_t - \widetilde{M}_{t-1})$ is a martingale increment, and hence belongs to $\mathcal{L}^1(P)$. If we let

$$\tau_n := \inf\{\, t \mid Z_t < 1/n \,\} \wedge T, \quad n = 2, 3, \dots,$$

it follows that

$$(\Lambda_t - \Lambda_{t-1})(\widetilde{M}_t - \widetilde{M}_{t-1})\, \mathrm{I}_{\{\tau_n \ge t\}} \in \mathcal{L}^1(P)\,.$$

In particular, the conditional expectations appearing in the statement of the theorem are P-a.s. well-defined. Moreover, the identity (10.13) implies that P-a.s. on $\{\tau_n \ge t\}$

$$\begin{aligned} &E[\, \widetilde{M}_t - \widetilde{M}_{t-1} \mid \mathcal{F}_{t-1} \,] \\ &= \frac{1}{Z_{t-1}} E\big[\, Z_t(\widetilde{M}_t - \widetilde{M}_{t-1}) \mid \mathcal{F}_{t-1} \,\big] - E[\, (\Lambda_t - \Lambda_{t-1})(\widetilde{M}_t - \widetilde{M}_{t-1}) \mid \mathcal{F}_{t-1} \,] \\ &= -E\big[\, (\Lambda_t - \Lambda_{t-1})(\widetilde{M}_t - \widetilde{M}_{t-1}) \mid \mathcal{F}_{t-1} \,\big]. \end{aligned}$$

Thus, we have identified the Doob decomposition of $\widetilde{M}$ under P. □

The preceding theorem allows us to characterize those equivalent measures $P^* \approx P$ which are martingale measures. Let

$$X = Y + B \tag{10.14}$$

denote the Doob decomposition of X under P, where Y is a d-dimensional P-martingale, and $(B_t)_{t=1,\dots,T}$ is a d-dimensional predictable process.

Corollary 10.26. *Let $P^* \approx P$ be such that $E^*[\,|X_t|\,] < \infty$ for each t, and denote by Λ the P-martingale arising in the representation* (10.12) *of the successive densities $Z_t := E[\,dP^*/dP \mid \mathcal{F}_t\,]$. Then P^* is an equivalent martingale measure if and only if the predictable process B in the Doob decomposition* (10.14) *satisfies*

$$\begin{aligned} B_t &= -\sum_{s=1}^{t} E\big[\,(\Lambda_s - \Lambda_{s-1})(Y_s - Y_{s-1}) \mid \mathcal{F}_{s-1}\,\big] \\ &= -\sum_{s=1}^{t} E\big[\,(\Lambda_s - \Lambda_{s-1})(X_s - X_{s-1}) \mid \mathcal{F}_{s-1}\,\big] \end{aligned}$$

P-a.s. for $t = 1, \dots, T$.

Proof. If P^* is an equivalent martingale measure, then our formula for B is an immediate consequence of Theorem 10.25. For the proof of the converse direction, we denote by

$$X = Y^* + B^*$$

the Doob decomposition of X under P^*. Then Y^* is a P^*-martingale. Using Theorem 10.25, we see that $\widetilde{Y}^* := Y^* + \widetilde{B}^*$ is a P-martingale where

$$\widetilde{B}^*_t = \sum_{s=1}^{t} E\big[\,(\Lambda_s - \Lambda_{s-1})(Y^*_s - Y^*_{s-1}) \mid \mathcal{F}_s\,\big] = -B_t.$$

On the other hand, $Y = X - B = Y^* + (B^* - B)$ is a P-martingale. It follows that the Doob decomposition of Y^* under P is given by $Y^* = Y + (B - B^*)$. Hence,

$$Y + (B - B^*) = Y^* = \widetilde{Y}^* - \widetilde{B}^* = \widetilde{Y}^* + B.$$

The uniqueness of the Doob decomposition implies $B^* \equiv 0$, so X is a P^*-martingale. □

We can now return to our initial task of characterizing a minimal martingale measure.

Theorem 10.27. *Let $\hat{P} \in \mathcal{P}$ be an equivalent martingale measure whose density $d\hat{P}/dP$ is square-integrable. Then $\hat{P}$ is a minimal martingale measure if and only if the P-martingale Λ of* (10.12) *admits a representation as a "stochastic integral" with respect to the P-martingale Y arising in the Doob decomposition of X:*

$$\Lambda_t = 1 + \sum_{s=1}^{t} \lambda_s \cdot (Y_s - Y_{s-1}), \quad t = 0, \dots, T, \tag{10.15}$$

for some d-dimensional predictable process λ.

Proof. To prove sufficiency of (10.15), we have to show that if $M \in \mathcal{H}^2$ is strongly orthogonal to X, then M is a $\hat{P}$-martingale. By Lemma 10.23 this follows if we can show that MZ is a P-martingale where

$$Z_t := E\Big[\frac{d\hat{P}}{dP} \,\Big|\, \mathcal{F}_t \Big].$$

Clearly, $M_t Z_t \in \mathcal{L}^1(P)$ since M and Z are both square-integrable.

For the next step, we introduce the stopping times

$$\tau_n := \inf\big\{ t \geq 0 \mid |\lambda_{t+1}| > n \big\}.$$

By stopping the martingale Λ at τ_n, we obtain the P-martingale Λ^{τ_n}. Since X is square-integrable, an application of Jensen's inequality yields that $E[\,|Y_t|^2\,] < \infty$ for all t. In particular, $M\Lambda^{\tau_n}$ is integrable. Furthermore, Lemma 10.15 shows that the strong orthogonality of M and Y implies that MY is a d-dimensional P-martingale. Hence,

$$\begin{aligned} &E\big[\, M_{t+1}(\Lambda^{\tau_n}_{t+1} - \Lambda^{\tau_n}_t) \mid \mathcal{F}_t \,\big] \\ &\qquad = \mathrm{I}_{\{t+1 \leq \tau_n\}} \lambda_{t+1} \cdot \big(E[\, M_{t+1} Y_{t+1} \mid \mathcal{F}_t \,] - E[\, M_{t+1} \mid \mathcal{F}_t \,] Y_t \big) = 0. \end{aligned}$$

Noting that

$$Z^{\tau_n}_t = \prod_{s=1}^{t} (1 + \Lambda^{\tau_n}_s - \Lambda^{\tau_n}_{s-1}),$$

and that Z^{τ_n} is square-integrable, we conclude that

$$E[\, M_{t+1} Z^{\tau_n}_{t+1} \mid \mathcal{F}_t \,] = Z^{\tau_n}_t E\big[\, M_{t+1}(1 + \Lambda^{\tau_n}_{t+1} - \Lambda^{\tau_n}_t) \mid \mathcal{F}_t \,\big] = Z^{\tau_n}_t M_t.$$

Thus, $Z^{\tau_n} M$ is a P-martingale for each n. By Doob's stopping theorem, the process

$$(Z^{\tau_n} M)^{\tau_n} = (ZM)^{\tau_n}$$

is also a P-martingale. Since $\tau_n \nearrow T$ P-a.s. and

$$\big| M^{\tau_n}_{t+1} Z^{\tau_n}_{t+1} \big| \leq \sum_{s=0}^{T} \big| M_s Z_s \big| \in \mathcal{L}^1(P),$$

we may apply the dominated convergence theorem for conditional expectations to obtain the desired martingale property of MZ:

$$E[\,M_{t+1}Z_{t+1}\mid \mathcal{F}_t\,] = \lim_{n\uparrow\infty} E[\,M^{\tau_n}_{t+1}Z^{\tau_n}_{t+1}\mid \mathcal{F}_t\,] = \lim_{n\uparrow\infty} M^{\tau_n}_t Z^{\tau_n}_t = M_t Z_t.$$

Thus, $\hat{P}$ is a minimal martingale measure.

For the proof of the converse assertion of the theorem, denote by

$$Z_t = 1 + \sum_{s=1}^{t} \eta_s \cdot (Y_s - Y_{s-1}) + L_t$$

the Kunita–Watanabe decomposition of the density process Z with respect to the measure P and the square-integrable martingale Y, as explained in Theorem 10.18. The process L is a square-integrable P-martingale strongly orthogonal to Y, and hence to X. Thus, the assumption that $\hat{P}$ is a minimal martingale measure implies that L is also a $\hat{P}$-martingale. Applying Lemma 10.23, it follows that

$$L_t Z_t = L_t + L_t \sum_{s=1}^{t} \eta_s \cdot (Y_s - Y_{s-1}) + L_t^2$$

is a P-martingale. According to Lemma 10.15, the strong orthogonality of L and Y yields that

$$L_t \sum_{s=1}^{t} \eta_s \cdot (Y_s - Y_{s-1})$$

is a P-martingale; recall that $\eta_s \cdot (Y_s - Y_{s-1}) \in \mathcal{L}^2(P)$ for all s. But then (L_t^2) must also be a martingale. In particular, the expectation of L_t^2 is independent of t and so

$$E[\,L_t^2\,] = L_0^2 = 0$$

from which we get that L vanishes P-almost surely. Hence, Z is equal to the "stochastic integral" of η with respect to Y, and we conclude that

$$\Lambda_{t+1} - \Lambda_t = \frac{Z_{t+1} - Z_t}{Z_t} = \frac{1}{Z_t}\eta_{t+1} \cdot (Y_{t+1} - Y_t),$$

so that (10.15) holds with $\lambda_t := \eta_t / Z_{t-1}$. □

Corollary 10.28. *There exists at most one minimal martingale measure.*

Proof. Let $\hat{P}$ and $\hat{P}'$ be two minimal martingale measures, and denote the martingales in the representation (10.12) by Λ and Λ', respectively. On the one hand, it follows from Corollary 10.26 that the martingale $N := \Lambda - \Lambda'$ is strongly orthogonal to Y. On the other hand, Theorem 10.27 implies that N admits a representation as a "stochastic integral" with respect to the P-martingale Y:

$$N_t = \sum_{s=1}^{t} (\lambda_s - \lambda'_s) \cdot (Y_s - Y_{s-1}), \quad t = 0, \dots, T.$$

Let $\tau_n := \inf\{t \mid |\lambda_{t+1} - \lambda'_{t+1}| > n\}$, so that N^{τ_n} is in $\mathcal{L}^2(P)$. Then it follows as in the proof of Corollary 10.14 that N^{τ_n} vanishes P-almost surely. Hence the densities of $\hat{P}$ and $\hat{P}'$ coincide. □

Recall that, for $d = 1$, we denote by

$$\sigma_t^2 = \text{var}(\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,)$$

the conditional variance of the increments of X.

Corollary 10.29. *In dimension $d = 1$, the following two conditions are implied by the existence of a minimal martingale measure $\hat{P}$:*

(a) *The predictable process λ arising in the representation formula* (10.15) *is of the form*

$$\lambda_t = \frac{-E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]}{\sigma_t^2} \qquad P\text{-a.s. on } \{\, \sigma_t^2 \neq 0\,\}. \tag{10.16}$$

(b) *For each t, P-a.s. on $\{\sigma_t^2 \neq 0\}$,*

$$(X_t - X_{t-1}) \cdot E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,] < E[\,(X_t - X_{t-1})^2 \mid \mathcal{F}_{t-1}\,].$$

Proof. (a): Denote by $X = Y + B$ the Doob decomposition of X with respect to P. According to Corollary 10.26, the P-martingale Λ arising in the representation (10.12) of the density $d\hat{P}/dP$ must satisfy

$$B_t - B_{t-1} = -E\big[\,(\Lambda_t - \Lambda_{t-1})(Y_t - Y_{t-1}) \mid \mathcal{F}_{t-1}\,\big].$$

Using that $\sigma_t^2 = E[\,(Y_t - Y_{t-1})^2 \mid \mathcal{F}_{t-1}\,]$ and that $B_t - B_{t-1} = E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]$ yields our formula for λ_t.

(b): By Proposition 10.24, the P-martingale Λ must be such that

$$\Lambda_t - \Lambda_{t-1} = \lambda_t \cdot (Y_t - Y_{t-1}) > -1 \quad P\text{-a.s. for all } t.$$

Given (a), this condition is equivalent to (b). □

Note that condition (b) of Corollary 10.29 is rather restrictive as it imposes an almost-sure bound on the $\mathcal{F}_t$-measurable increment $X_t - X_{t-1}$ in terms of $\mathcal{F}_{t-1}$-measurable quantities.

Theorem 10.30. *Consider a market model with a single risky asset satisfying condition* (b) *of Corollary* 10.29 *and the assumption*

$$\big(E[\,X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]\big)^2 \le C \cdot \sigma_t^2, \quad P\text{-a.s. for all } t \ge 1,$$

of bounded mean-variance trade-off. Then there exists a unique minimal martingale measure $\hat{P}$ whose density $d\hat{P}/dP = Z_T$ is given via (10.16), (10.15), *and* (10.12).

Proof. Denote by $X = Y + B$ the Doob decomposition of X under P. For λ_t defined via (10.16), the assumption of bounded mean-variance trade-off yields that

$$E\big[\,\big(\lambda_t \cdot (Y_t - Y_{t-1})\big)^2 \mid \mathcal{F}_{t-1}\,\big] \le C \quad P\text{-a.s.} \tag{10.17}$$

Hence, Λ_t defined according to (10.15) is a square-integrable P-martingale. As observed in the second part of the proof of Corollary 10.29, its condition (b) holds if and only if $\Lambda_t - \Lambda_{t-1} > -1$ for all t, so that Z defined by

$$Z_t = \prod_{s=1}^{t}(1 + \Lambda_s - \Lambda_{s-1}) = \prod_{s=1}^{t}\big(1 + \lambda_s \cdot (Y_s - Y_{s-1})\big)$$

is P-a.s. strictly positive. Moreover, the bound (10.17) guarantees that Z is a square-integrable P-martingale. We may thus conclude from Proposition 10.24 that Z is the density process of a probability measure $\hat{P} \approx P$ with a square-integrable density $d\hat{P}/dP$. In particular, X_t is $\hat{P}$-integrable for all t. Our choice of λ implies that

$$\begin{aligned} E\big[\,(\Lambda_t - \Lambda_{t-1})(Y_t - Y_{t-1}) \mid \mathcal{F}_{t-1}\,\big] &= -E[\,X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,] \\ &= -(B_t - B_{t-1}), \end{aligned}$$

and so $\hat{P}$ is an equivalent martingale measure by Corollary 10.26. Finally, Theorem 10.27 states that $\hat{P}$ is a minimal martingale measure, while uniqueness was already established in Corollary 10.28. □

Example 10.31. Let us consider again the market model of Example 10.13 with independent and identically distributed returns $R_t \in \mathcal{L}^2(P)$. We have seen that the condition of bounded mean-variance trade-off is satisfied without further assumptions. Let $\widetilde{m} := E[\,R_1\,]$ and $\widetilde{\sigma}^2 := \operatorname{var}(R_1)$. A short calculation using the formulas for $E[\,X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,]$ and $\operatorname{var}(\,X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,)$ obtained in Example 10.13 shows that the crucial condition (b) of Corollary 10.29 is equivalent to

$$(\widetilde{m} - r)R_1 < \widetilde{\sigma}^2 + \widetilde{m}(\widetilde{m} - r) \quad P\text{-a.s.} \tag{10.18}$$

Hence, (10.18) is equivalent to the existence of the minimal martingale measure. For $\widetilde{m} > r$ the condition (10.18) is an upper bound on R_1, while we obtain a lower bound for $\widetilde{m} < r$. In the case $\widetilde{m} = r$, the measure P is itself the minimal martingale measure, and the condition (10.18) is void. If the distribution of R_1 is given, and if

$$a \le R_1 \le b \quad P\text{-a.s.}$$

for certain constants $a > -1$ and $b < \infty$, then (10.18) is satisfied for all r in a certain neighborhood of $\widetilde{m}$. ◇

Remark 10.32. The purpose of condition (b) of Corollary 10.29 is to ensure that the density Z defined via

$$Z_t = \prod_{s=1}^{t} \big(1 + \lambda_s \cdot (Y_s - Y_{s-1})\big)$$

is strictly positive. In cases where this condition is violated, Z may still be a square-integrable P-martingale and can be regarded as the density of a *signed* measure

$$d\hat{P} = Z_T \, dP,$$

which shares some properties with the minimal martingale measure of Definition 10.21; see, e.g., [186]. ◇

In the remainder of this section, we consider briefly another quadratic criterion for the risk of an $\mathcal{L}^2$-admissible strategy.

Definition 10.33. The *remaining conditional risk* of an $\mathcal{L}^2$-admissible strategy (ξ^0, ξ) with cost process C is given by the process

$$R_t^{\text{rem}}(\xi^0, \xi) := E[\,(C_T - C_t)^2 \mid \mathcal{F}_t\,], \quad t = 0, \dots, T.$$

We say that an $\mathcal{L}^2$-admissible strategy (ξ^0, ξ) *minimizes the remaining conditional risk* if

$$R_t^{\text{rem}}(\xi^0, \xi) \le R_t^{\text{rem}}(\eta^0, \eta) \qquad P\text{-a.s.}$$

for all t and for each $\mathcal{L}^2$-admissible strategy (η^0, η) which coincides with (ξ^0, ξ) up to time t.

The next result shows that minimizing the remaining conditional risk for a martingale measure is the same as minimizing the local risk. In this case, Corollary 10.20 yields formulas for the value process and the cost process of a minimizing strategy.

Proposition 10.34. *For $P \in \mathcal{P}$, an $\mathcal{L}^2$-admissible strategy minimizes the remaining conditional risk if and only if it is locally risk minimizing.*

Proof. Let $(\hat{\xi}^0, \hat{\xi})$ be a locally risk-minimizing strategy, which exists by Corollary 10.20, and write $\hat{V}$ and $\hat{C}$ for its value and cost processes. Take another $\mathcal{L}^2$-admissible strategy (η^0, η) whose value and cost processes are denoted by V and C. Since $\hat{V}_T = H = V_T$, the cost process C satisfies

$$\begin{aligned} C_T - C_t &= V_T - V_t - \sum_{k=t+1}^{T} \eta_k \cdot (X_k - X_{k-1}) \\ &= \hat{V}_t - V_t + \sum_{k=t+1}^{T} (\hat{\xi}_k - \eta_k) \cdot (X_k - X_{k-1}) + \hat{C}_T - \hat{C}_t. \end{aligned}$$

Since X and $\hat{C}$ are strongly orthogonal martingales, the remaining conditional risk of (η^0, η) satisfies

$$\begin{aligned} &R_t^{\text{rem}}(\eta^0, \eta) \\ &= (\hat{V}_t - V_t)^2 + E\Big[\sum_{k=t+1}^{T} (\hat{\xi}_k - \eta_k)^2 \, (X_k - X_{k-1})^2 \;\Big|\; \mathcal{F}_t \Big] + E[\, (\hat{C}_T - \hat{C}_t)^2 \mid \mathcal{F}_t \,], \end{aligned}$$

and this expression is minimal if and only if $V_t = \widetilde{V}_t$ and $\eta_k = \hat{\xi}_k$ for all $k \geq t+1$ P-almost surely. □

In general, however, an $\mathcal{L}^2$-admissible strategy minimizing the remaining conditional risk does not exist, as will be shown in the following Section 10.3.

10.3 Variance-optimal hedging

Let $H \in \mathcal{L}^2(P)$ be a square-integrable discounted claim. Throughout this section, we assume that the discounted price process X of the risky asset is square-integrable with respect to P:

$$E[\, |X_t|^2 \,] < \infty \quad \text{for all } t.$$

As in the previous section, there is no need to exclude the existence of arbitrage opportunities, even though the cases of interest will of course be arbitrage-free.

Informally, the problem of variance-minimal hedging is to minimize the quadratic hedging error defined as the squared $L^2(P)$-distance

$$\| H - V_T \|_2^2 = E[\, (H - V_T)^2 \,]$$

between H and the terminal value of the value process V of a self-financing trading strategy.

Remark 10.35. Mean-variance hedging is closely related to the discussion in the preceding sections, where we considered the problems of minimizing the local conditional risk or the remaining conditional risk within the class of $\mathcal{L}^2$-admissible strategies for H. To see this, let (ξ^0, ξ) be an $\mathcal{L}^2$-admissible strategy for H in the sense of Definition 10.4, and denote by V, G, and C the resulting value, gains, and cost processes. The quantity

$$R(\xi^0, \xi) := E\big[\,(C_T - C_0)^2\,\big]$$

may be called the "global quadratic risk" of (ξ^0, ξ). It coincides with the initial value of the process $R_0^{\text{rem}}(\xi^0, \xi)$ of the remaining conditional risk introduced in Definition 10.33. Note that

$$R(\xi^0, \xi) = E\big[\,(H - V_0 - G_T)^2\,\big]$$

is independent of the values of the numéraire component ξ^0 at the times $t = 1, \dots, T$. Thus, the global quadratic risk of the generalized trading strategy (ξ^0, ξ) coincides with the quadratic hedging error

$$E\big[\,(H - \widetilde{V}_T)^2\,\big]$$

where $\widetilde{V}$ is the value process of the *self-financing* trading strategy arising from the d-dimensional predictable process ξ and the initial investment $\widetilde{V}_0 = V_0 = \xi_0^0$. ◇

Let us rephrase the problem of mean-variance hedging in a form which can be interpreted both within the class of self-financing trading strategies and within the context of Section 10.1. For a d-dimensional predictable process ξ we denote by $G(\xi)$ the gains process

$$G_t(\xi) = \sum_{s=1}^{t} \xi_s \cdot (X_s - X_{s-1}), \quad t = 0, \dots, T,$$

associated with ξ. Let us introduce the class

$$\mathcal{S} := \big\{\, \xi \mid \xi \text{ is predictable and } G_t(\xi) \in \mathcal{L}^2(P) \text{ for all } t \,\big\}.$$

Definition 10.36. A pair (V_0^*, ξ^*) where $V_0^* \in \mathbb{R}$ and $\xi^* \in \mathcal{S}$ is called a *variance-optimal strategy* for the discounted claim H if

$$E\big[\,\big(H - V_0^* - G_T(\xi^*)\big)^2\,\big] \le E\big[\,\big(H - V_0 - G_T(\xi)\big)^2\,\big]$$

for all $V_0 \in \mathbb{R}$ and all $\xi \in \mathcal{S}$.

Our first result identifies a variance-optimal strategy in the case $P \in \mathcal{P}$.

Proposition 10.37. *Assume that $P \in \mathcal{P}$, and let $(\hat{\xi}^0, \hat{\xi})$ be a locally risk-minimizing $\mathcal{L}^2$-admissible strategy as constructed in Corollary* 10.20. *Then $(V_0^*, \xi^*) := (\hat{\xi}_0^0, \hat{\xi})$ is a variance-optimal strategy.*

Proof. Recall from Remark 10.35 that if (ξ^0, ξ) is an $\mathcal{L}^2$-admissible strategy for H with value process V, then the expression

$$E\big[\big(H - V_0 - G_T(\xi)\big)^2\big]$$

is equal to the initial value $R_0^{\text{rem}}(\xi^0, \xi)$ of the remaining conditional risk process of (ξ^0, ξ). But according to Proposition 10.34, R_0^{rem} is minimized by $(\hat{\xi}^0, \hat{\xi})$. □

The general case where X is not a P-martingale will be studied under the simplifying assumption that the market model contains only one risky asset. We will first derive a general existence result, and then determine an explicit solution in a special setting. The key idea for showing the existence of a variance-optimal strategy is to minimize the functional

$$\mathcal{S} \ni \xi \longmapsto E\big[\big(H - V_0 - G_T(\xi)\big)^2\big],$$

first for fixed V_0, and then to vary the parameter V_0. The first step will be accomplished by projecting $H - V_0$ onto the space of "stochastic integrals"

$$\mathcal{G}_T := \big\{\, G_T(\xi) \mid \xi \in \mathcal{S} \,\big\}.$$

Clearly, $\mathcal{G}_T$ is a linear subspace of $L^2(P)$. Thus, we can obtain the optimal $\xi = \xi(V_0)$ by using the orthogonal projection of $H - V_0$ on $\mathcal{G}_T$ as soon as we know that $\mathcal{G}_T$ is *closed* in $L^2(P)$. In order to formulate a criterion for the closedness on $\mathcal{G}_T$, we denote by

$$\sigma_t^2 := \operatorname{var}(\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,), \quad t = 1, \dots, T,$$

the conditional variance, and by

$$\alpha_t := E[\, X_t - X_{t-1} \mid \mathcal{F}_{t-1}\,], \quad t = 1, \dots, T,$$

the conditional mean of the increments of X.

Proposition 10.38. *Suppose that $d = 1$, and assume the condition of bounded mean-variance trade-off*

$$\alpha_t^2 \le C \cdot \sigma_t^2 \quad P\text{-a.s. for all } t. \tag{10.19}$$

Then $\mathcal{G}_T$ is a closed linear subspace of $L^2(P)$.

Proof. Let $X = M + A$ be the Doob decomposition of X into a P-martingale M and a process A such that $A_0 = 0$ and $(A_t)_{t=1,\dots,T}$ is predictable. Since

$$\sigma_T^2 = \operatorname{var}(\, X_T - X_{T-1} \mid \mathcal{F}_{T-1}\,) = E\big[\,(M_T - M_{T-1})^2 \mid \mathcal{F}_{T-1}\,\big],$$

we get for $\xi \in \mathcal{S}$

$$\begin{aligned} E\big[G_T(\xi)^2\big] &= E\big[\big(G_{T-1}(\xi) + \xi_T \cdot (X_T - X_{T-1})\big)^2\big] \\ &= E\big[\big(G_{T-1}(\xi) + \xi_T \cdot (A_T - A_{T-1})\big)^2\big] + E\big[\big(\xi_T \cdot (M_T - M_{T-1})\big)^2\big] \\ &\ge E\big[\xi_T^2 \cdot \sigma_T^2\big]. \end{aligned} \tag{10.20}$$

Suppose now that ξ^n is a sequence in $\mathcal{S}$ such that $G_T(\xi^n)$ converges in $L^2(P)$ to some Y. Applying the inequality (10.20) to $G_T(\xi^n) - G_T(\xi^m) = G_T(\xi^n - \xi^m)$, we find that $\phi^n := \xi_T^n \cdot \sigma_T$ is a Cauchy sequence in $L^2(P)$. Denote $\phi := \lim_n \phi^n$, and let

$$\xi_T := \mathrm{I}_{\{\sigma_T > 0\}} \frac{\phi}{\sigma_T}.$$

By using our assumption (10.19), we obtain

$$\begin{aligned} E\big[\big(\xi_T^n \cdot (X_T - X_{T-1}) - \xi_T \cdot (X_T - X_{T-1})\big)^2 \big] & \\ &= E\big[(\xi_T^n - \xi_T)^2 \cdot E[\,(X_T - X_{T-1})^2 \mid \mathcal{F}_{T-1}\,]\big] \\ &= E\big[(\xi_T^n - \xi_T)^2 \cdot (\sigma_T^2 + \alpha_T^2)\big] \\ &\le (1 + C)\, E[\,(\xi_T^n \sigma_T - \xi_T \sigma_T)^2\,] \\ &= (1 + C)\, E[\,(\phi^n - \phi)^2\,]. \end{aligned}$$

Since the latter term converges to 0, it follows that

$$G_{T-1}(\xi^n) = G_T(\xi^n) - \xi_T^n \cdot (X_T - X_{T-1}) \longrightarrow Y - \xi_T \cdot (X_T - X_{T-1})$$

in $L^2(P)$. A backward iteration of this argument yields a predictable process $\xi \in \mathcal{S}$ such that $Y = G_T(\xi)$. Hence, $\mathcal{G}_T$ is closed in $L^2(P)$. □

Theorem 10.39. *In dimension $d = 1$, the condition* (10.19) *of bounded mean-variance trade-off guarantees the existence of a variance-optimal strategy (V_0^*, ξ^*). Such a strategy is P-a.s. unique up to modifications of ξ_t^* on $\{\sigma_t = 0\}$.*

Proof. Let $p : L^2(P) \to \mathcal{G}_T$ denote the orthogonal projection onto the closed subspace $\mathcal{G}_T$ of the Hilbert space $L^2(P)$, i.e., $p : L^2(P) \to \mathcal{G}_T$ is a linear operator such that

$$E\big[\big(Y - p(Y)\big)^2 \big] = \min_{Z \in \mathcal{G}_T} E\big[(Y - Z)^2 \big] \tag{10.21}$$

for all $Y \in L^2(P)$.

For any $V_0 \in \mathbb{R}$ we choose some $\xi(V_0) \in \mathcal{S}$ such that $G_T\big(\xi(V_0)\big) = p(H - V_0)$. The identity (10.21) shows that $\xi(V_0)$ minimizes the functional

$$E\big[\big(H - V_0 - G_T(\xi)\big)^2 \big]$$

among all $\xi \in \mathcal{S}$. Note that

$$V_0 \mapsto G_T\big(\xi(V_0)\big) = p(H - V_0) = p(H) - V_0 \cdot p(1)$$

is an affine mapping. Hence,

$$E\big[\big(H - V_0 - G_T\big(\xi(V_0)\big)\big)^2 \big]$$

is a quadratic function of V_0 and there exists a minimizer V_0^*. For any $V_0 \in \mathbb{R}$ and $\xi \in \mathcal{S}$ we clearly have

$$E\big[\big(H - V_0 - G_T(\xi)\big)^2\big] \geq E\big[\big(H - V_0 - G_T\big(\xi(V_0)\big)\big)^2\big] \geq E\big[\big(H - V_0^* - G_T\big(\xi(V_0^*)\big)\big)^2\big].$$

Hence $(V_0^*, \xi^*) := (V_0^*, \xi(V_0^*))$ is a variance-optimal strategy. Uniqueness follows from (10.20) and an induction argument. □

Under the additional assumption that

$$\frac{\alpha_t^2}{\sigma_t^2} \text{ is deterministic for each } t \tag{10.22}$$

(here we use the convention $\frac{0}{0} := 0$), the variance-optimal strategy (ξ^*, V_0^*) can be determined explicitly. It turns out that (ξ^*, V_0^*) is closely related to the locally risk-minimizing strategy $(\hat{\xi}^0, \hat{\xi})$ for the discounted claim H. Recall from Proposition 10.10 that $(\hat{\xi}^0, \hat{\xi})$ and its value process $\hat{V}$ are determined by the following recursion:

$$\hat{V}_T = H,$$
$$\hat{\xi}_{t+1} = \frac{\operatorname{cov}(\hat{V}_{t+1}, X_{t+1} - X_t \mid \mathcal{F}_t)}{\sigma_{t+1}^2} \cdot \mathrm{I}_{\{\sigma_{t+1} \neq 0\}},$$
$$\hat{V}_t = E[\hat{V}_{t+1} \mid \mathcal{F}_t] - \hat{\xi}_{t+1} \cdot E[X_{t+1} - X_t \mid \mathcal{F}_t];$$

the numéraire component $\hat{\xi}^0$ is given by $\hat{\xi}_t^0 := \hat{V}_t - \hat{\xi}_t \cdot X_t$.

Theorem 10.40. *Under condition* (10.22)*,* $V_0^* := \hat{V}_0$ *and*

$$\xi_t^* := \hat{\xi}_t + \frac{\alpha_t}{\alpha_t^2 + \sigma_t^2}\big(\hat{V}_{t-1} - \hat{V}_0 - G_{t-1}(\xi^*)\big), \quad t = 1, \dots, T,$$

defines a variance-optimal strategy (ξ^*, V_0^*)*. Moreover,*

$$E\big[\big(H - V_0^* + G_T(\xi^*)\big)^2\big] = \sum_{t=1}^{T} \gamma_t \cdot E\big[(\hat{C}_t - \hat{C}_{t-1})^2\big],$$

where $\hat{C}$ *denotes the cost process of* $(\hat{\xi}^0, \hat{\xi})$*, and* γ_t *is given by*

$$\gamma_t := \prod_{k=t+1}^{T} \frac{\sigma_k^2}{\sigma_k^2 + \alpha_k^2}.$$

Proof. We prove the assertion by induction on T. For $T = 1$ the problem is just a particular case of Proposition 10.10, which yields $\xi_1^* = \hat{\xi}_1$ and $V_0^* = \hat{V}_0$.

For $T > 1$ we use the orthogonal decomposition

$$H = \hat{V}_T = \hat{V}_0 + \sum_{t=1}^{T} \hat{\xi}_t \cdot (X_t - X_{t-1}) + \hat{C}_T - \hat{C}_0, \tag{10.23}$$

of the discounted claim H as constructed in Corollary 10.14. Suppose that the assertion is proved for $T - 1$. Let us consider the minimization of

$$\xi_T \mapsto E\big[\big(\hat{V}_T - V_{T-1} - \xi_T \cdot (X_T - X_{T-1})\big)^2\big], \tag{10.24}$$

where V_{T-1} is any random variable in $\mathcal{L}^2(\Omega, \mathcal{F}_{T-1}, P)$. By (10.23) and Theorem 10.9, we may write $\hat{V}_T$ as

$$\hat{V}_T = \hat{V}_{T-1} + \hat{\xi}_T \cdot (X_T - X_{T-1}) + \hat{C}_T - \hat{C}_{T-1},$$

where $\hat{C}$ is a P-martingale strongly orthogonal to X. Thus,

$$\begin{aligned} &E\big[\big(\hat{V}_T - V_{T-1} - \xi_T \cdot (X_T - X_{T-1})\big)^2\big] \\ &\qquad = E\big[\big(\hat{V}_{T-1} - V_{T-1} + (\hat{\xi}_T - \xi_T) \cdot (X_T - X_{T-1}) + \hat{C}_T - \hat{C}_{T-1}\big)^2\big]. \end{aligned}$$

The expectation conditional on $\mathcal{F}_{T-1}$ of the integrand on the right-hand side is equal to

$$\begin{aligned} &(\hat{V}_{T-1} - V_{T-1})^2 + 2(\hat{V}_{T-1} - V_{T-1})(\hat{\xi}_T - \xi_T) \cdot \alpha_T \\ &\qquad + (\hat{\xi}_T - \xi_T)^2(\sigma_T^2 + \alpha_T^2) + E\big[(\hat{C}_T - \hat{C}_{T-1})^2 \mid \mathcal{F}_{T-1}\big]. \end{aligned} \tag{10.25}$$

This expression is minimized by

$$\xi_T(V_{T-1}) := \hat{\xi}_T + \frac{\alpha_T}{\alpha_T^2 + \sigma_T^2}\,(\hat{V}_{T-1} - V_{T-1}), \tag{10.26}$$

which must also be the minimizer in (10.24). The minimal value in (10.25) is given by

$$(\hat{V}_{T-1} - V_{T-1})^2 \frac{1}{1 + \alpha_T^2/\sigma_T^2} + E\big[(\hat{C}_T - \hat{C}_{T-1})^2 \mid \mathcal{F}_{T-1}\big].$$

Using our assumption (10.22) that α_T^2/σ_T^2 is constant, we can compute the expectation of the latter expression, and we arrive at the following identity:

$$\begin{aligned} &E\big[\big(\hat{V}_T - V_{T-1} - \xi_T(V_{T-1}) \cdot (X_T - X_{T-1})\big)^2\big] \\ &\qquad = E\big[(\hat{V}_{T-1} - V_{T-1})^2\big] \cdot \frac{\sigma_T^2}{\sigma_T^2 + \alpha_T^2} + E\big[(\hat{C}_T - \hat{C}_{T-1})^2\big]. \end{aligned} \tag{10.27}$$

So far, we have not specified V_{T-1}. Let us now consider the minimization of

$$E\Big[\big(\hat{V}_T - V_{T-1} - \xi_T(V_{T-1}) \cdot (X_T - X_{T-1})\big)^2\Big]$$

with respect to V_{T-1}, when V_{T-1} is of the form $V_{T-1} = V_0 + G_{T-1}(\xi)$ for $\xi \in \mathcal{S}$ and $V_0 \in \mathbb{R}$. According to our identity (10.27), this problem is equivalent to the minimization of

$$(V_0, \xi) \longmapsto E\Big[\big(H_{T-1} - V_0 + G_{T-1}(\xi)\big)^2\Big],$$

where $H_{T-1} := \hat{V}_{T-1}$. By the induction hypotheses, this problem is solved by V_0^* and $(\xi_t^*)_{t=1,\dots,T-1}$ as defined in the assertion. Inserting our formula (10.26) for $\xi_T\big(V_0^* + G_{T-1}(\xi^*)\big)$ completes the induction argument. □

Remark 10.41. The martingale property of $\hat{C}$ implies

$$E\big[\,(\hat{C}_T - \hat{C}_0)^2\,\big] = \sum_{t=1}^{T} E\big[\,(\hat{C}_t - \hat{C}_{t-1})^2\,\big].$$

If $\hat{C} \not\equiv \hat{C}_0$ and if $\gamma_1 < 1$, then $E[\,(\hat{C}_T - \hat{C}_0)^2\,]$ must be strictly larger than the minimal global risk

$$E\Big[\big(H - (V_0^* + G_T(\xi^*))\big)^2\Big] = \sum_{t=1}^{T} \gamma_t \cdot E\big[\,(\hat{C}_t - \hat{C}_{t-1})^2\,\big]. \qquad \diamond$$

Remark 10.42. If follows from Theorem 10.40 as well as from the preceding remark that the component $\hat{\xi}$ of a variance-optimal strategy $(\hat{\xi}^0, \hat{\xi})$ will differ from the corresponding ξ^* of a locally risk-minimizing strategy, if α_t does not vanish for all t, i.e., if P is not a martingale measure. This explains why there may be no strategy which minimizes the remaining conditional risk in the sense of Definition 10.33: For the minimality of

$$R_0^{\mathrm{rem}}(\xi^0, \xi) = E\big[\,(C_T - C_0)^2\,\big] = E\big[\,(H - \xi_0^0 - G_T(\xi))^2\,\big]$$

we need that $\xi = \xi^*$, while the minimality of

$$R_{T-1}^{\mathrm{rem}}(\xi^0, \xi) = E\big[\,(C_T - C_{T-1})^2 \mid \mathcal{F}_{T-1}\,\big] = R_{T-1}^{\mathrm{loc}}(\xi^0, \xi)$$

requires $\xi_T = \hat{\xi}_T$. Hence, the two minimality requirements are in general incompatible. $\diamond$

Appendix

A.1 Convexity

This section contains a few basic facts on convex functions and on convex sets in Euclidean space. Denote by

$$|x| := \sqrt{x \cdot x}$$

the Euclidean norm of $x \in \mathbb{R}^n$.

Proposition A.1. *Suppose that $\mathcal{C} \subset \mathbb{R}^n$ is a non-empty convex set with $0 \notin \mathcal{C}$. Then there exists $\eta \in \mathbb{R}^n$ with $\eta \cdot x \geq 0$ for all $x \in \mathcal{C}$, and with $\eta \cdot x_0 > 0$ for at least one $x_0 \in \mathcal{C}$. Moreover, if $\inf_{x \in \mathcal{C}} |x| > 0$, then one can find $\eta \in \mathbb{R}^n$ with $\inf_{x \in \mathcal{C}} \eta \cdot x > 0$.*

Proof. First we consider the case in which $\inf_{x \in \mathcal{C}} |x| > 0$. This infimum is attained by some y in the closure $\overline{\mathcal{C}}$ of $\mathcal{C}$. Since the set $\overline{\mathcal{C}}$ is also convex, $|y + \alpha(x-y)|^2 \geq |y|^2$ for each $x \in \mathcal{C}$ and all $\alpha \in [0, 1]$. Thus

$$2\alpha \, y \cdot (x - y) + \alpha^2 \, |x - y|^2 \geq 0.$$

For $\alpha \downarrow 0$ we obtain $y \cdot x \geq y \cdot y > 0$, and so we can take $\eta := y$.

Now let $\mathcal{C}$ be any non-empty convex subset of $\mathbb{R}^n$ such that $0 \notin \mathcal{C}$. In a first step, we will show that $\overline{\mathcal{C}}$ is a proper subset of $\mathbb{R}^n$. To this end, let $\{e_1, \dots, e_k\}$ be a maximal collection of linearly independent vectors in $\mathcal{C}$. Then each $x \in \mathcal{C}$ can be expressed as a linear combination of $e_1, \dots, e_k$. We claim that

$$z := -\sum_{i=1}^{k} e_i$$

is not contained in $\overline{\mathcal{C}}$. We assume by way of contradiction that $z \in \overline{\mathcal{C}}$. Then there are $z_n \in \mathcal{C}$ converging to z. If we write $z_n = \sum_{i=1}^{k} \lambda_n^i e_i$, then $z_n \to z$ is equivalent to the convergence $\lambda_n^i \to -1$ for all i. It follows that for some $n_0 \in \mathbb{N}$ all coefficients $\lambda_{n_0}^i$ are strictly negative. Let

$$\alpha_0 := \frac{1}{1 - \sum_{i=1}^{k} \lambda_{n_0}^i} \quad \text{and} \quad \alpha_j := \frac{-\lambda_{n_0}^j}{1 - \sum_{i=1}^{k} \lambda_{n_0}^i} \quad \text{for } j = 1, \dots, k.$$

Then the α_j's are non-negative and sum up to 1. Thus, the convexity of $\mathcal{C}$ implies that

$$0 = \alpha_0 z_{n_0} + \alpha_1 e_1 + \cdots + \alpha_k e_k \in \mathcal{C},$$

which is a contradiction. Hence, z is not contained in $\overline{\mathcal{C}}$.

Now we are in a position to prove the existence of a separating η in the case in which 0 is a boundary point of $\mathcal{C}$, and thus $\inf_{x\in\mathcal{C}} |x| = 0$. We may assume without loss of generality that the linear hull of $\mathcal{C}$ is the full space $\mathbb{R}^n$. Since we already know that $\mathcal{C}$ is not dense in $\mathbb{R}^n$, we may choose a sequence $(z_m) \subset \mathbb{R}^n$ such that $\inf_{x\in\mathcal{C}} |x - z_m| > 0$ and $z_m \to 0$. Then $\mathcal{C}_m := \mathcal{C} - z_m$ satisfies $\inf_{x\in\mathcal{C}_m} |x| > 0$, and the first part of the proof yields corresponding vectors η_m. We may assume that $|\eta_m| = 1$ for all m. By compactness of the $(n-1)$-dimensional unit sphere, there exists a convergent subsequence (η_{m_k}) with limit η, which satisfies

$$\eta \cdot x = \lim_{k\uparrow\infty} \eta_{m_k} \cdot x = \lim_{k\uparrow\infty} \eta_{m_k} \cdot (x - z_{m_k}) \geq 0$$

for all $x \in \mathcal{C}$. Since η is also a unit vector and $\mathcal{C}$ is not contained in a proper linear subspace of $\mathbb{R}^n$, the case $\eta \cdot x = 0$ for all $x \in \mathcal{C}$ cannot occur, and so there must be some $x_0 \in \mathcal{C}$ with $\eta \cdot x_0 > 0$. □

Definition A.2. Let A be any subset of a linear space E. The *convex hull* of A is defined as

$$\text{conv } A = \left\{ \sum_{i=1}^n \alpha_i x_i \;\middle|\; x_i \in A,\ \alpha_i \geq 0,\ \sum_{i=1}^n \alpha_i = 1,\ n \in \mathbb{N} \right\}.$$

It is straightforward to check that conv A is the smallest convex set containing A. Let us now turn to convex functions on $\mathbb{R}$.

Definition A.3. A function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is called a *proper convex function* if $f(x) < \infty$ for some $x \in \mathbb{R}$ and if

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha\, f(x) + (1-\alpha)\, f(y)$$

for $x, y \in \mathbb{R}$ and $\alpha \in [0, 1]$. The *effective domain* of f, denoted by dom f, consists of all $x \in \mathbb{R}$ such that $f(x) < \infty$.

Clearly, the effective domain of a proper convex function f is a real interval $S = \text{dom } f$. If considered as a function $f : S \to \mathbb{R}$, the function f is *convex* in the usual sense. Conversely, any convex function $f : S \to \mathbb{R}$ defined on some non-empty interval S may be viewed as a proper convex function defined by $f(x) := +\infty$ for $x \in \mathbb{R}\backslash S$. The following proposition summarizes continuity and differentiability properties of a proper convex function on its effective domain.

Proposition A.4. *Let f be a proper convex function, and denote by D the interior of* dom f.

(a) *f is upper semicontinuous on* dom f *and locally Lipschitz continuous on D.*

(b) *f admits left- and right-hand derivatives*

$$f'_-(y) := \lim_{x\uparrow y} \frac{f(x) - f(y)}{x - y} \quad \textit{and} \quad f'_+(y) := \lim_{z\downarrow y} \frac{f(z) - f(y)}{z - y}$$

at each $y \in D$. Both f'_+ and f'_- are increasing functions and satisfy $f'_- \leq f'_+$.

(c) *The right-hand derivative* f'_+ *is right-continuous, the left-hand derivative* f'_- *is left-continuous.*

(d) f *is differentiable a.e. in* D*, and for any* $x_0 \in D$

$$f(x) = f(x_0) + \int_{x_0}^{x} f'_+(y)\,dy = f(x_0) + \int_{x_0}^{x} f'_-(y)\,dy, \quad x \in D.$$

Proof. We first prove part (b). For $x, y, z \in D$ with $x < y < z$, we take $\alpha \in (0, 1)$ such that $y = \alpha z + (1 - \alpha)x$. Using the convexity of f, one gets

$$\frac{f(y) - f(x)}{y - x} \le \frac{f(z) - f(x)}{z - x} \le \frac{f(z) - f(y)}{z - y}. \tag{A.1}$$

Thus, the difference quotient

$$\frac{f(x) - f(y)}{x - y}$$

is an increasing function of x, which shows the existence of the left- and right-hand derivatives. Moreover, we get $f'_-(y) \le f'_+(y) \le f'_-(z)$ for $y < z$.

(a): Let $z \in \operatorname{dom} f$, and take a sequence $(x_n) \subset \operatorname{dom} f$ such that $x_n \to z$. Without loss of generality, we may assume that $x_n \downarrow z$ or $x_n \uparrow z$. In either case, $x_n = \delta_n x_1 + (1 - \delta_n)z$, where $\delta_n \downarrow 0$. Convexity of f yields

$$\limsup_{n\uparrow\infty} f(x_n) \le \limsup_{n\uparrow\infty} \big(\delta_n f(x_1) + (1 - \delta_n) f(z)\big) = f(z),$$

and so f is upper semicontinuous. To prove local Lipschitz continuity, take $a \le x < y \le b$ such that $[a, b] \subset D$. We get from part (b) that

$$f'_+(a) \le f'_+(x) \le \frac{f(x) - f(y)}{x - y} \le f'_-(y) \le f'_+(b).$$

Hence, f is Lipschitz continuous on $[a, b]$ with Lipschitz constant $L := |f'_+(a)| \vee |f'_-(b)|$.

(c): Continuity of f shows that for $x < z$

$$\frac{f(z) - f(x)}{z - x} = \lim_{y\downarrow x} \frac{f(z) - f(y)}{z - y} \ge \limsup_{y\downarrow x} f'_+(y).$$

Taking $z \downarrow x$ yields $f'_+(x) \ge \limsup_{y\downarrow x} f'_+(y)$. Since f'_+ is increasing, we must in fact have $f'_+(y) \to f'_+(x)$ as $y \downarrow x$. In the same way, one shows left-continuity of f'_-.

(d): Since the function f is Lipschitz continuous, it is absolutely continuous. By Lebesgue's differentiation theorem, f is hence a.e. differentiable and equal to the integral of its derivative, which is equal to $f'_-(x) = f'_+(x)$ for a.e. $x \in D$. □

Definition A.5. The *Fenchel–Legendre transform* of a function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$f^*(y) := \sup_{x \in \mathbb{R}} \big(y\,x - f(x)\big), \quad y \in \mathbb{R}.$$

If $f \not\equiv +\infty$, then f^* is a convex and lower semicontinuous as the supremum of the affine functions $y \mapsto y\,x - f(x)$. In particular, f^* is a proper convex function which is continuous on its effective domain. If f is itself a proper convex function, then f^* is also called the *conjugate function* of f.

Proposition A.6. *Let f be a proper convex function.*

(a) *For all $x, y \in \mathbb{R}$,*

$$xy \le f(x) + f^*(y) \tag{A.2}$$

with equality if x belongs to the interior of dom f *and if* $y \in [f'_-(x), f'_+(x)]$.

(b) *If f is lower semicontinuous, then $f^{**} = f$, i.e.,*

$$f(x) = \sup_{y \in \mathbb{R}} \big(x\,y - f^*(y)\big), \quad x \in \mathbb{R}.$$

Proof. (a): The inequality (A.2) is obvious. Now suppose that x_0 belongs to the interior of dom f. Proposition A.4 yields $f(x) \ge f(x_0) + f'_\pm(x_0)(x - x_0)$ for all x in the interior of dom f and, by upper semi-continuity, for all $x \in \mathbb{R}$. Hence, $f(x) \ge f(x_0) + y_0\,(x - x_0)$ whenever $y_0 \in [f'_-(x_0), f'_+(x_0)]$. This shows that $x\,y_0 - f(x) \le x_0\,y_0 - f(x_0)$ for all $x \in \mathbb{R}$, i.e.,

$$x_0\,y_0 - f(x_0) = \sup_{x \in \mathbb{R}} \big(x\,y_0 - f(x)\big) = f^*(y_0).$$

(b): We first show the following auxiliary claim: If $\beta < f(x_0)$, then there exists an affine function h such that $h(x_0) = \beta$ and $h(x) < f(x)$ for all x. For the proof of this claim let

$$\mathcal{C} := \{\, (x, a) \in \mathbb{R}^2 \mid f(x) \le a \,\}.$$

$\mathcal{C}$ is usually called the *epigraph* of f. Lower semicontinuity of f implies that $\mathcal{C}$ is closed. The point (x_0, β) does not belong to $\mathcal{C}$, and Proposition A.1 thus yields some $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2$ such that

$$\inf_{x \in \operatorname{dom} f} (\eta_1 x + \eta_2 f(x)) \ge \delta := \inf_{(x,a) \in \mathcal{C}} (\eta_1 x + \eta_2 a) > \eta_1 x_0 + \eta_2 \beta.$$

If $f(x_0) < \infty$, we get $\eta_1 x_0 + \eta_2 f(x_0) > \eta_1 x_0 + \eta_2 \beta$. Hence $\eta_2 > 0$, and one checks that

$$h(x) := -\frac{\eta_1}{\eta_2}(x - x_0) + \beta$$

is as desired. If $f(x_0) = \infty$ and $\eta_2 > 0$, then the same definition works. Now assume that $f(x_0) = \infty$ and $\eta_2 = 0$. Letting $\widetilde{h}(x) := \delta - \eta_1 x$ we have $\widetilde{h}(x_0) > 0$

and $\widetilde{h}(x) \leq 0$ for $x \in \operatorname{dom} f$. Since f is proper, the first step of the proof of our claim allows us to construct an affine function g with $g < f$. If $g(x_0) \geq \beta$, then $h := g + \beta - g(x_0)$ is as desired. Otherwise, we let $h(x) := g(x) + \lambda \widetilde{h}(x)$ for $\lambda := (\beta - g(x_0))/\widetilde{h}(x_0)$. This concludes the proof of our auxiliary claim.

Now we can prove part (b) of the assertion. It is clear from the definition that $f \geq f^{**}$. Suppose there exists a point x_0 such that $f(x_0) > f^{**}(x_0)$. Take β strictly between $f^{**}(x_0)$ and $f(x_0)$. By the auxiliary claim, there exists an affine function $h < f$ such that $h(x_0) = \beta$. Let us write $h(x) = y_0 x + \alpha$. Then it follows that $f^*(y_0) < -\alpha$ and hence

$$f^{**}(x_0) \geq y_0 x_0 - f^*(y_0) > h(x_0) = \beta,$$

which is a contradiction. □

A.2 Absolutely continuous probability measures

Suppose that P and Q are two probability measures on a measurable space $(\Omega, \mathcal{F})$.

Definition A.7. Q is said to be *absolutely continuous* with respect to P on the σ-algebra $\mathcal{F}$, and we write $Q \ll P$, if for all $A \in \mathcal{F}$,

$$P[\,A\,] = 0 \quad \Longrightarrow \quad Q[\,A\,] = 0.$$

If both $Q \ll P$ and $P \ll Q$ hold, we will say that Q and P are *equivalent*, and we will write $Q \approx P$.

The following characterization of absolute continuity is known as the Radon–Nikodym theorem:

Theorem A.8 (Radon–Nikodym). *Q is absolutely continuous with respect to P on $\mathcal{F}$ if and only if there exists an $\mathcal{F}$-measurable function $\varphi \geq 0$ such that*

$$\int F \, dQ = \int F\varphi \, dP \quad \textit{for all } \mathcal{F}\textit{-measurable functions } F \geq 0. \tag{A.3}$$

Proof. See, e.g., §17 of [20]. □

The function φ is called the *density* or *Radon–Nikodym derivative* of Q with respect to P, and we will write

$$\frac{dQ}{dP} := \varphi.$$

Clearly, the Radon–Nikodym derivative is uniquely determined through (A.3).

Corollary A.9. *If $Q \ll P$ on $\mathcal{F}$, then*

$$Q \approx P \quad \Longleftrightarrow \quad \frac{dQ}{dP} > 0 \ \ P\text{-a.s.}$$

In this case, the density of P with respect to Q is given by

$$\frac{dP}{dQ} = \left(\frac{dQ}{dP}\right)^{-1}.$$

Proof. Suppose that $Q \ll P$, let $\varphi := dQ/dP$. Take an $\mathcal{F}$-measurable function $F \geq 0$. Then

$$\int F\,dQ = \int_{\{\varphi>0\}} F\varphi\,dP = \int_{\{\varphi>0\}} F\,dQ.$$

In particular, $Q[\,\varphi = 0\,] = 0$. Replacing F with $F\varphi^{-1}$ yields

$$\int F\varphi^{-1}\,dQ = \int_{\{\varphi>0\}} F\varphi^{-1}\,dQ = \int_{\{\varphi>0\}} F\varphi^{-1}\varphi\,dP.$$

Note that the term on the right-hand side equals $\int F\,dP$ for all F if and only if $P[\,\varphi = 0\,] = 0$. This proves the result. $\square$

Remark A.10. Let us stress that absolute continuity depends on the underlying σ-field $\mathcal{F}$. For example, let P be the Lebesgue measure on $\Omega := [0, 1)$. Then every probability measure Q is absolutely continuous with respect to P on a σ-algebra $\mathcal{F}_0$ which is generated by finitely many intervals $[a_{i-1}, a_i)$ with $0 = a_0 < a_1 < \cdots < a_n = 1$. However, if we take for $\mathcal{F}$ the Borel σ-algebra on Ω, then for instance a Dirac point mass $Q = \delta_x$ is clearly not absolutely continuous with respect to P on $\mathcal{F}$. $\diamond$

While the preceding example shows that, in general, absolute continuity is not preserved under an enlargement of the underlying σ-algebra, the next proposition states that it is safe to take smaller σ-algebras. This proposition involves the notion of a *conditional expectation*

$$E[\,F \mid \mathcal{F}_0\,]$$

of an $\mathcal{F}$-measurable function $F \geq 0$ with respect to a probability measure P and a σ-algebra $\mathcal{F}_0 \subset \mathcal{F}$. Recall that $E[\,F \mid \mathcal{F}_0\,]$ may be defined as the P-a.s. unique $\mathcal{F}_0$-measurable random variable F_0 such that

$$E[\,F\,;\ A_0\,] = E[\,F_0\,;\ A_0\,] \quad \text{for all } A_0 \in \mathcal{F}_0; \tag{A.4}$$

see, e.g., § 15 of [19]. Note also our shorthand convention of writing

$$E[\,F\,;\ A_0\,] := E[\,F\,\mathrm{I}_{A_0}\,].$$

Clearly, we can replace in (A.4) the class of all indicator functions of sets in $\mathcal{F}_0$ by the class of all bounded $\mathcal{F}_0$-measurable functions or by the class of all non-negative $\mathcal{F}_0$-measurable functions.

Proposition A.11. *Suppose that Q and P are two probability measures on the measurable space $(\Omega, \mathcal{F})$ and that $Q \ll P$ on $\mathcal{F}$ with density φ. If $\mathcal{F}_0$ is a σ-algebra contained in $\mathcal{F}$, then $Q \ll P$ on $\mathcal{F}_0$, and the corresponding density is given by*

$$\frac{dQ}{dP}\Big|_{\mathcal{F}_0} = E[\,\varphi \mid \mathcal{F}_0\,] \quad P\text{-a.s.}$$

Proof. $Q \ll P$ on $\mathcal{F}_0$ follows immediately from the definition of absolute continuity. Since φ is the density on $\mathcal{F} \supseteq \mathcal{F}_0$, it follows for $A \in \mathcal{F}_0$ that

$$Q[\,A\,] = \int_A \varphi\, dP = \int_A E[\,\varphi \mid \mathcal{F}_0\,]\, dP.$$

Therefore the $\mathcal{F}_0$-measurable random variable $E[\,\varphi \mid \mathcal{F}_0\,]$ must coincide with the density on $\mathcal{F}_0$. □

Now we prove a formula for computing a conditional expectation $E_Q[\,F \mid \mathcal{F}_0\,]$ under a measure Q in terms of conditional expectations with respect to another measure P with $Q \ll P$.

Proposition A.12. *Suppose that $Q \ll P$ on $\mathcal{F}$ with density φ, and that $\mathcal{F}_0 \subseteq \mathcal{F}$ is another σ-algebra. Then, for any $\mathcal{F}$-measurable $F \geq 0$,*

$$E_Q[\,F \mid \mathcal{F}_0\,] = \frac{1}{E[\,\varphi \mid \mathcal{F}_0\,]} \cdot E[\,F\varphi \mid \mathcal{F}_0\,] \quad Q\text{-a.s.}$$

Proof. Suppose that $G_0 \geq 0$ is $\mathcal{F}_0$-measurable. Then

$$E_Q[\,G_0 F\,] = E[\,G_0 F\varphi\,] = E\big[\,G_0 \cdot E[\,F\varphi \,\big|\mathcal{F}_0\,]\big].$$

Let $\varphi_0 := E[\varphi \mid \mathcal{F}_0]$. Proposition A.11 implies that $\varphi_0 > 0$ Q-almost surely. Hence, we may assume that $G_0 = 0$ P-a.s. on $\{\varphi_0 = 0\}$, and Corollary A.9 yields

$$E\big[\,G_0 \cdot E[\,F\varphi\,\big|\mathcal{F}_0\,]\big] = E_Q\bigg[\,G_0 \cdot \frac{1}{E[\,\varphi \mid \mathcal{F}_0\,]} E[\,F\varphi \mid \mathcal{F}_0\,]\bigg].$$

This proves the assertion. □

If neither $Q \ll P$ nor $P \ll Q$ holds, one can use the following *Lebesgue decomposition* of P with respect to Q.

Theorem A.13. *For any two probability measures Q and P on $(\Omega, \mathcal{F})$, there exists a set $N \in \mathcal{F}$ with $Q[\,N\,] = 0$ and a $\mathcal{F}$-measurable function $\varphi \geq 0$ such that*

$$P[\,A\,] = P[\,A \cap N\,] + \int_A \varphi\, dQ \quad \text{for all } A \in \mathcal{F}.$$

One writes

$$\frac{dP}{dQ} := \begin{cases} \varphi & \text{on } N^c, \\ +\infty & \text{on } N. \end{cases}$$

Proof. Let $R := \frac{1}{2}(Q + P)$. Then both Q and P are absolutely continuous with respect to R with respective densities dQ/dR and dP/dR. Let

$$N := \left\{ \frac{dQ}{dR} = 0 \right\}.$$

Then $Q[\,N\,] = 0$. We define

$$\frac{dP}{dQ} := \varphi := \begin{cases} \dfrac{dP}{dR} \cdot \left(\dfrac{dQ}{dR}\right)^{-1} & \text{on } N^c, \\ +\infty & \text{on } N. \end{cases}$$

Then, for $\mathscr{F}$-measurable $f \geq 0$,

$$\begin{aligned} \int f\,dP &= \int_N f\,dP + \int_{N^c} f\,\frac{dP}{dR}\,dR \\ &= \int_N f\,dP + \int_{N^c} f\,\frac{dP}{dR} \cdot \left(\frac{dQ}{dR}\right)^{-1} dQ \\ &= \int_N f\,dP + \int f\,\varphi\,dQ, \end{aligned}$$

where we have used the fact that $Q[\,N\,] = 0$ in the last step. □

A.3 Quantile functions

Suppose that $F : (a, b) \to \mathbb{R}$ is an increasing function which is not necessarily strictly increasing. Let

$$c := \lim_{x \downarrow a} F(x) \quad \text{and} \quad d := \lim_{x \uparrow b} F(x).$$

Definition A.14. A function $q : (c, d) \to (a, b)$ is called an *inverse function* for F if

$$F\big(q(s)-\big) \leq s \leq F\big(q(s)+\big) \quad \text{for all } s \in (c, d).$$

The functions

$$q^-(s) := \sup\{\, x \in \mathbb{R} \mid F(x) < s \,\} \quad \text{and} \quad q^+(s) := \inf\{\, x \in \mathbb{R} \mid F(x) > s \,\}$$

are called the *left-* and *right-continuous inverse functions*.

The following lemma explains the reason for calling q^- and q^+ the left- and right-continuous inverse functions of F.

Lemma A.15. *A function $q : (c, d) \to (a, b)$ is an inverse function for F if and only if*

$$q^-(s) \le q(s) \le q^+(s) \quad \textit{for all } s \in (c, d).$$

In particular, q^- and q^+ are inverse functions. Moreover, q^- is left-continuous, q^+ is right-continuous, and every inverse function q is increasing and satisfies $q(s-) = q^-(s)$ and $q(s+) = q^+(s)$ for all $s \in (c, d)$. In particular, any two inverse functions coincide a.e. on (c, d).

Proof. We have $q^- \le q^+$, and any inverse function q satisfies $q^- \le q \le q^+$, due to the definitions of q^- and q^+. Hence, the first part of the assertion follows if we can show that $F(q^+(s)-) \le s \le F(q^-(s)+)$ for all s. But $x < q^+(s)$ implies $F(x) \le s$ and $y > q^-(s)$ implies $F(y) \ge s$, which gives the result.

Next, the set $\{ x \mid F(x) > s \}$ is the union of the sets $\{ x \mid F(x) > s + \varepsilon \}$ for $\varepsilon < 0$, and so q^+ is right-continuous. An analogous argument shows the left-continuity of q^-. It is clear that both q^- and q^+ are increasing, so that the second part of the assertion follows. □

Remark A.16. The left- and right-continuous inverse functions can also be represented as

$$q^-(s) = \inf\{ x \in \mathbb{R} \mid F(x) \ge s \} \quad \text{and} \quad q^+(s) = \sup\{ x \in \mathbb{R} \mid F(x) \le s \}.$$

To see this, note first that $q^-(s)$ is clearly dominated by the infimum on the right. On the other hand, $y > q^-(s)$ implies $F(y) \ge s$, and we get $q^-(s) \ge \inf\{ x \in \mathbb{R} \mid F(x) \ge s \}$. The proof for q^+ is analogous. ◇

Lemma A.17. *Let q be an inverse function for F. Then F is an inverse function for q. In particular,*

$$F(x+) = \inf\{ s \in (c, d) \mid q(s) > x \} \quad \textit{for } x \textit{ with } F(x) < d. \tag{A.5}$$

Proof. If $s > F(x)$ then $q(s) \ge q^-(s) \ge x$, and hence $q(F(x)+) \ge x$. Conversely, $s < F(x)$ implies $q(s) \le q^+(s) \le x$, and thus $q(F(x)-) \le x$. This proves that F is an inverse function for q. □

Remark A.18. By defining $q(d) := b$ we can extend (A.5) to

$$F(x+) = \inf\{ s \in (c, d] \mid q(s) > x \} \quad \text{for all } x \in (a, b).$$ ◇

From now on we will assume that

$$F : \mathbb{R} \longrightarrow [0, 1] \text{ is increasing and right-continuous}$$

and that F is normalized in the sense that $c = 0$ and $d = 1$. This assumption always holds if F is the *distribution function* of a random variable X on some probability space $(\Omega, \mathcal{F}, P)$, i.e., F is given by $F(x) = P[\, X \le x\,]$. The following lemma shows in particular that also the converse is true: any normalized increasing right-continuous functions $F : \mathbb{R} \to [0,1]$ is the distribution function of some random variable. By considering the laws of random variables, we also obtain the one-to-one correspondence $F(x) = \mu((-\infty, x])$ between all Borel probability measures μ on $\mathbb{R}$ and all normalized increasing right-continuous functions $F : \mathbb{R} \to [0,1]$.

Lemma A.19. *Let U be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ with a uniform distribution on $(0,1)$, i.e., $P[\, U \le s\,] = s$ for all $s \in (0,1)$. If q is an inverse function of a normalized increasing right-continuous function $F : \mathbb{R} \to [0,1]$, then*

$$X(\omega) := q\big(U(\omega)\big)$$

has the distribution function F.

Proof. First note that any inverse function for F is measurable because it coincides with the measurable function q^+ outside the countable set $\{\, s \in (0,1) \mid q^-(s) < q^+(s)\,\}$. Since $q(F(x)-) \le x$, we have $q(s) \le x$ for $s < F(x)$. Moreover, Lemma A.17 shows that $q(s) \le x$ implies $F(x) \ge F(q(s)) = F(q(s)+) \ge s$. It follows that

$$(0, F(x)) \subseteq \{\, s \in (0,1) \mid q(s) \le x\,\} \subseteq (0, F(x)].$$

Hence,

$$\begin{aligned} F(x) = P\big[\, U \in \big(0, F(x)\big)\,\big] &\le P\big[\, U \in \{\, s \mid q(s) \le x\,\}\,\big] \\ &\le P\big[\, U \in (0, F(x)]\,\big] = F(x). \end{aligned}$$

The assertion now follows from the identity $P[\, U \in \{\, s \mid q(s) \le x\,\}\,] = P[\, X \le x\,]$. □

Definition A.20. An inverse function $q : (0,1) \to \mathbb{R}$ of a distribution function F is called a *quantile function*. That is, q is a function with

$$F\big(q(s)-\big) \le s \le F\big(q(s)\big) \quad \text{for all } s \in (0,1).$$

The left- and rightcontinuous inverses,

$$q^-(s) = \sup\{\, x \in \mathbb{R} \mid F(x) < s\,\} \quad \text{and} \quad q^+(s) = \inf\{\, x \in \mathbb{R} \mid F(x) > s\,\},$$

are called the *lower* and *upper quantile functions*.

We will often use the generic notation F_X for the distribution function of a random variable X. When the emphasis is on the law μ of X, we will also write F_μ. In the same manner, we will write q_X or q_μ for the corresponding quantile functions. The value $q_X(\lambda)$ of a quantile function at a given level $\lambda \in (0,1)$ is often called a *λ-quantile* of X.

The following result complements Lemma A.19. It implies that a probability space supports a random variable with uniform distribution on $(0,1)$ if and only if it supports any non-constant random variable X with a continuous distribution.

Lemma A.21. *Let X be a random variable with a continuous distribution function F_X and with quantile function q_X. Then $U := F_X(X)$ is uniformly distributed on $(0, 1)$, and $X = q_X(U)$ P-almost surely.*

Proof. Let $(\widetilde{\Omega}, \widetilde{F}, \widetilde{P})$ be a probability space that supports a random variable $\widetilde{U}$ with a uniform distribution on $(0, 1)$. Then $\widetilde{X} := q_X(\widetilde{U})$ has the same distribution as X due to Lemma A.19. Hence, $F_X(X)$ and $F_X(\widetilde{X})$ also have the same distribution. On the other hand, if F_X is continuous, then $F_X\big(q_X(s)\big) = s$ and thus $F_X(\widetilde{X}) = \widetilde{U}$.

To show that $X = q_X(U)$ P-a.s., note first that $q_X^+(F(t)) \geq t$ and hence $q_X(U) = q_X^+(U) \geq X$ P-almost surely. Now let $f : \mathbb{R} \to (0, 1)$ be a strictly increasing function. Since $q_X(U)$ and X have the same law, we have $E[\, f(q_X(U))\,] = E[\, f(X)\,]$ and get $P[\, q_X(U) > X\,] = 0$. □

The following lemma uses the concept of the Fenchel–Legendre transform of a convex function as introduced in Definition A.5.

Lemma A.22. *Let X be a random variable with distribution function F_X and quantile function q_X such that $E[\,|X|\,] < \infty$. Then the Fenchel–Legendre transform of the convex function*

$$\Psi(x) := \int_{-\infty}^{x} F_X(z)\,dz = E[\,(x - X)^+\,]$$

is given by

$$\Psi^*(y) = \sup_{x\in\mathbb{R}}(xy - \Psi(x)) = \begin{cases} \int_0^y q_X(t)\,dt & \text{if } 0 \leq y \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Moreover, for $0 < y < 1$, the supremum above is attained in x if and only if x is a y-quantile of X.

Proof. Note first that, by Fubini's theorem and Lemma A.19,

$$\Psi(x) = E\Big[\int_{-\infty}^{x} \mathrm{I}_{\{X\leq z\}}\,dz\Big] = E[\,(x - X)^+\,] = \int_0^1 (x - q_X(t))^+\,dt. \tag{A.6}$$

It follows that $\Psi^*(y) = +\infty$ for $y < 0$, $\Psi^*(0) = -\inf_x \Psi(x) = 0$,

$$\Psi^*(1) = \sup_{x\in\mathbb{R}}(x - \Psi(x)) = \lim_{x\uparrow\infty}\int_0^1 x - (x - q_X(t))^+\,dt = \int_0^1 q_X(t)\,dt,$$

and $\Psi^*(y) = \infty$ for $y > 1$. To prove our formula for $0 < y < 1$, note that the right-hand and left-hand derivatives of the concave function $f(x) = xy - \Psi(x)$ are given by $f'_+(x) = y - F_X(x)$ and $f'_-(x) = y - F_X(x-)$. A point x is a maximizer of

f if $f'_+(x) \le 0$ and $f'_-(x) \ge 0$, which is equivalent to x being a y-quantile. Taking $x = q_X(y)$ and using (A.6) gives

$$\Psi(x) = \int_0^y (x - q_X(t))\,dt = xy - \int_0^y q_X(t)\,dt,$$

and our formula follows. □

Lemma A.23. *If $X = f(Y)$ for an* increasing *function f and q_Y is a quantile function for Y, then $f(q_Y(t))$ is a quantile function for X. In particular,*

$$q_X(t) = q_{f(Y)}(t) = f(q_Y(t)) \quad \textit{for a.e. } t \in (0,1),$$

for any quantile function q_X of X.

If f is decreasing*, then $f(q_Y(1-t))$ is a quantile function for X. In particular,*

$$q_X(t) = q_{f(Y)}(t) = f(q_Y(1-t)) \quad \textit{for a.e. } t \in (0,1).$$

Proof. If f is decreasing, then $q(t) := f(q_Y(1-t))$ satisfies

$$\begin{aligned} F_X(q(t)) &= P\big[\, f(Y) \le f\big(q_Y(1-t)\big)\,\big] \\ &\ge P[\, Y \ge q_Y(1-t)\,] \\ &\ge t \\ &\ge P[\, Y > q_Y(1-t)\,] \\ &\ge F_X(q(t)-), \end{aligned}$$

since $F_Y(q_Y(1-t)-) \le 1-t \le F_Y(q_Y(1-t))$ by definition. Hence $q(t) = f(q_Y(1-t))$ is a quantile function. A similar argument applies to an increasing function f. □

The following theorem is a version of the *Hardy–Littlewood inequalities*. They estimate the expectation $E[\,XY\,]$ in terms of quantile functions q_X and q_Y.

Theorem A.24. *Let X and Y be two random variables on $(\Omega, \mathcal{F}, P)$ with quantile functions q_X and q_Y. Then,*

$$\int_0^1 q_X(1-s)q_Y(s)\,ds \le E[\,XY\,] \le \int_0^1 q_X(s)q_Y(s)\,ds,$$

provided that all integrals are well defined. If $X = f(Y)$ and the lower (upper) bound is finite, then the lower (upper) bound is attained if and only if f can be chosen as a decreasing (increasing) function.

Proof. We first prove the result for $X, Y \geq 0$. By Fubini's theorem,

$$E[XY] = E\Big[\int_0^\infty \mathrm{I}_{\{X>x\}}\,dx \int_0^\infty \mathrm{I}_{\{Y>y\}}\,dy\Big] = \int_0^\infty \int_0^\infty P[X > x, Y > y]\,dx\,dy.$$

Since

$$P[X > x, Y > y] \geq (P[X > x] - P[Y \leq y])^+ = \int_0^1 \mathrm{I}_{\{F_Y(y)\leq s\}}\,\mathrm{I}_{\{s\leq 1-F_X(x)\}}\,ds,$$

and since

$$q_Z^+(s) = \sup\{x \geq 0 \mid F_Z(x) \leq s\} = \int_0^\infty \mathrm{I}_{\{F_Z(x)\leq s\}}\,dx$$

for any random variable $Z \geq 0$, another application of Fubini's theorem yields

$$E[XY] \geq \int_0^1 q_X^+(1-s)\,q_Y^+(s)\,ds = \int_0^1 q_X(1-s)\,q_Y(s)\,ds.$$

In the same way, the upper estimate follows from the inequality

$$P[X > x, Y > y] \leq P[X > x] \wedge P[Y > y] = \int_0^1 \mathrm{I}_{\{F_X(x)\leq s\}}\,\mathrm{I}_{\{F_Y(y)\leq s\}}\,ds.$$

For $X = f(Y)$,

$$E[XY] = E[f(Y)Y] = \int_0^1 f(q_Y(t))q_Y(t)\,dt, \tag{A.7}$$

due to Lemma A.19, and so Lemma A.23 implies that the upper and lower bounds are attained for increasing and decreasing functions, respectively.

Conversely, assume that $X = f(Y)$, and that the upper bound is attained and finite:

$$E[f(Y)Y] = \int_0^1 q_X(t)q_Y(t)\,dt < \infty. \tag{A.8}$$

Our aim is to show that

$$X = f(Y) = \widetilde{f}(Y) \quad P\text{-a.s.},$$

where $\widetilde{f}$ is the increasing function on $[0, \infty)$ defined by

$$\widetilde{f}(x) := q_X(F_Y(x))$$

if x is a continuity point of F_Y, and by

$$\widetilde{f}(x) := \frac{1}{F_Y(x) - F_Y(x-)} \int_{F_Y(x-)}^{F_Y(x)} q_X(t)\,dt$$

otherwise. Note that

$$\widetilde{f}(q_Y) = E_\lambda[\, q_X \mid q_Y \,], \tag{A.9}$$

where $E_\lambda[\,\cdot\,|\, q_Y\,]$ denotes the conditional expectation with respect to q_Y under the Lebesgue measure λ on $(0, 1)$. Hence, (A.8) takes the form

$$\infty > \int_0^1 f\big(q_Y(t)\big) q_Y(t)\,dt = \int_0^1 \widetilde{f}\big(q_Y(t)\big) q_Y(t)\,dt, \tag{A.10}$$

where we have used Lemma A.19. Let ν denote the distribution of Y. By introducing the positive measures $d\mu = f\,d\nu$ and $d\widetilde{\mu} = \widetilde{f}\,d\nu$, (A.10) can be written as

$$\int_0^\infty \mu([y,\infty))\,dy = \int x\,\mu(dx) = \int x\,\widetilde{\mu}(dx) = \int_0^\infty \widetilde{\mu}([y,\infty))\,dy. \tag{A.11}$$

On the other hand, with g denoting the increasing function $I_{[y,\infty)}$, the upper Hardy–Littlewood inequality, Lemma A.23, and (A.9) yield

$$\begin{aligned} \mu([y,\infty)) &= E[\, g(Y) f(Y) \,] \\ &\le \int_0^1 q_{g(Y)}(t) q_X(t)\,dt \\ &= \int_0^1 g\big(q_Y(t)\big) \widetilde{f}\big(q_Y(t)\big)\,dt \\ &= \widetilde{\mu}([y,\infty)). \end{aligned}$$

In view of (A.11), we obtain $\mu = \widetilde{\mu}$, hence $f = \widetilde{f}$ ν-a.s. and $X = \widetilde{f}(Y)$ P-almost surely. An analogous argument applies to the lower bound, and the proof for $X, Y \ge 0$ is concluded.

The result for general X and Y is reduced to the case of non-negative random variables by separately considering the positive and negative parts of X and Y:

$$\begin{aligned} E[\, XY \,] &= E[\, X^+Y^+ \,] - E[\, X^+Y^- \,] - E[\, X^-Y^+ \,] + E[\, X^-Y^- \,] \\ &\le \int_0^1 q_{X^+}(t) q_{Y^+}(t)\,dt - \int_0^1 q_{X^+}(t) q_{Y^-}(1-t)\,dt \\ &\quad - \int_0^1 q_{X^-}(1-t) q_{Y^+}(t)\,dt + \int_0^1 q_{X^-}(t) q_{Y^-}(t)\,dt, \end{aligned} \tag{A.12}$$

where we have used the upper Hardy–Littlewood inequality on the positive terms and the lower one on the negative terms. Since $q_{Z^+}(t) = (q_Z(t))^+$ and $q_{Z^-}(t) =$

$(q_Z(1-t))^-$ for all random variables due to Lemma A.23, one checks that the right-hand side of (A.12) is equal to $\int_0^1 q_X(t)q_Y(t)\,dt$, and we obtain the general form of the upper Hardy–Littlewood inequality. The same argument also works for the lower one.

Now suppose that $X = f(Y)$. We first note that (A.7) still holds, and so Lemma A.23 implies that the upper and lower bounds are attained for increasing and decreasing functions, respectively. Conversely, let us assume that the upper Hardy–Littlewood inequality is an identity. Then all four inequalities used in (A.12) must also be equalities. Using the fact that $XY^+ = f(Y^+)Y^+$ and $XY^- = f(-Y^-)Y^-$, the assertion is reduced to the case of non-negative random variables, and one checks that f can be chosen as an increasing function. The same argument applies if the lower Hardy–Littlewood inequality is attained. □

Remark A.25. For indicator functions of two sets A and B in $\mathcal{F}$, the Hardy–Littlewood inequalities reduce to the elementary inequalities

$$\big(P[\,A\,] + P[\,B\,] - 1\big)^+ \le P[\,A \cap B\,] \le P[\,A\,] \wedge P[\,B\,]; \tag{A.13}$$

note that these estimates were used in the preceding proof. Applied to the sets $\{X \le x\}$ and $\{Y \le y\}$, where X and Y are random variables with distribution functions F_X and F_Y and joint distribution function $F_{X,Y}$ defined by $F_{X,Y}(x, y) = P[\,X \le x,\ Y \le y\,]$, they take the form

$$\big(F_X(x) + F_Y(y) - 1\big)^+ \le F_{X,Y}(x, y) \le F_X(X) \wedge F_Y(y). \tag{A.14}$$

The estimates (A.13) and (A.14) are often called *Fréchet bounds*, and the Hardy–Littlewood inequalities provide their natural extension from sets to random variables. ◇

Definition A.26. A probability space $(\Omega, .\mathcal{F}, P)$ is called *atomless* if it contains no atoms. That is, there is no set $A \in \mathcal{F}$ such that $P[\,A\,] > 0$ and $P[\,B\,] = 0$ or $P[\,B\,] = P[\,A\,]$ whenever $B \in \mathcal{F}$ is a subset of A.

Proposition A.27. *For any probability space, the following conditions are equivalent.*

(a) $(\Omega, \mathcal{F}, P)$ *is atomless.*

(b) *There exists an i.i.d. sequence* $X_1, X_2, \dots$ *of random variables with Bernoulli distribution*
$$P[\,X_1 = 1\,] = P[\,X_1 = 0\,] = \frac{1}{2}.$$

(c) *For any* $\mu \in \mathcal{M}_1(\mathbb{R})$ *there exist i.i.d. random variables* $Y_1, Y_2, \dots$ *with common distribution* μ.

(d) $(\Omega, \mathcal{F}, P)$ *supports a random variable with a continuous distribution.*

Proof. (a)$\Rightarrow$(b): We need the following intuitive fact from measure theory: If $(\Omega, \mathcal{F}, P)$ is atomless, then for every $A \in \mathcal{F}$ and all δ with $0 \le \delta \le P[\,A\,]$ there exists a measurable set $B \subset A$ such that $P[\,B\,] = \delta$; see Theorem 9.51 of [2]. Thus, we may take a set $A \in \mathcal{F}$ such that $P[\,A\,] = 1/2$ and define $X_1 := 1$ on A and $X_1 := 0$ on A^c. Now suppose that $X_1, \dots, X_n$ have already been constructed. Then

$$P[\,X_1 = x_1, \dots, X_n = x_n\,] = 2^{-n}$$

for all $x_1, \dots, x_n \in \{0, 1\}$, and this property is equivalent to $X_1, \dots, X_n$ being independent with the desired symmetric Bernoulli distribution. For all $x_1, \dots, x_n \in \{0, 1\}$ we may choose a set

$$B \subset \{X_1 = x_1, \dots, X_n = x_n\}$$

such that $P[\,B\,] = 2^{-(n+1)}$ and define $X_{n+1} := 1$ on B and $X_{n+1} := 0$ on B^c. Clearly, the collection $X_1, \dots, X_{n+1}$ is again i.i.d. with a symmetric Bernoulli distribution.

(b)$\Rightarrow$(c): By relabeling the sequence $X_1, X_2, \dots,$ we may obtain a double-indexed sequence $(X_{i,j})_{i,j\in\mathbb{N}}$ of independent Bernoulli-distributed random variables. If we let

$$U_i := \sum_{n=1}^{\infty} 2^{-n} X_{i,n},$$

then it is straightforward to check that U_i has a uniform distribution. Let q be a quantile function for μ. Lemma A.19 shows that the i.i.d. sequence $Y_i := q(U_i)$, $i = 1, 2, \dots,$ has common distribution μ.

The proofs of the implications (c)$\Rightarrow$(d) and (d)$\Rightarrow$(a) are straightforward. $\square$

A.4 The Neyman–Pearson lemma

Suppose that P and Q are two probability measures on $(\Omega, \mathcal{F})$, and denote by

$$P[\,A\,] = P[\,A \cap N\,] + \int_A \frac{dP}{dQ}\, dQ, \qquad A \in \mathcal{F},$$

the Lebesgue decomposition of P with respect to Q as in Theorem A.13. For fixed $c \ge 0$, we let

$$A^0 := \left\{ \frac{dP}{dQ} > c \right\},$$

where we make use of the convention that $dP/dQ = \infty$ on N.

Proposition A.28 (Neyman–Pearson lemma). *If $A \in \mathcal{F}$ is such that $Q[\,A\,] \le Q[\,A^0\,]$, then $P[\,A\,] \le P[\,A^0\,]$.*

Proof. Let $F := \mathrm{I}_{A^0} - \mathrm{I}_A$. Then $F \geq 0$ on N, and $F \cdot (dP/dQ - c) \geq 0$. Hence

$$
\begin{aligned}
P[\,A^0\,] - P[\,A\,] &= \int F\,dP \\
&= \int_N F\,dP + \int F \cdot \frac{dP}{dQ}\,dQ \\
&\geq c \int F\,dQ \\
&= c(Q[\,A^0\,] - Q[\,A\,]).
\end{aligned}
$$

This proves the proposition. □

Remark A.29. In statistical test theory, A^0 is interpreted as the *likelihood quotient test* of the null hypothesis Q against the alternative hypothesis P: If the outcome ω of a statistical experiment is in A^0, then the null hypothesis is rejected. There are two possible kinds of error which can occur in such a test. A *type* 1 *error* occurs if the null hypotheses is rejected despite the fact that Q is the "true" probability. Similarly, a *type* 2 *error* occurs when the null hypothesis is not rejected, although Q is not the "true" probability. The probability of a type 1 error is given by $Q[\,A^0\,]$. This quantity is usually called the *size* or the *significance level* of the statistical test A^0. A type 2 error occurs with probability $P[\,(A^0)^c\,]$. The complementary probability $P[\,A^0\,] = 1 - P[\,(A^0)^c\,]$ is called the *power* of the test A^0. In this setting, the set A of Proposition A.28 can be regarded as another statistical test to which our likelihood quotient test is compared. The proposition can thus be restated as follows: *A likelihood quotient test has maximal power on its significance level.* ◇

Indicator functions of sets take only the values 0 and 1. We now generalize Proposition A.28 by considering $\mathcal{F}$-measurable functions $\psi : \Omega \to [0,1]$; let $\mathcal{R}$ denote the set of all such functions.

Theorem A.30. *Let* $\Pi := \frac{1}{2}(P + Q)$*, and define the density* $\varphi := dP/dQ$ *as above.*

(a) *Take* $c \geq 0$*, and suppose that* $\psi^0 \in \mathcal{R}$ *satisfies* Π*-a.s.*

$$
\psi^0 = \begin{cases} 1 & \text{on } \{\,\varphi > c\,\}, \\[2ex] 0 & \text{on } \{\,\varphi < c\,\}. \end{cases} \tag{A.15}
$$

Then, for any $\psi \in \mathcal{R}$*,*

$$
\int \psi\,dQ \leq \int \psi^0\,dQ \quad \Longrightarrow \quad \int \psi\,dP \leq \int \psi^0\,dP. \tag{A.16}
$$

(b) *For any* $\alpha_0 \in (0,1)$ *there is some* $\psi^0 \in \mathcal{R}$ *of the form* (A.15) *such that* $\int \psi^0 \, dQ = \alpha_0$. *More precisely, if* c *is an* $(1-\alpha_0)$*-quantile of* φ *under* Q*, we can define* ψ^0 *by*

$$\psi^0 = \mathrm{I}_{\{\varphi > c\}} + \kappa \mathrm{I}_{\{\varphi = c\}},$$

where κ *is defined as*

$$\kappa := \begin{cases} 0 & \text{if } Q[\,\varphi = c\,] = 0, \\ \dfrac{\alpha_0 - Q[\,\varphi > c\,]}{Q[\,\varphi = c\,]} & \text{otherwise.} \end{cases}$$

(c) *Any* $\psi^0 \in \mathcal{R}$ *satisfying* (A.16) *is of the form* (A.15) *for some* $c \geq 0$.

Proof. (a): Take $F := \psi^0 - \psi$ and repeat the proof of Proposition A.28.

(b): Let F denote the distribution function of φ under Q. Then $Q[\,\varphi > c\,] = 1 - F(c) \leq \alpha_0$ and

$$\begin{aligned} Q[\,\varphi = c\,] &= F(c) - F(c-) \\ &\geq F(c) - 1 + \alpha_0 \\ &= \alpha_0 - Q[\,\varphi > c\,]. \end{aligned}$$

Hence $0 \leq \kappa \leq 1$ and ψ^0 belongs to $\mathcal{R}$. The fact that $\int \psi^0 \, dQ = \alpha_0$ is obvious.

(c): Suppose that ψ^* satisfies

$$\int \psi \, dQ \leq \int \psi^* \, dQ \quad \Longrightarrow \quad \int \psi \, dP \leq \int \psi^* \, dP.$$

The cases in which $\alpha_0 := \int \psi^* \, dQ$ equals 0 or 1 are trivial. For $0 < \alpha_0 < 1$, we can take ψ^0 as in part (b). Then $\alpha_0 = \int \psi^* \, dQ = \int \psi^0 \, dQ$. One also has that $\int \psi^* \, dP = \int \psi^0 \, dP$, as can be seen by applying (A.16) to both ψ^* and ψ^0 with reversed roles. Hence, for $f := \psi^0 - \psi^*$ and $N = \{\varphi = \infty\}$,

$$0 = \int f \, dP - c \int f \, dQ = \int_N f \, dP + \int f \cdot (\varphi - c) \, dQ.$$

But (A.15) implies that both $f \geq 0$ P-a.s. on N, and $f \cdot (\varphi - c) \geq 0$ Q-a.s. Hence f vanishes Π-a.s. on $\{\varphi \neq c\}$. □

Remark A.31. In the context of Remark A.29, an element ψ of $\mathcal{R}$ is interpreted as a *randomized statistical test*: If ω is the outcome of a statistical experiment and $p := \psi(\omega)$, then the null hypothesis is rejected with probability p, i.e., after performing an independent random coin toss with success probability p. *Significance level* and *power* of a randomized test are defined as above, and a test of the form (A.15) is called a *generalized likelihood quotient test*. Thus, the general Neyman–Pearson lemma in the form of Theorem A.30 can be stated as follows: *A randomized test has maximal power on its significance level, if and only if it is a generalized likelihood quotient test.* ◇

A.5 The essential supremum of a family of random variables

In this section, we discuss the essential supremum of an arbitrary family Φ of random variables on a given probability space $(\Omega, \mathscr{F}, P)$. Consider first the case in which the set Φ is countable. Then $\varphi^*(\omega) := \sup_{\varphi\in\Phi} \varphi(\omega)$ will also be a random variable, i.e., φ^* is measurable. Measurability of the pointwise supremum, however, is not guaranteed if Φ is uncountable. Even if the pointwise supremum is measurable, it may not be the right concept, when we focus on almost sure properties. This can be illustrated by taking P as the Lebesgue measure on $\Omega := [0, 1]$ and $\Phi := \{\mathrm{I}_{\{x\}} \mid 0 \le x \le 1\}$. Then $\sup_{\varphi\in\Phi} \varphi(x) \equiv 1$ whereas $\varphi = 0$ P-a.s. for each single $\varphi \in \Phi$. This suggests the following notion of an *essential supremum* defined in terms of almost sure inequalities.

Theorem A.32. *Let Φ be any set of random variables on $(\Omega, \mathscr{F}, P)$.*

(a) *There exists a random variable φ^* such that*

$$\varphi^* \ge \varphi \quad P\text{-a.s. for all } \varphi \in \Phi. \tag{A.17}$$

Moreover, φ^* is almost surely unique in the following sense: Any other random variable ψ with property (A.17) *satisfies $\psi \ge \varphi^*$ P-almost surely.*

(b) *Suppose that Φ is directed upwards, i.e., for φ, $\widetilde{\varphi} \in \Phi$ there exists $\psi \in \Phi$ with $\psi \ge \varphi \vee \widetilde{\varphi}$. Then there exists an increasing sequence $\varphi_1 \le \varphi_2 \le \cdots$ in Φ such that $\varphi^* = \lim_n \varphi_n$ P-almost surely.*

Definition A.33. The random variable φ^* in Theorem A.32 is called the *essential supremum* of Φ with respect to P, and we write

$$\operatorname{ess\,sup} \Phi = \operatorname*{ess\,sup}_{\varphi\in\Phi} \varphi := \varphi^*.$$

The *essential infimum* of Φ with respect to P is defined as

$$\operatorname{ess\,inf} \Phi = \operatorname*{ess\,inf}_{\varphi\in\Phi} \varphi := -\operatorname*{ess\,sup}_{\varphi\in\Phi}(-\varphi).$$

Proof of Theorem A.32. Without loss of generality, we may assume that each $\varphi \in \Phi$ takes values in $[0, 1]$; otherwise we may consider $\widetilde{\Phi} := \{ f \circ \varphi \mid \varphi \in \Phi \}$ with $f : \mathbb{R} \to [0, 1]$ strictly increasing.

If $\Psi \subset \Phi$ is countable, let $\varphi_\Psi(\omega) := \sup_{\varphi\in\Psi} \varphi(\omega)$. Then φ_Ψ is measurable. We claim that the upper bound

$$c := \sup\{ E[\, \varphi_\Psi \,] \mid \Psi \subset \Phi \text{ countable } \}$$

is attained by some countable $\Psi^* \subset \Phi$. To see this, take Ψ_n with $E[\, \varphi_{\Psi_n} \,] \to c$ and let $\Psi^* := \bigcup_n \Psi_n$. Then Ψ^* is countable and $E[\, \varphi_{\Psi^*} \,] = c$.

We now show that $\varphi^* := \varphi_{\Psi^*}$ satisfies (A.17). Suppose that (A.17) does not hold. Then there exists $\varphi \in \Phi$ such that $P[\, \varphi > \varphi^*] > 0$. Hence $\Psi' := \Psi^* \cup \{\varphi\}$ satisfies

$$E[\, \varphi_{\Psi'} \,] > E[\, \varphi_{\Psi^*} \,] = c,$$

in contradiction to the definition of c. Furthermore, if ψ is any other random variable satisfying (A.17), then obviously $\psi \geq \varphi^*$.

Finally, the construction shows that φ_{Ψ^*} can be approximated by an increasing sequence if Ψ is directed upwards. □

Remark A.34. For a given random variable X let Φ be the set of all constants c such that $P[\, X > c\,] > 0$. The number

$$\operatorname{ess\,sup} X := \operatorname{ess\,sup} \Phi$$

is the smallest constant $c \leq +\infty$ such that $X \leq c$ P-a.s. and called the *essential supremum of* X with respect to P. The *essential infimum* of X is defined as

$$\operatorname{ess\,inf} X := \operatorname{ess\,sup}(-X). \qquad \diamond$$

A.6 Spaces of measures

Let S be a topological space. S is called *metrizable* if there exists a metric d on S which generates the topology of S. That is, the open d-balls

$$B_\varepsilon(x) := \{\, y \in S \mid d(x, y) < \varepsilon\,\}, \quad x \in S,\ \varepsilon > 0,$$

form a *base* for the topology of S in the sense that a set $U \subset S$ is open if and only if it can be written as a union of such d-balls. A convenient feature of metrizable spaces is that their topological properties can be characterized via convergent sequences. For instance, a subset A of the metrizable space S is closed if and only if for every convergent sequence in A its limit point is also contained in A. Moreover, a function $f : S \to \mathbb{R}$ is continuous at $y \in S$ if and only if $f(y_n)$ converges to $f(y)$ for every sequence (y_n) converging to y. We write

$$C_b(S)$$

for the set of all bounded and continuous functions on S.

The metrizable space S is called *separable* if there exists a countable dense subset $\{x_1, x_2, \dots\}$ of S. In this case, the Borel σ-algebra $\mathcal{S}$ of S is generated by the open d-balls $B_\varepsilon(x)$ with radii $\varepsilon > 0$, $\varepsilon \in \mathbb{Q}$, and centered in $x \in \{x_1, x_2, \dots\}$. In what follows, we will always assume that S is separable and metrizable. If, moreover, the metric d can be chosen to be complete, i.e., if every Cauchy sequence with respect to d converges to some point in S, then S is called a *Polish space*. Clearly, $\mathbb{R}^d$ with the Euclidean distance is a complete and separable metric space, hence a Polish space.

Let us denote by

$$\mathcal{M}(S) := \mathcal{M}(S, \mathcal{S})$$

the set of all non-negative finite measures on $(S, \mathscr{S})$. Every $\mu \in \mathcal{M}(S)$ is of the form $\mu = \alpha\,\nu$ for some factor $\alpha \in [0, \infty)$ and some probability measure ν on the measurable space $(S, \mathscr{S})$. The space of all probability measures on $(S, \mathscr{S})$ is denoted by

$$\mathcal{M}_1(S) = \mathcal{M}_1(S, \mathscr{S}).$$

Definition A.35. The *weak topology* on $\mathcal{M}(S)$ is the coarsest topology for which all mappings

$$\mathcal{M}(S) \ni \mu \mapsto \int f\,d\mu, \quad f \in C_b(S),$$

are continuous.

It follows from this definition that the sets

$$U_\varepsilon\big(\mu; f_1, \dots, f_n\big) := \bigcap_{i=1}^{n} \Big\{ \nu \in \mathcal{M}(S) \;\Big|\; \Big| \int f_i\,d\nu - \int f_i\,d\mu \Big| < \varepsilon \Big\} \tag{A.18}$$

for $\mu \in \mathcal{M}(S)$, $\varepsilon > 0$, $n \in \mathbb{N}$, and $f_1, \dots, f_n \in C_b(S)$ form a base for the weak topology on $\mathcal{M}(S)$; for details see, e.g., Section 2.13 of [2]. Since the constant function 1 is continuous,

$$\mathcal{M}_1(S) = \Big\{ \mu \in \mathcal{M}(S) \;\Big|\; \mu(S) = \int 1\,d\mu = 1 \Big\}$$

is a closed subset of $\mathcal{M}(S)$. A well-known example for weak convergence of probability measures is the classical *central limit theorem*; the following version is needed in Section 5.7.

Theorem A.36. *Suppose that for each $N \in \mathbb{N}$ we are given N independent random variables $Y_1^{(N)}, \dots, Y_N^{(N)}$ on $(\Omega_N, \mathcal{F}_N, P_N)$ which satisfy the following conditions:*

- *There are constants γ_N such that $\gamma_N \to 0$ and $|Y_k^{(N)}| \le \gamma_N$ P_N-a.s.*
- $\displaystyle\sum_{k=1}^{N} E_N[\,Y_k^{(N)}\,] \longrightarrow m.$
- $\displaystyle\sum_{k=1}^{N} \mathrm{var}_N\big(Y_k^{(N)}\big) \longrightarrow \sigma^2$*, where* var_N *denotes the variance with respect to* P_N.

Then the distributions of

$$Z_N := \sum_{k=1}^{N} Y_k^{(N)}, \quad N = 1, 2, \dots,$$

converge weakly to the normal distribution with mean m and variance σ^2.

Proof. See, for instance, the corollary to Theorem 7.1.2 of [40]. □

The following theorem allows us to examine the weak topology in terms of weakly converging *sequences* of measures.

Theorem A.37. *The space $\mathcal{M}(S)$ is separable and metrizable for the weak topology. If S is Polish, then so is $\mathcal{M}(S)$. Moreover, if S_0 is a dense subset of S, then the set*

$$\Big\{ \sum_{i=1}^{n} \alpha_i \delta_{x_i} \;\Big|\; \alpha_i \in \mathbb{Q}^+,\ x_i \in S_0,\ n \in \mathbb{N} \Big\}$$

of simple measures on S_0 with rational weights is dense in $\mathcal{M}(S)$ for the weak topology.

Proof. In most textbooks on measure theory, the previous result is proved for $\mathcal{M}_1(S)$ instead of $\mathcal{M}(S)$; see, e.g., Theorem 14.12 of [2]. The general case requires only minor modifications. It is treated in full generality in Chapter IX, §5, of [27]. □

The following characterization of weak convergence in $\mathcal{M}(S)$ is known as the "portmanteau theorem".

Theorem A.38. *For any sequence $\mu, \mu_1, \mu_2, \dots$ of measures in $\mathcal{M}(S)$, the following conditions are equivalent:*

(a) *The sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ.*

(b) *$\mu_n(S) \to \mu(S)$ and*

$$\limsup_{n\uparrow\infty} \mu_n(A) \le \mu(A) \quad \textit{for every closed set } A \subset S.$$

(c) *$\mu_n(S) \to \mu(S)$ and*

$$\liminf_{n\uparrow\infty} \mu_n(U) \ge \mu(U) \quad \textit{for every open set } U \subset S.$$

(d) *$\mu_n(B) \to \mu(B)$ for every Borel set B whose boundary ∂B is not charged by μ in the sense that $\mu(\partial B) = 0$.*

(e) *$\int f\, d\mu_n \to \int f\, d\mu$ for every bounded measurable function f which is μ-a.e. continuous.*

(f) *$\int f\, d\mu_n \to \int f\, d\mu$ for every bounded and uniformly continuous function f.*

Proof. The result is proved for $\mathcal{M}_1(S)$ in [2], Theorem 14.3. The general case requires only minor modifications; see Chapter IX of [27]. □

Remark A.39. It follows from the portmanteau theorem that, on $S = \mathbb{R}$, weak convergence of μ_n to μ is equivalent to the condition

$$F(x-) \leq \liminf_{n\uparrow\infty} F_n(x) \leq \limsup_{n\uparrow\infty} F_n(x) \leq F(x)$$

for the corresponding distribution functions (F_n) and F, or to the pointwise convergence of $F_n(x)$ to $F(x)$ in any continuity point of F. It is also equivalent to the condition

$$q_\mu^-(t) \leq \liminf_{n\uparrow\infty} q_n(t) \leq \limsup_{n\uparrow\infty} q_n(t) \leq q_\mu^+(t)$$

for any choice of the quantile functions q_n of μ_n, or to the pointwise convergence of $q_n(t)$ to $q_\mu^+(t)$ in any continuity point of q_μ^+. ◇

The next theorem can be regarded as a stability result for weak convergence.

Theorem A.40 (Slutsky). *Suppose that, for $n \in \mathbb{N}$, X_n and Y_n are real-valued random variables on $(\Omega_n, \mathcal{F}_n, P_n)$ such that the laws of X_n converge weakly to the law of X, and the laws of Y_n converge weakly to δ_y for some $y \in \mathbb{R}$. Then:*

(a) *The laws of $X_n + Y_n$ converge weakly to the law of $X + y$.*

(b) *The laws of $X_n \cdot Y_n$ converge weakly to the law of $X \cdot y$.*

Proof. See, for instance, Section 8.1 of [39]. □

We turn now to the fundamental characterization of the relative compact subsets of $\mathcal{M}(S)$ known as *Prohorov's theorem.*

Theorem A.41 (Prohorov). *Let S be a Polish space. A subset M of $\mathcal{M}(S)$ is relatively compact for the weak topology if and only if*

$$\sup_{\mu\in M} \mu(S) < \infty$$

and if M is tight, i.e., if for every $\varepsilon > 0$ there exists a compact subset K of S such that

$$\sup_{\mu\in M} \mu(K^c) \leq \varepsilon.$$

In particular, $\mathcal{M}_1(S)$ is weakly compact if S is a compact metric space.

Proof. For a proof in the context of probability measures, see for instance Theorem 1 in §III.2 of [187]. The general case requires only minor modifications; see Chapter IX of [27]. □

Example A.42. Take for S the positive half axis $[0, \infty)$ and define

$$\mu_n := \frac{n-1}{n}\,\delta_0 + \frac{1}{n}\,\delta_n \quad \text{and} \quad \mu := \delta_0,$$

where δ_x denotes the Dirac point mass in $x \in S$, i.e., $\delta_x(A) = \mathrm{I}_A(x)$. Clearly,

$$\int f\,d\mu_n \longrightarrow \int f\,d\mu \quad \text{for all } f \in C_b(S)$$

so that μ_n converges weakly to μ. However, if we take the continuous but unbounded function $f(x) = x$, then $\int f\,d\mu_n = 1$ for all n so that

$$\lim_{n\uparrow\infty} \int f\,d\mu_n = 1 \neq \int f\,d\mu. \qquad \diamond$$

The preceding example shows that the weak topology is not an appropriate topology for ensuring the convergence of integrals against unbounded test functions. Let us introduce a suitable transformation of the weak topology which will allow us to deal with certain classes of unbounded functions.

We fix a continuous function

$$\psi : S \mapsto [1, \infty)$$

which will serve as a *gauge function*, and we denote by

$$C_\psi(S)$$

the linear space of all continuous functions f on S for which there exists a constant c such that

$$|f(x)| \le c \cdot \psi(x) \quad \text{for all } x \in S.$$

Furthermore, we denote by

$$\mathcal{M}^\psi(S)$$

the set of all measures $\mu \in \mathcal{M}(S)$ such that $\int \psi\,d\mu < \infty$.

Definition A.43. The *ψ-weak topology* on $\mathcal{M}^\psi(S)$ is the coarsest topology for which all mappings

$$\mathcal{M}^\psi(S) \ni \mu \mapsto \int f\,d\mu, \quad f \in C_\psi(S),$$

are continuous.

Since the gauge function ψ takes values in $[1, \infty)$, every bounded continuous function f belongs to $C_\psi(S)$. It follows that all mappings

$$\mathcal{M}^\psi(S) \ni \mu \mapsto \int f\,d\mu, \quad f \in C_b(S),$$

are continuous. In particular, the set

$$\mathcal{M}_1^{\psi}(S) := \{\, \mu \in \mathcal{M}^{\psi}(S) \mid \mu(S) = 1 \,\}$$

of all Borel probability measures in $\mathcal{M}^{\psi}(S)$ is closed for the ψ-weak topology.

As in the case of the weak topology, it follows that the sets

$$U_\varepsilon^{\psi}\big(\mu; f_1, \dots, f_n\big) := \bigcap_{i=1}^{n} \Big\{\, \nu \in \mathcal{M}^{\psi}(S) \;\Big|\; \Big| \int f_i \, d\nu - \int f_i \, d\mu \Big| < \varepsilon \,\Big\}$$

for $\mu \in \mathcal{M}^{\psi}(S)$, $\varepsilon > 0$, $n \in \mathbb{N}$, and $f_1, \dots, f_n \in C_{\psi}(S)$ form a base for the ψ-weak topology on $\mathcal{M}^{\psi}(S)$.

Let us define a mapping

$$\Psi : \mathcal{M}(S) \longrightarrow \mathcal{M}^{\psi}(S)$$

by

$$d\Psi(\mu) := \frac{1}{\psi}\, d\mu, \qquad \mu \in \mathcal{M}(S).$$

Clearly, Ψ is a bijective mapping between the two sets $\mathcal{M}(S)$ and $\mathcal{M}^{\psi}(S)$. Moreover, if we apply Ψ to an open neighborhood for the weak topology as in (A.18), we get

$$\Psi\big(U_\varepsilon(\mu; f_1, \dots, f_n)\big) = U_\varepsilon^{\psi}\big(\Psi(\mu); f_1\psi, \dots, f_n\psi\big).$$

Since $f\psi \in C_{\psi}(S)$ for each bounded and continuous function f, and since every function in $C_{\psi}(S)$ arises in this way, we conclude that a subset U of $\mathcal{M}(S)$ is weakly open if and only if $\Psi(U)$ is open for the ψ-weak topology. Hence, Ψ is a homeomorphism. This observation allows us to translate statements for the weak topology into results for the ψ-weak topology:

Corollary A.44. *For separable and metrizable S, the space $\mathcal{M}^{\psi}(S)$ is separable and metrizable for the ψ-weak topology. If S is Polish, then so is $\mathcal{M}^{\psi}(S)$. Moreover, if S_0 is a dense subset of S, then the set*

$$\Big\{\, \sum_{i=1}^{n} \alpha_i \delta_{x_i} \;\Big|\; \alpha_i \in \mathbb{Q}^+,\ x_i \in S_0,\ n \in \mathbb{N} \,\Big\}$$

of simple measures on S_0 with rational weights is dense in $\mathcal{M}^{\psi}(S)$ for the ψ-weak topology.

The preceding corollary implies in particular that it suffices to consider ψ-weakly converging sequences when studying the ψ-weak topology. The following corollary is implied by the portmanteau theorem.

Corollary A.45. *A sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathcal{M}^\psi(S)$ converges ψ-weakly to μ if and only if*

$$\int f\,d\mu_n \longrightarrow \int f\,d\mu$$

for every measurable function f which is μ-a.e. continuous and for which exists a constant c such that $|f| \le c\cdot\psi$ μ-almost everywhere.

Prohorov's theorem translates as follows to our present setting:

Corollary A.46. *Let S be a Polish space and M be a subset of $\mathcal{M}^\psi(S)$. The following conditions are equivalent:*

(a) *M is relatively compact for the ψ-weak topology.*

(b) *We have*

$$\sup_{\mu\in M}\int \psi\,d\mu < \infty\,,$$

and for every $\varepsilon > 0$ there exists a compact subset K of S such that

$$\sup_{\mu\in M}\int_{K^c} \psi\,d\mu \le \varepsilon\,.$$

(c) *There exists a measurable function $\phi : S \to [1,\infty]$ such that each set*

$$\{\, x\in S \mid \phi(x) \le n\,\psi(x)\,\}\,,\quad n\in\mathbb{N}\,,$$

is relatively compact in S, and such that

$$\sup_{\mu\in M}\int \phi\,d\mu < \infty\,.$$

Proof. (a)⇔(b): This follows immediately from Theorem A.41 and the fact that Ψ is a homeomorphism.

(b)⇒(c): Take an increasing sequence $K_1 \subset K_2 \subset \cdots$ of compact sets in S such that

$$\sup_{\mu\in M}\int_{K_n^c} \psi\,d\mu \le 2^{-n},$$

and define ϕ by

$$\phi(x) := \psi(x) + \sum_{n=1}^{\infty} \mathrm{I}_{K_n^c}(x)\psi(x).$$

Then $\{\phi \le n\psi\} \subset K_n$. Moreover,

$$\sup_{\mu\in M}\int \phi\,d\mu \le \sup_{\mu\in M}\int \psi\,d\mu + 1 < \infty.$$

(c) ⇒(b): Since $\{\phi \le \psi\}$ is relatively compact, we have that

$$c := \sup\{\, \psi(x) \mid x \in S,\ \phi(x) \le \psi(x)\,\} < \infty,$$

and hence

$$\sup_{\mu\in M} \int \psi \, d\mu \le (1+c) \sup_{\mu\in M} \int \phi \, d\mu < \infty.$$

Moreover, for $n \ge \varepsilon^{-1} \sup_{\mu\in M} \int \phi \, d\mu$, the relatively compact set $K := \{\phi \le n\psi\}$ satisfies

$$\sup_{\mu\in M} \int_{K^c} \psi \, d\mu \le \frac{1}{n} \sup_{\mu\in M} \int_{K_n^c} \phi \, d\mu \le \varepsilon,$$

and so condition (b) is satisfied. □

We turn now to the task of identifying a linear functional on a space of functions as the integral with respect to a suitable measure.

Theorem A.47 (Riesz). *Let Ω be a compact metric space and suppose that I is a linear functional on $C(\Omega)$ that is non-negative in the sense that $f \ge 0$ everywhere on Ω implies $I(f) \ge 0$. Then there exists a unique positive Borel measure μ on Ω such that*

$$I(f) = \int f \, d\mu \quad \textit{for all } f \in C(\Omega).$$

To state a general version of the preceding theorem, we need the notion of a *vector lattice* of real-valued functions on an arbitrary set Ω. This is a linear space $\mathcal{L}$ that is stable under the operation of taking the pointwise maximum: for $f, g \in \mathcal{L}$ also $f \vee g \in \mathcal{L}$. One example is the space of all bounded measurable functions on $(\Omega, \mathcal{F})$. Another one is the space $C_b(\Omega)$ of all bounded continuous functions on a separable metric space Ω. In this case, the σ-algebra $\sigma(\mathcal{L})$ generated by $\mathcal{L}$ coincides with the Borel σ-algebra of the underlying metric space. Note that Theorem A.47 is implied by the following result, together with Dini's lemma as recalled in Remark 4.26.

Theorem A.48 (Daniell–Stone). *Let I be a linear functional on a vector lattice $\mathcal{L}$ of functions on Ω such that the following conditions hold:*

(a) *I is non-negative in the sense that $f \ge 0$ everywhere on Ω implies $I(f) \ge 0$.*

(b) *If (f_n) is a sequence in $\mathcal{L}$ such that $f_n \searrow 0$, then $I(f_n) \searrow 0$.*

Then there exists a unique positive measure μ on the measurable space $\big(\Omega, \sigma(\mathcal{L})\big)$ such that

$$I(f) = \int f \, d\mu \quad \textit{for all } f \in \mathcal{L}.$$

Proof. See, e.g., Theorem 4.5.2 of [72] or Satz 40.5 in [18]. □

Without the continuity assumption (b), the preceding result takes a different form, as we will discuss now.

Definition A.49. Let $(\Omega, \mathcal{F})$ be a measurable space. A mapping $\mu : \mathcal{F} \to \mathbb{R}$ is called a *finitely additive set function* if $\mu(\emptyset) = 0$, and if for any finite collection $A_1, \dots, A_n \in \mathcal{F}$ of mutually disjoint sets

$$\mu\Big(\bigcup_{i=1}^n A_i\Big) = \sum_{i=1}^n \mu(A_i).$$

We denote by $\mathcal{M}_{1,f} := \mathcal{M}_{1,f}(\Omega, \mathcal{F})$ the set of all those finitely additive set functions $\mu : \mathcal{F} \to [0, 1]$ which are normalized to $\mu(\Omega) = 1$. The *total variation* of a finitely additive set function μ is defined as

$$\|\mu\|_{\mathrm{var}} := \sup\Big\{ \sum_{i=1}^n |\mu(A_i)| \;\Big|\; A_1, \dots, A_n \text{ disjoint sets in } \mathcal{F}, n \in \mathbb{N} \Big\}.$$

The space of all finitely additive measures μ whose total variation is finite is denoted by $ba(\Omega, \mathcal{F})$.

We will now give a brief outline of the integration theory with respect to a measure $\mu \in ba := ba(\Omega, \mathcal{F})$; for details we refer to Chapter III in [76]. The space $\mathcal{X}$ of all bounded measurable functions on $(\Omega, \mathcal{F})$ is a Banach space if endowed with the *supremum norm*,

$$\|F\| := \sup_{\omega \in \Omega} \big|F(\omega)\big|, \qquad F \in \mathcal{X}.$$

Let $\mathcal{X}_0$ denote the linear subspace of all finitely valued step functions which can be represented in the form

$$F = \sum_{i=1}^n \alpha_i \, \mathrm{I}_{A_i},$$

for some $n \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, and disjoint sets $A_1, \dots, A_n \in \mathcal{F}$. For this F we define

$$\int F \, d\mu := \sum_{i=1}^n \alpha_i \, \mu(A_i),$$

and one can check that this definition is independent of the particular representation of F. Moreover,

$$\Big| \int F \, d\mu \Big| \le \|F\| \cdot \|\mu\|_{\mathrm{var}}. \tag{A.19}$$

Since $\mathcal{X}_0$ is dense in $\mathcal{X}$ with respect to $\| \cdot \|$, this inequality allows us to define the integral on the full space $\mathcal{X}$ as the extension of the continuous linear functional

$\mathcal{X}_0 \ni F \mapsto \int F\,d\mu$. Clearly, $\mathcal{M}_{1,f}$ is contained in ba, and we will denote the integral of a function $F \in \mathcal{X}$ with respect to $Q \in \mathcal{M}_{1,f}$ by

$$E_Q[\,F\,] := \int F\,dQ.$$

Theorem A.50. *The integral*

$$\ell(F) = \int F\,d\mu, \quad F \in \mathcal{X},$$

defines a one-to-one correspondence between continuous linear functionals ℓ on $\mathcal{X}$ and finitely additive set functions $\mu \in ba$.

Proof. By definition of the integral and by (A.19), it is clear that any $\mu \in ba$ defines a continuous linear functional on $\mathcal{X}$. Conversely, if a continuous linear functional ℓ is given, then we can define a finitely additive set function μ on $(\Omega, \mathcal{F})$ by

$$\mu(A) := \ell(\mathrm{I}_A), \quad A \in \mathcal{F}.$$

If $L \geq 0$ is such that $\ell(F) \leq L$ for $\|F\| \leq 1$, then $\|\mu\|_{\mathrm{var}} \leq L$, and so $\mu \in ba$. One then checks that the integral with respect to μ coincides with ℓ on $\mathcal{X}_0$. Since $\mathcal{X}_0$ is dense in $\mathcal{X}$, we see that $\int F d\mu$ and $\ell(F)$ coincide for all $F \in \mathcal{X}$. □

Remark A.51. Theorem A.50 yields in particular a one-to-one correspondence between set functions $Q \in \mathcal{M}_{1,f}$ and continuous linear functionals ℓ on $\mathcal{X}$ such that $\ell(1) = 1$ and $\ell(X) \geq 0$ for $X \geq 0$. ◇

Example A.52. Clearly, the set $\mathcal{M}_{1,f}$ coincides with the set $\mathcal{M}_1 := \mathcal{M}_1(\Omega, \mathcal{F})$ of all σ-additive probability measures if $(\Omega, \mathcal{F})$ can be reduced to a finite set, in the sense that $\mathcal{F}$ is generated by a finite partition of Ω. Otherwise, $\mathcal{M}_{1,f}$ is strictly larger than $\mathcal{M}_1$. Suppose in fact that there are infinitely many disjoint sets $A_1, A_2, \ldots \in \mathcal{F}$, take $\omega_n \in A_n$, and define

$$\ell_n(X) := \frac{1}{n}\sum_{i=1}^{n} X(\omega_i), \quad n = 1, 2, \ldots$$

The continuous linear functionals ℓ_n on $\mathcal{X}$ belong to the unit ball B_1 in the dual Banach space $\mathcal{X}'$. By Theorem A.62, there exists a cluster point ℓ of (ℓ_n). For any $X \in \mathcal{X}$ there is a subsequence (n_k) such that $\ell_{n_k}(X) \to \ell(X)$. This implies that $\ell(X) \geq 0$ for $X \geq 0$ and $\ell(1) = 1$. Hence, Theorem A.50 allows us to write $\ell(X) = E_Q[\,X\,]$ for some $Q \in \mathcal{M}_{1,f}$. But Q is not σ-additive, since $Q[\,A_n\,] = \ell(\mathrm{I}_{A_n}) = 0$ and $Q\big[\,\bigcup_n A_n\,\big] = 1$. ◇

A.7 Some functional analysis

Numerous arguments in this book involve *infinite-dimensional* vector spaces. Typical examples are the spaces L^p for $0 \le p \le \infty$, which we will introduce below. To this end, we first take $p \in (0, \infty]$ and denote by $\mathcal{L}^p(\Omega, \mathcal{F}, P)$ the set of all $\mathcal{F}$-measurable functions Z on $(\Omega, \mathcal{F}, P)$ such that $\|Z\|_p < \infty$, where

$$\|Z\|_p := \begin{cases} E[\,|Z|^p\,]^{1/p}, & \text{if } 0 < p < \infty, \\ \\ \inf\big\{\, c \ge 0 \mid P[\,|Z| > c\,] = 0 \,\big\}, & \text{if } p = \infty. \end{cases} \tag{A.20}$$

Let us also introduce the space $\mathcal{L}^0(\Omega, \mathcal{F}, P)$, defined as the set of all P-a.s. finite random variables. If no ambiguity with respect to σ-algebra and measure can arise, we may sometimes write $\mathcal{L}^p(P)$ or just $\mathcal{L}^p$ instead of $\mathcal{L}^p(\Omega, \mathcal{F}, P)$. For $p \in [0, \infty]$, the space $L^p(\Omega, \mathcal{F}, P)$, or just L^p, is obtained from $\mathcal{L}^p$ by identifying random variables which coincide up to a P-null set. Thus, L^p consists of all equivalence classes with respect to the equivalence relation

$$Z \sim \widetilde{Z} \quad :\Longleftrightarrow \quad Z = \widetilde{Z} \;\; P\text{-a.s.} \tag{A.21}$$

If $p \in [1, \infty]$ then the vector space L^p is a *Banach space* with respect to the norm $\|\cdot\|_p$ defined in (A.20), i.e., every Cauchy sequence with respect to $\|\cdot\|_p$ converges to some element in L^p. In principle, one should distinguish between a random variable $Z \in \mathcal{L}^p$ and its associated equivalence class $[Z] \in L^p$, of which Z is a representative element. In order to keep things simple, we will follow the usual convention of identifying Z with its equivalence class, i.e., we will just write $Z \in L^p$.

On the space L^0, we use the topology of convergence in P-measure. This topology is generated by the metric

$$d(X, Y) := E[\,|X - Y| \wedge 1\,], \quad X, Y \in L^0. \tag{A.22}$$

Note, however, that d is not a norm.

Definition A.53. A linear space E which carries a topology is called a *topological vector space* if every singleton $\{x\}$ for $x \in E$ is a closed set, and if the vector space operations are continuous in the following sense:

$$(x, y) \longmapsto x + y$$

is a continuous mapping from $E \times E$ into E, and

$$(\alpha, x) \longmapsto \alpha x$$

is a continuous mapping from $\mathbb{R} \times E$ into E.

Clearly, every Banach space is a topological vector space. The following result is a generalization of the separation argument in Proposition A.1 to an infinite-dimensional setting.

Theorem A.54. *In a topological vector space E, any two disjoint convex sets $\mathcal{B}$ and $\mathcal{C}$, one of which has an interior point, can be separated by a non-zero continuous linear functional ℓ on E, i.e.,*

$$\ell(x) \le \ell(y) \quad \textit{for all } x \in \mathcal{C} \textit{ and all } y \in \mathcal{B}. \tag{A.23}$$

Proof. See [76], Theorem V.2.8. □

If one wishes to *strictly* separate two convex sets by a linear functional in the sense that one has a strict inequality in (A.23), then one needs additional conditions both on the convex sets and on the underlying space E.

Definition A.55. A topological vector space E is called a *locally convex space* if its topology has a base consisting of convex sets.

If E is a Banach space with norm $\|\cdot\|$, then the open balls

$$\{\, y \in E \mid \|y - x\| < r \,\}, \quad x \in E,\ r > 0,$$

form by definition a base for the topology of E. Since such balls are convex sets, any Banach space is locally convex. The space $L^0(\Omega, \mathcal{F}, P)$ with the topology of convergence in P-measure, however, is *not* locally convex if $(\Omega, \mathcal{F}, P)$ has no atoms; see, e.g., Theorem 12.41 of [2].

The following theorem is one variant of the classical Hahn–Banach theorem on the existence of "separating hyperplanes".

Theorem A.56 (Hahn–Banach). *Suppose that $\mathcal{B}$ and $\mathcal{C}$ are two non-empty, disjoint, and convex subsets of a locally convex space E. Then, if $\mathcal{B}$ is compact and $\mathcal{C}$ is closed, there exists a continuous linear functional ℓ on E such that*

$$\sup_{x \in \mathcal{C}} \ell(x) < \inf_{y \in \mathcal{B}} \ell(y).$$

Proof. See, for instance, [178], p. 65, or [76], Theorem V.2.10. □

One corollary of the preceding result is that, on a locally convex space E, the collection

$$E' := \{\, \ell : E \to \mathbb{R} \mid \ell \text{ is continuous and linear} \,\}$$

separates the points of E, i.e., for any two distinct points $x, y \in E$ there exists some $\ell \in E'$ such that $\ell(x) \neq \ell(y)$. The space E' is called the *dual* or the *dual space* of E. For instance, if $p \in [1, \infty)$ it is well-known that the dual of $L^p(\Omega, \mathcal{F}, P)$ is given by $L^q(\Omega, \mathcal{F}, P)$, where $\frac{1}{p} + \frac{1}{q} = 1$. The following definition describes a natural way in which locally convex topologies often arise.

Definition A.57. Let E be linear space, and suppose that F is a linear class of linear functionals on E which separates the points of E. The F*-topology* on E, denoted by $\sigma(E, F)$, is the topology on E which is obtained by taking as a base all sets of the form

$$\{ y \in E \mid |\ell_i(y) - \ell_i(x)| < r,\ i = 1, \dots, n \},$$

where $n \in \mathbb{N}$, $x \in E$, $\ell_i \in F$, and $r > 0$. If E already carries a locally convex topology, then the E'-topology $\sigma(E, E')$ is called the *weak topology* on E.

If E is infinite-dimensional, then E is typically *not* metrizable in the F-topology. In this case, it may not suffice to consider converging sequences when making topological assertions; see, however, Theorem A.65 below. The following proposition summarizes a few elementary properties of the F-topology.

Proposition A.58. *Consider the situation of the preceding definition. Then:*

(a) *E is a locally convex space for the F-topology.*

(b) *The F-topology is the coarsest topology on E for which every $\ell \in F$ is continuous.*

(c) *The dual of E for the F-topology is equal to F.*

Proof. See, e.g., Section V.3 of [76]. □

Theorem A.59. *Suppose that E is a locally convex space and that $\mathcal{C}$ is a convex subset of E. Then $\mathcal{C}$ is weakly closed if and only if $\mathcal{C}$ is closed in the original topology of E.*

Proof. If the convex set $\mathcal{C}$ is closed in the original topology then, by Theorem A.56, it is the intersection of the halfspaces $H = \{\ell \le c\}$ such that $H \supset \mathcal{C}$, and thus closed in the weak topology $\sigma(E, E')$. The converse is clear. □

For a given locally convex space E we can turn things around and consider E as a set of linear functionals on the dual space E' by letting $x(\ell) := \ell(x)$ for $\ell \in E'$ and $x \in E$. The E-topology $\sigma(E', E)$ obtained in this way is called the *weak* topology* on E'. According to part (c) of Proposition A.58, E is then the topological dual of $(E', \sigma(E', E))$. For example, the Banach space $L^\infty := L^\infty(\Omega, \mathcal{F}, P)$ is the dual of L^1, but the converse is generally not true. However, L^1 becomes the dual of L^∞ if we endow L^∞ with the weak* topology $\sigma(L^\infty, L^1)$.

The mutual duality between E and E' allows us to state a general version of part (b) of Proposition A.6. As in the one-dimensional situation of Definition A.3, a convex function $f : E \to \mathbb{R} \cup \{+\infty\}$ is called a *proper convex function* if $f(x) < \infty$ for some $x \in \mathbb{R}$.

Definition A.60. The *Fenchel–Legendre transform* of a function $f : E \to \mathbb{R} \cup \{+\infty\}$ is the function f^* on E' defined by

$$f^*(\ell) := \sup_{x \in E} \big(\ell(x) - f(x)\big).$$

If $f \not\equiv +\infty$, then f^* is a proper convex and lower semicontinuous function as the supremum of affine functions. If f is itself a proper convex function, then f^* is also called the *conjugate function* of f.

Theorem A.61. *Let f be a proper convex function on a locally convex space E. If f is lower semicontinuous with respect to $\sigma(E, E')$, then $f = f^{**}$.*

It is straightforward to adapt the proof we gave in the one-dimensional case of Proposition A.6 to the infinite-dimensional situation of Theorem A.61; all one has to do is to replace the separating hyperplane lemma by the Hahn–Banach separation theorem in the form of Theorem A.56.

One of the reasons for considering the weak topology on a Banach space or, more generally, on a locally convex space is that typically more sets are compact for the weak topology than for the original topology. The following result shows that the unit ball in the dual of a Banach space is weak* compact. Here we use the fact that a Banach space $(E, \|\cdot\|_E)$ defines the following norm on its dual E':

$$\|\ell\|_{E'} := \sup_{\|x\|_E \le 1} \ell(x), \qquad \ell \in E'.$$

Theorem A.62 (Banach–Alaoglu). *Let E be a Banach space with dual E'. Then $\{ x \in E' \mid \|x\|_{E'} \le r \}$ is weak* compact for every $r \ge 0$.*

Proof. See, e.g., Theorem IV.21 in [164]. □

Theorem A.63 (Krein–Šmulian). *Let E be a Banach space and suppose that $\mathcal{C}$ is a convex subset of the dual space E'. Then $\mathcal{C}$ is weak* closed if and only if*

$$\mathcal{C} \cap \{ x \in E' \mid \|x\|_{E'} \le r \}$$

is weak closed for each $r > 0$.*

Proof. See Theorem V.5.7 in [76]. □

The preceding theorem implies the following characterization of weak* closed sets in L^∞.

Lemma A.64. *A convex subset $\mathcal{C}$ of L^∞ is weak* closed if for every $r > 0$*

$$\mathcal{C}_r := \mathcal{C} \cap \{ X \in L^\infty \mid \|X\|_\infty \le r \}$$

is closed in L^1.

Proof. Since $\mathcal{C}_r$ is convex and closed in L^1, it is weakly closed in L^1 by Theorem A.59. Since the natural injection

$$\big(L^\infty, \sigma(L^\infty, L^1)\big) \longrightarrow \big(L^1, \sigma(L^1, L^\infty)\big)$$

is continuous, $\mathcal{C}_r$ is $\sigma(L^\infty, L^1)$-closed in L^∞. Thus, $\mathcal{C}$ is weak* closed due to the Krein–Šmulian theorem. □

Finally, we state a few fundamental results on *weakly compact sets*.

Theorem A.65 (Eberlein–Šmulian). *For any subset A of a Banach space E, the following conditions are equivalent:*

(a) *A is* weakly sequentially compact*, i.e., any sequence in A has a subsequence which converges weakly in E.*

(b) *A is* weakly relatively compact*, i.e., the weak closure of A is weakly compact.*

Proof. See [76], Theorem V.6.1. □

Theorem A.66 (James). *In a Banach space E, a bounded and weakly closed convex subset A is weakly compact if and only if every continuous linear functional attains its supremum on A.*

Proof. See, for instance, [86]. □

The following result characterizes the weakly relatively compact subsets of the Banach space $L^1 := L^1(\Omega, \mathcal{F}, P)$. It implies, in particular, that a set of the form $\{ f \in L^1 \mid |f| \le g \}$ with given $g \in L^1$ is weakly compact in L^1.

Theorem A.67 (Dunford–Pettis). *A subset A of L^1 is weakly relatively compact if and only if it is bounded and uniformly integrable.*

Proof. See, e.g., Theorem IV.8.9 or Corollary IV.8.11 in [76]. □

Notes

In these notes, we do not make any attempt to give a systematic account of all the sources which have been relevant for the development of the field. We simply mention a number of references which had a direct influence on our decisions how to present the topics discussed in this book. A more comprehensive list of references can be found, e.g., in [129].

Chapter 1: The proof of Theorem 1.6 is based on Dalang, Morton, and Willinger [51]. Remark 1.17 and Example 1.18 are taken from Schachermayer [175]. Section 1.6 is mainly based on [175], with the exception of Lemma 1.63, which is taken from Kabanov and Stricker [122]. Our proof of Lemma 1.67 combines ideas from [122] with the original argument in [175], as suggested to us by Irina Penner. For a historical overview of the development of arbitrage pricing and for an outlook to continuous-time developments, we refer to Schachermayer [177]. For some mathematical connections between superhedging of call options as discussed in Section 1.3 and bounds on stop-loss premiums in insurance see Chapter 5 of Goovaerts et al. [105].

Chapter 2: The results on the structure of preferences developed in this chapter are, to a large extent, standard topics in mathematical economics. We refer to textbooks on expected utility theory such as Fishburn [84], [85], Kreps [138], or Savage [174], and to the survey articles in [10], [17]. The ideas and results of Section 2.1 go back to classical references such as Debreu [57], Eilenberg [78], Milgram [151], and Rader [163]. The theory of affine numerical representations in Section 2.2 was initiated by von Neumann and Morgenstern [155] and further developed by Herstein and Milnor [112]. The drastic consequences of the assumption that a favorable bet is rejected at any level of wealth, as explained in Lemma 2.50, were stressed by Rabin [162]. The discussion of the partial orders $\succcurlyeq_{\text{uni}}$, $\succcurlyeq_{\text{mon}}$, and $\succcurlyeq_{\text{bal}}$ in Sections 2.4 and 2.6 has a long history. A first version of Theorem 2.58 is already contained in Hardy, Littlewood, and Polya [107]. A complete treatment was given by Strassen [191]; this paper is also the source for Section 2.6. The economic interpretation of Theorem 2.58 was developed by Rothschild and Stiglitz [170], [171]. The analysis of robust preferences in Section 2.5 is mainly based on the ideas of Savage [174], Anscombe and Aumann [5], and Gilboa and Schmeidler [104]. See Gilboa [103] for an alternative axiomatic approach to the characterization in part (b) of Theorem 2.80, and [131] for a survey of related developments. In the context of robust statistics, a special case of Proposition 2.85 appears in Huber [114].

Chapter 3: Given a preference relation of von Neumann–Morgenstern type, the analysis of optimal portfolios in Section 3.1 or, more generally, of optimal asset profiles in Section 3.3 is a standard exercise, both in microeconomic theory and in convex optimization. Section 3.1 shows that the existence of a solution is equivalent to the absence to arbitrage; here we follow Rogers [168]. In the special case of exponential utility, the construction of the optimal portfolio is equivalent to the minimization of relative entropy as discussed in Section 3.2. This may be viewed as the financial interpretation of general results on entropy minimization in Csiszar [43], [44]. The methods for characterizing optimal asset profiles in Section 3.3 in terms of "first-order conditions" are well-known; see, for example, [79] or [129], where they are developed in greater generality. The optimization problem in Theorem 3.44, which is formulated in terms of the partial order $\succcurlyeq_{\text{uni}}$ and involves the Hardy–Littlewood inequalities of Theorem A.24, is less standard. Our discussion is based on Dybvig [77], Jouini and Kallal [120] and on a proof of Dana and Meilijson, and we are obliged to Rose-Anne Dana for introducing us to this topic; see [53] for further developments. As to the existence of Arrow–Debreu equilibria discussed in Section 3.4, we refer to the classical version in Debreu [56] and to the survey articles in [10]. In our financial context, equilibrium allocations do no longer involve commodity bundles in Euclidean space as in [56] but asset profiles described by random variables on a probability space. This formulation of the equilibrium problem goes back to Borch [26], where it was motivated by the problem of risk exchange in a reinsurance market. The systematic analysis of the equilibrium problem in an infinite-dimensional setting was developed by Bewley, Mas-Colell, and others; see [23] and, for example, [147], [148], [3]. In our introductory approach, the existence proof is reduced to an application of Brouwer's fixed point theorem. Here we benefitted from discussions with Peter Bank; see also Dana [52]. Examples 3.51 and Example 3.52 are based on Bühlmann [31]; see also [32]. As mentioned in Remark 3.58, the equilibrium discussion of interest rates requires an intertemporal setting; for a systematic discussion see, e.g., Duffie [73] and, in a different conceptual framework, Bank and Riedel [16].

Chapter 4: The axiomatic approach to coherent measures of risk and their acceptance sets was initiated by Artzner, Delbaen, Eber, and Heath [11], and most results of Section 4.1 are based on this seminal paper. The extension to convex measures of risk was given independently by Heath [110], Heath and Ku [111], Föllmer and Schied [93], and Frittelli and Rosazza Gianin [100]. The robust representation theorems in Section 4.2 are taken from [94]; the discussion of convex risk measures on a space of continuous functions corrects an error in [94] and in the first edition of this book; see also Krätschmer [134] for a further analysis. The representation theory on L^∞ as presented in Section 4.3 was developed by Delbaen [58], [59]; for the connection to the general duality theory as explained in Remarks 4.17 and 4.39 see [58], [59], [100], [101]. Among the results on Value at Risk and its various modifications in Section 4.4, Proposition 4.42 and Theorem 4.61 are taken from [11] and [58]. Average Value at Risk is discussed, e.g., by Acerbi and Tasche [1], Delbaen [58], and Rockafellar and Uryasev [167]. Remark 4.44 was pointed out to us by Ruszczynski; see [157]. The notation $V@R$ is taken from Pflug and Ruszczynski [158]. The representations of law-invariant risk measures given in Section 4.5 were first obtained in the coherent case by Kusuoka

[141]; see also Kunze [140] and Frittelli and Rosazza Gianin [101] for the extension to the general convex case. Theorem 4.64 in Section 4.6 was first proved in [141]. The representations of the core of a concave distortion in Theorem 4.73 and Corollary 4.74 are due to Carlier and Dana [34]. See also [35] for further applications. The study of Choquet integrals with respect to general set functions as used in Section 4.7 was started by Choquet [38]. The connections with coherent risk measures were observed by Delbaen [58], [59]. The two directions in Theorem 4.82 are due to Dellacherie [64] and Schmeidler [182], respectively. The proof via Lemma 4.83, which we give here, is taken from Denneberg [67]. Theorem 4.87 is due to Kusuoka [141]. The equivalence between (b) and (d) in Theorem 4.88 was first proved in [38], item (c) was added in [182]. The proof given here is based on [67]. The first part of Section 4.8 is based on [93], the second on Carr, Geman, and Madan [36]. Section 4.9 is taken from [93]. Theorem 4.106 is an extension of a classical result for Orlicz spaces; see Krasnoselskii and Rutickii [136]. For some mathematical connections between risk measures and premium principles in insurance see, e.g., Denneberg [68] and Wang and Dhaene [196]. Market equilibria in terms of risk measures are studied in Heath and Ku [111].

Chapter 5: Martingales in Finance have a long history; see, e.g., Samuelson [172]. In the context of dynamic arbitrage theory, martingales and martingale measures are playing a central role, both in discrete and continuous time; for a historical overview we refer again to Schachermayer [177]. The first four sections of this chapter are based on Harrison and Kreps [108], Kreps [137], Harrison and Pliska [109], Dalang, Morton, and Willinger [51], Stricker [193], Schachermayer [175], Jacka [117], Rogers [168], Ansel and Stricker [9], and Kabanov and Kramkov [121]. The binomial model of Section 5.5 was introduced by Cox, Ross, and Rubinstein in [42]. Geometric Brownian motion, which appears in Section 5.7 as the diffusion limit of binomial models, was proposed since the late 1950s by Samuelson and others as a model for price fluctuations in continuous time, following the re-discovery of the linear Brownian motion model of Bachelier [15]; see Samuelson [173] and Cootner [41]. The corresponding dynamic theory of arbitrage pricing in continuous time goes back to Black and Scholes [24] and Merton [150]. A Black–Scholes type formula for option pricing appears in Sprenkle [190] in an ad hoc manner, without the arbitrage argument introduced by Black and Scholes. The approximation of Black–Scholes prices for various options by arbitrage-free prices in binomial models goes back to Cox, Ross, and Rubinstein [42]. A functional version of Theorem 5.53, based on Donsker's invariance principle, can be found in [73].

Chapter 6: The dynamic arbitrage theory for American options begins with Bensoussan [21] and Karatzas [124]. A survey is given in Myeni [154]. The theory of optimal stopping problems as presented in Sections 6.1 and 6.2 was initiated by Snell [189]; see [156] for a systematic introduction. Stability under pasting as discussed in Section 6.4 has appeared under several names in various contexts; see Delbaen [60] for a number of references and for an extension to continuous time. Our discussion of upper and lower Snell envelopes in Section 6.5 uses ideas from Karatzas and Kou [127] and standard techniques from dynamic programming. The application to the time consistency of dynamic coherent risk measures recovers results by Artzner et al. [12], [13], [14] and Riedel [165]; see also Delbaen [60].

Chapter 7: Optional decompositions, or uniform Doob decompositions as we call them, and the resulting construction of superhedging strategies were first obtained by El Karoui and Quenez [80] in a jump-diffusion model. In a general semimartingale setting, the theory was developed by Kramkov [135] and Föllmer and Kabanov [89]. From a mathematical point of view, the existence of martingale measures with marginals determined by given option prices in Theorem 7.25 is a corollary of Strassen [191]; continuous-time analoga were proved by Doob [70] and Kellerer [132]. For the economic interpretation, see, e.g., Breeden and Litzenberger [28]. The results on superhedging of exotic derivatives by means of plain vanilla options stated in Theorems 7.27, 7.31, 7.33, and Corollary 7.34 are due to Hobson [113] and Brown, Hobson, and Rogers [29]; they are related to martingale inequalities of Dubins and Gilat [71].

Chapter 8: The analysis of quantile hedging was triggered by a talk of D. Heath in March 1995 at the Isaac Newton Institute on the results in Kulldorf [139], where an optimization problem for Brownian motion with drift is reduced to the Neyman–Pearson lemma. Section 8.1 is based on Föllmer and Leukert [91]; see also Karatzas [125], Cvitanic and Spivak [50], Cvitanic and Karatzas [49], and Browne [30]. The results in Section 8.2 on minimizing the shortfall risk are taken from Föllmer and Leukert [92]; see also Leukert [145], Cvitanic and Karatzas [48], Cvitanic [45], and Pham [159].

Chapter 9: In continuous-time models, dynamic arbitrage pricing with portfolio constraints was considered by Cvitanic and Karatzas [46], [47]. In a discrete-time model with convex constraints, absence of arbitrage was characterized by Carrassus, Pham, and Touzi [33]. In a general semimartingale setting, Föllmer and Kramkov [90] proved a uniform Doob decomposition and superhedging duality theorems for a predictably convex set of admissible trading strategies and for American contingent claims; see also Karatzas and Kou [127].

Chapter 10: The idea of quadratic risk minimization for hedging strategies goes back to Föllmer and Sondermann [97], where the optimality criterion was formulated with respect to a martingale measure. Extensions to the general case and the construction of minimal martingale measures were developed by Föllmer and Schweizer [96] and Schweizer; see, e.g., [183], [184]. Our exposition also uses arguments from Föllmer and Schweizer [95], Schäl [179], and Li and Xia [146]. Variance-optimal hedging was introduced by Duffie and Richardson [75] and further developed by Schweizer and others; the discrete-time theory as presented in Section 10.3 is based on Schweizer [185]. Melnikov and Nechaev [149] give an explicit formula for a variance-optimal strategy without condition (10.22); in fact, they show that their formula always defines a variance-optimal strategy if one does not insist on the square-integrability of the gains process at intermediate times. For a survey on related results, we refer to [186].

Bibliography

[1] Acerbi, C., Tasche, D., On the coherence of expected shortfall. J. Banking & Finance 26 (7), 2002, 1487–1503.

[2] Aliprantis, C. D., Border, K. C., Infinite dimensional Analysis. A hitchhiker's guide, 2nd edition. Springer-Verlag, Berlin, 1999.

[3] Aliprantis, C., Brown, D., Burkinshaw, O., Existence and optimality of competitive equilibria. Springer-Verlag, Berlin, 1989.

[4] Allais, M., Le comportement de l'homme rationell devant de risque: Critique des postulats et axiomes de l'école Américaine. Econometrica 21 (1953), 503–546 .

[5] Anscombe, F. J., Aumann, R. J., A definition of subjective probability. Ann. Math. Statistics 34 (1963), 199–205.

[6] Ansel, J.-P., Stricker, C., Quelques remarques sur un théorème de Yan. Séminaire de probabilités XXIV, 1988/89 (J. Azéma, P. André and M. Yor, eds.), Lecture Notes in Math. 1426, Springer-Verlag, Berlin, 1990, 266–274.

[7] Ansel, J.-P., Stricker, C., Lois de martingale, densités et décomposition de Föllmer-Schweizer. Ann. Inst. H. Poincaré Probab. Statist. 28 (1992), 375–392.

[8] Ansel, J.-P., Stricker, C., Unicité et existence de la loi minimale. In: Séminaire de probabilités XXVII (J. Azéma et al., eds.). Lecture Notes in Math. 1557, Springer-Verlag, Berlin, 1993, 22–29.

[9] Ansel, J.-P., Stricker, C., Couverture des actifs contingents et prix maximum. Ann. Inst. H. Poincaré Probab. Statist. 30 (1994), 303–315.

[10] Arrow, K., Hildebrandt, W., Intriligator, D., Sonnenschein, H. (eds.), Handbook of Mathematical Economics, Vol. I–IV, Elsevier Science Publishers, Amsterdam, 1991.

[11] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Coherent measures of risk. Math. Finance 9 (1999), 203–228.

[12] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H., Coherent multiperiod risk adjusted values and Bellman's priciple. Preprint, ETH Zürich (2003).

[13] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H., Coherent multiperiod risk measurement. Preprint, ETH Zürich (2003).

[14] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H., Multiperiod risk and coherent multiperiod risk measurement. Preprint, ETH Zürich (2003).

[15] Bachelier, L., Théorie de la spéculation. Ann. Sci. École Norm. Sup. 17 (1900), 21–86.

[16] Bank, P., Riedel. F., Existence and Structure of Stochastic Equilibria with Intertemporal Substitution. Finance Stoch. 5 (2001), 487–509.

[17] Barbera, S., Hammond, P. J., Seidl, C. (eds.), Handbook of Utility Theory, Vol. 1: Principles, Kluwer Academic Publishers, 1998.

[18] Bauer, H., Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. Third edition. Walter de Gruyter, Berlin–New York, 1978.

[19] Bauer, H., Probability theory. De Gruyter Stud. Math. 23, Walter de Gruyter, Berlin–New York, 1996.

[20] Bauer, H., Measure and integration theory. De Gruyter Stud. Math. 26, Walter de Gruyter, Berlin–New York, 2001.

[21] Bensoussan, A., On the theory of option pricing. Acta Appl. Math. 2 (1984), 139–158.

[22] Bernoulli, D., Specimen theoriae novae de mensura sortis. Commentarii Academiae Scientiarum Imperialis Petropolitanae 5 (1738), 175–192. Translated by L. Sommer: Econometrica 22 (1954), 23–36.

[23] Bewley, T., Existence of equilibria in economies with infinitely many commodities. J. Econom. Theory 43 (1972), 514–540.

[24] Black, F., Scholes, M., The pricing of options and corporate liabilities. J. Polit. Econom. 81 (1973), 637–654.

[25] Blackwell, D., Dubins, L., A Converse to the dominated convergence theorem. Illinois J. Math. 7 (1963), 508–514.

[26] Borch, K., Equilibrium in a re-insurance market. Econometrica 30 (1962), 424–444.

[27] Bourbaki, N., Intégration. Hermann, Paris, 1969.

[28] Breeden, D., Litzenberger, R., Prices of state-contingent claims implicit in option prices. J. Business 51 (1978), 621–651.

[29] Brown, H., Hobson, D., Rogers, L. C. G., Robust hedging of barrier options. Math. Finance 11 (2001), 285–314.

[30] Browne, S., The risks and rewards of minimizing shortfall probability. J. Portfolio Management 25 (1999), 76–85.

[31] Bühlmann, H., The general economic premium priciple. Astin Bulletin 14 (1984), 13–21.

[32] Bühlmann, H., Jewell, S., Optimal risk exchanges. Astin Bulletin 10 (1979), 243–262.

[33] Carassus, L., Pham, H., Touzi, N., No-arbitrage in discrete-time under portfolio constraints. Math. Finance 11 (2001), 315–329.

[34] Carlier, G., Dana, R. A., Core of convex distortions of a probability. J. Econom. Theory 113 (2) (2003), 199–222.

[35] Carlier, G., Dana, R. A., Insurance contracts with deductibles and upper limits. Preprint, Ceremade, Université Paris Dauphine (2002).

[36] Carr, P., Geman, H., Madan, D., Pricing and hedging in incomplete markets. J. Financial Econom. 62 (2001), 131–167.

[37] Cheridito, P., Delbaen, F., Kupper, M., Coherent and convex monetary risk measures for bounded càdlàg processes. Stochastic Process. Appl. 112 (1) (2004), 1–22.

[38] Choquet, G., Theory of capacities. Ann. Inst. Fourier 5 (1953/54), 131–295.

[39] Chow, Y. S., Teicher, H., Probability Theory: Independence, interchangeability, martingales. Springer-Verlag, New York–Heidelberg–Berlin, 1978.

[40] Chung, K.-L., A course in probability theory. 2nd ed. Academic Press, New York–London, 1974.

[41] Cootner, P. H. (ed.), The random character of stock market prices. MIT Press, Cambridge, MA, 1964.

[42] Cox, J., Ross, S., Rubinstein, M., Option pricing: a simplified approach. J. Financial Econom. 7 (1979), 229–263.

[43] Csisźar, I., I-divergence geometry of probability distributions and minimization problems. Ann. Probab. 3 (1975), 146–158.

[44] Csiszár, I., Sanov property, generalized I-projection and a conditional limit theorem. Ann. Probab. 12 (1984), 768–793.

[45] Cvitanic, J., Minimizing expected loss of hedging in incomplete and constrained markets. SIAM J. Control Optim. 38 (2000), 1050–1066.

[46] Cvitanic, J., Karatzas, I., Convex duality in constrained portfolio optimization. Ann. Appl. Probab. 2 (1992), 767–818.

[47] Cvitanic, J., Karatzas, I., Hedging contingent claims with constrained portfolios. Ann. Appl. Probab. 3 (1993), 652–681.

[48] Cvitanic, J., Karatzas, I., On dynamic measures of risk. Finance Stoch. 3 (1999), 451–482.

[49] Cvitanic, J., Karatzas, I., Generalized Neyman-Pearson lemma via convex duality. Bernoulli 7 (2001), 79–97.

[50] Cvitanic, J., Spivak, G., Maximizing the probability of perfect hedge. Ann. Appl. Probab. 9 (1999), 1303–1328.

[51] Dalang, R., Morton, A., Willinger, W., Equivalent martingale measures and no-arbitrage in stochastic securities market models. Stochastics Stochastics Rep. 29 (1990), 185–201.

[52] Dana, R.-A., Existence and uniqueness of equilibria when preferences are additively separable and gross substitute. Econometrica 61 (1993), 953–957.

[53] Dana, R.-A., Stochastic dominance, individual decision making, and equilibrium asset pricing. Preprint, Ceremade, Université Paris Dauphine (2002).

[54] Dana, R.-A., Jeanblanc-Piqué, M., Marchés financiers en temps continu. Economica (1994), Paris.

[55] Davis, M. H. A., Option pricing in incomplete markets. In: Mathematics of derivative securities (Dempster, M. A. H. and Pliska, S. R., eds.), Publ. Newton Inst. 15, Cambridge University Press, Cambridge, 1997, 216–226.

[56] Debreu, G., Theory of value: an axiomatic analysis of economic equilibrium. Cowles Foundation for Research in Economics at Yale University, Monograph 17, John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London, 1959.

[57] Debreu, G., Continuity properties of paretian utility. International Econ. Rev. 5 (1964), 285–293.

[58] Delbaen, F., Coherent risk measures on general probability spaces. In: Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann, Springer-Verlag, Berlin, 2002, 1–37.

[59] Delbaen, F., Coherent risk measures. Cattedra Galileiana. Scuola Normale Superiore, Classe di Scienze, Pisa, 2000.

[60] Delbaen, F., The structure of m-stable sets and in particular of the set of riskneutral measures. Preprint, ETH Zürich (2003).

[61] Delbaen, F., Grandits, P., Rheinländer, T., Samperi, D., Schweizer, M. and Stricker, C., Exponential Hedging and Entropic Penalties. Math. Finance 12 (2002), 99–123.

[62] Delbaen, F., Schachermayer, W., A general version of the fundamental theorem of asset pricing. Math. Ann. 300 (1994), 463–520.

[63] Delbaen, F., Schachermayer, W., The fundamental theorem of asset pricing for unbounded stochastic processes. Math. Ann. 312 (1998), 215–250.

[64] Dellacherie, C., Quelques commentaires sur les prolongements de capacités. Séminaire Probabilités V, Sttrasbourg, Lecture Notes in Math. 191, Springer-Verlag, Berlin, 1970.

[65] Dellacherie, C., Meyer, P.-A., Probabilités et potentiel. Chapitres IX à XI. Théorie discrète du potential. Revised edition. Publications de l'Institut de Mathématiques de l'Université de Strasbourg, XVIII. Actualités Sci. Indust. 1410, Hermann, Paris, 1983

[66] Dembo, A., Zeitouni, O., Large deviations techniques and applications. Second edition. Appl. Math. 38. Springer-Verlag, New York, 1998.

[67] Denneberg, D., Non-additive measure and integral. Theory Decis. Lib. Ser. B Math. Statist. Meth. 27, Kluwer Academic Publishers, Dordrecht, 1994.

[68] Denneberg, D., Distorted probabilities and insurance premium. Methods Operations Research 52 (1990), 21–42.

[69] Deuschel, J.-D., Stroock, D., Large deviations. Pure Appl. Math. 137, Academic Press, Inc., Boston, MA, 1989.

[70] Doob, J., Generalized sweeping-out and probability. J. Funct. Anal. (1968), 207–225.

[71] Dubins, L., Gilat, D., On the distribution of maxima of martingales. Proc. Amer. Math. Soc. 68 (1978), 337–338.

[72] Dudley, R., Real analysis and probability. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1989.

[73] Duffie, D., Security markets. Econom. Theory Econometrics Math. Econom., Academic Press, Boston, MA, 1988.

[74] Duffie, D., Dynamic asset pricing theory. 2nd ed. Princeton University Press, Princeton, NJ, 1996.

[75] Duffie, D., Richardson, H. R., Mean-variance hedging in continuous time. Ann. Appl. Probab. 1 (1991), 1–15.

[76] Dunford, N., Schwartz, J., Linear Operators. Part I: General Theory. Interscience Publishers, New York, 1958.

[77] Dybvig, P., Distributional analysis of portfolio choice. J. Business 61 (1988), 369–393.

[78] Eilenberg, S., Ordered topological spaces. Amer. J. Math. LXIII (1941), 39–45.

[79] Ekeland, I., Temam, R., Convex analysis and variational problems. Stud. Math. Appl. 1, North-Holland Publishing Co., Amsterdam–Oxford; American Elsevier Publishing Co., Inc., New York, 1976.

[80] El Karoui, N., Quenez, M.-C., Dynamic programming and pricing of contingent claims in an incomplete market. SIAM J. Control Optim. 33 (1995), 29–66.

[81] Elliott, R., Föllmer, H., Orthogonal martingale representation. In: Stochastic analysis. Liber amicorum for Moshe Zakai, Academic Press, Boston, MA, 1991, 139–152.

[82] Elliott, R., Kopp, P., Mathematics of financial markets. Springer Finance, Springer-Verlag, New York, 1999.

[83] Feller, W., An introduction to probability theory and its applications. 3rd edition. John Wiley & Sons, New York, 1968.

[84] Fishburn, P., Utility theory for decision making. Publ. Operations Res. 18, John Wiley, New York, 1970.

[85] Fishburn, P. C., Nonlinear preference and utility theory. The John Hopkins University Press, Baltimore, MD, 1988.

[86] Floret, K., Weakly compact sets. Lecture Notes in Math. 801, Springer-Verlag, Berlin, 1980.

[87] Föllmer, H., Calcul d'Ito sans probabilités. Seminaire de probabilites XV, Univ. Strasbourg 1979/80, Lect. Notes Math. 850 (1981), 143-150.

[88] Föllmer, H., Probabilistic Aspects of Financial Risk. Plenary Lecture at the Third European Congress of Mathematics. Proceedings of the European Congress of Mathematics, Barcelona 2000, Birkhäuser, Basel, 2001.

[89] Föllmer, H., Kabanov, Yu. M., Optional decomposition and Lagrange multipliers. Finance Stoch. 2 (1998), 69–81.

[90] Föllmer, H., Kramkov, D., Optional decompositions under constraints. Probab. Theory Related Fields 109 (1997), 1–25.

[91] Föllmer, H., Leukert, P., Quantile hedging. Finance Stoch. 3 (1999), 251–273.

[92] Föllmer, H., Leukert, P., Efficient hedging: cost versus shortfall risk. Finance Stoch. 4 (2000), 117–146.

[93] Föllmer, H., Schied, A., Convex measures of risk and trading constraints. Finance Stoch. 6 (4) (2002).

[94] Föllmer, H., Schied, A., Robust representation of convex measures of risk. In: Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann, Springer-Verlag, Berlin, 2002, 39–56.

[95] Föllmer, H., Schweizer, M., Hedging by sequential regression: An introduction to the mathematics of option trading. Astin Bulletin 18 (1989), 147–160.

[96] Föllmer, H., Schweizer, M., Hedging of contingent claims under incomplete information. Applied stochastic analysis, London, Stochastic Monogr. 5 (1990), 389–414.

[97] Föllmer, H., Sondermann, D., Hedging of non-redundant contingent claims. Contributions to mathematical economics, Hon. G. Debreu, 206–223 (1986).

[98] Frittelli, M., Introduction to a theory of value coherent with the no arbitrage principle, Finance Stoch. 4 (3) (2000), 275–297.

[99] Frittelli, M., The minimal entropy martingale measure and the valuation problem in incomplete markets, Math. Finance 10 (2000), 39–52.

[100] Frittelli, M., Rosazza Gianin, E., Putting order in risk measures. J. Banking & Finance 26 (2002), 1473–1486.

[101] Frittelli, M., Rosazza Gianin, E., Dynamic convex risk measures. In: New Risk Measures in Investment and Regulation, G. Szegö ed., John Wiley & Sons, New York, 2003.

[102] Frittelli, M., Scandolo, G., Risk measures and capital requirements for processes. Preprint, Università di Firenze (2004).

[103] Gilboa, I., Expected utility with purely subjective non-additive probabilities. J. Math. Econom. 16 (1987), 65–88.

[104] Gilboa, I., Schmeidler, D., Maxmin expected utility with non-unique prior. J. Math. Econom. 18 (1989), 141–153.

[105] Goovaerts, M. J., De Vylder, F., Haezendonck, J., Insurance premiums. Theory and applications. North-Holland Publishing Co., Amsterdam, 1984.

[106] Halmos, P., Savage, L., Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann. Math. Stat. 20 (1949), 225–241.

[107] Hardy, G. H., Littlewood, J. E., Pólya, G., Inequalities. 2$^{\text{nd}}$ edition, Cambridge University Press, Cambridge, 1952.

[108] Harrison, J., Kreps, D., Martingales and arbitrage in multiperiod securities markets. J. Econom. Theory 20 (1979), 381–408.

[109] Harrison, J., Pliska, S., Martingales and stochastic integrals in the theory of continuous trading. Stochastic Process. Appl. 11 (1981), 215–260.

[110] Heath, D., Back to the future. Plenary lecture, First World Congress of the Bachelier Finance Society, Paris, 2000.

[111] Heath, D., Ku, H., Pareto equilibria with coherent measures of risk. Math. Finance 14 (2) (2004), 163–172.

[112] Herstein, I. N., Milnor J., An axiomatic approach to measurable utility. Econometrica 21 (1953), 291–297.

[113] Hobson, D., Robust hedging of the loockback option. Finance Stoch. 2 (1998), 329–347.

[114] Huber, P., Robust statistics. Wiley Ser. Probab. Math. Statist., Wiley, New York, 1981.

[115] Huber, P., Strassen, V., Minimax tests and the Neyman-Pearson lemma for capacities. Ann. Statist. 1 (1973), 251–263.

[116] Hull, J., Options, futures, and other derivatives. 4th edition. Prentice Hall, London, 2000.

[117] Jacka, S., A martingale representation result and an application to incomplete financial markets. Math. Finance 2 (1992), 239–250.

[118] Jaschke, S., Küchler, U., Coherent Risk Measures and Good-Deal Bounds. Finance Stoch. 5 (2) (2001), 181–200

[119] Jouini, E., Kallal, H., Arbitrage in securities markets with short-sales constraints. Math. Finance 3 (1995), 197–232.

[120] Jouini, E., Kallal, H., Efficient trading strategies in the presence of market frictions. The Review of Financial Studies 14 (2001), 343–369.

[121] Kabanov, Yu. M., Kramkov, D. O., No-arbitrage and equivalent martingale measures: An elementary proof of the Harrison-Pliska theorem. Theory Probab. Appl. 39 (1994), 523–527.

[122] Kabanov, Yu. M., Stricker, C., A teacher's note on no-arbitrage criteria. Sém. Probab. 35, Lecture Notes in Math. 1755, Springer-Verlag, 2001, Berlin, 149–152.

[123] Kahneman, D., Tversky, A., Prospect theory: An analysis of decision under risk. Econometrica 47 (1979), 263–291.

[124] Karatzas, I., On the pricing of American options. App. Math. Optimization 17 (1988), 37–60.

[125] Karatzas, I., Lectures on the mathematics of finance. CRM Monogr. Ser. 8, American Mathematical Society, Providence, RI, 1996.

[126] Karatzas, I., Kou, S. G., On the pricing of contingent claims under constraints. Ann. Appl. Probab. 6 (1996), 321–369.

[127] Karatzas, I., Kou, S. G., Hedging American contingent claims with constraint portfolios. Finance Stoch. 2 (1998), 215–258.

[128] Karatzas, I., Shreve, S., Brownian motion and stochastic calculus. Grad. Texts in Math. 113, Springer-Verlag, New York, 1991.

[129] Karatzas, I., Shreve, S., Methods of mathematical finance. Appl. Math., Springer-Verlag, Berlin, 1998.

[130] Karni, E., Schmeidler, D., Vind, K., On state-dependent preferences and subjective probabilities. Econometrica 51 (1983), 1021–1031.

[131] Karni, E., Schmeidler, D., Utility theory with uncertainty. In: Handbook of Mathematical Economics, Vol. IV (W. Hildebrandt and H. Sonnenschein, eds.), Elsevier Science Publishers, Amsterdam, 1991, 1763–1831.

[132] Kellerer, H., Markov-Komposition und eine Anwendung auf Martingale. Math. Ann. 198 (1972), 99–122.

[133] Komlos, J., A generalization of a problem of Steinhaus. Acta Math. Acad. Sci. Hung. 18 (1967), 217–229.

[134] Krätschmer, V., Robust representation of convex risk measures by probability. Preprint, Universität des Saarlandes (2004).

[135] Kramkov, D. O., Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets. Probab. Theory Related Fields 105 (1996), 459–479.

[136] Krasnosel'skii, M., Rutickii, Ya., Convex functions and Orlicz spaces. Gordon and Breach Science Publishers, 1961.

[137] Kreps, D., Arbitrage and equilibrium in economies with infinitely many commodities. J. Math. Econom. 8 (1981) 15–35.

[138] Kreps, D., Notes on the theory of choice. Westview Press, Boulder, CO, 1988.

[139] Kulldorff, M., Optimal control of favorable games with a time limit. SIAM J. Control Optim. 31 (1993), 52–69.

[140] Kunze, M., Verteilungsinvariante konvexe Risikomaße. Diplomarbeit, Humboldt-Universität zu Berlin, 2003.

[141] Kusuoka, S., On law invariant coherent risk measures. Adv. Math. Econ. 3 (2001), 83–95.

[142] Lamberton, D., Lapeyre, B., Introduction to stochastic calculus applied to finance. Chapman & Hall, London, 1995.

[143] Lamberton, D., Lapeyre, B., Introduction au calcul stochastique appliqué à la finance. 2ème éd. Ellipses, Paris, 1997.

[144] Leland, H., Who should buy portolio insurance? J. Finance 35 (1980), 581–594.

[145] Leukert, P., Absicherungsstrategien zur Minimierung des Verlustrisikos. Dissertation, Humboldt-Universität zu Berlin (1999).

[146] Li, P., Xia, J., Minimal martingale measures in discrete time. Acta Math. Appl. Sinica (2002).

[147] Mas-Colell, A., The theory of general equilibrium: A differentiable approach. Cambridge University Press, Cambridge, 1985.

[148] Mas-Colell, A., Zame, W., Equilibrium theory in infinite dimensional spaces. In: Handbook of Mathematical Economics IV (W. Hildebrandt and H. Sonnenschein, eds.), Elsevier Science Publishers, Amsterdam, 1991, 1835–1898.

[149] Melnikov, A. V., Nechaev, M. L., On the mean-variance hedging problem. Theory Probab. Appl. 43 (1998), 588–603; translation from Teor. Veroyatnost. i Primenen. 43 (1998), 672–691.

[150] Merton, R., Theory of rational option pricing. Bell J. Econom. Manag. Sci. 4 (1973), 141–183.

[151] Milgram, A. N., Partially ordered sets, separating systems and inductiveness. Rep. Math. Colloq. II Ser. 1, University of Notre Dame, 1939, 18–30.

[152] Monat, P., Stricker, C., Föllmer-Schweizer decomposition and mean-variance hedging for general claims. Ann. Probab. 23 (1995), 605–628.

[153] Musiela, M., Rutkowski, M., Martingale methods in financial modelling. Appl. Math., Springer-Verlag, Berlin, 1997.

[154] Myeni, R., The pricing of the American option. Ann. Appl. Probab. 2 (1992), 1–23.

[155] von Neumann, J., Morgenstern, O., Theory of games and economic behavior. 2nd. Ed. Princeton University Press, Princeton, NJ, 1947.

[156] Neveu, J., Discrete-parameter martingales. North-Holland, Amsterdam–Oxford, 1975.

[157] Ogryczak, W., Ruszczynski, A., Dual stochastic dominance and related mean-risk models. SIAM J. Optim. 13 (1) (2002), 60–78.

[158] Pflug, G., Ruszczynski, A., Risk measures for income streams. Preprint, Universität Wien (2001).

[159] Pham, H., Dynamic L^p-hedging in discrete time under cone constraints. SIAM J. Control Optim. 38 (2000), 665–682.

[160] Pliska, S., Introduction to Mathematical Finance. Blackwell Publishers, 1997.

[161] Quiggin, J., A Theory of Anticipated Utility. J. Economic Behavior and Organization 3 (1982), 225–243.

[162] Rabin, M., Risk aversion and expected-utility theory: A calibration theorem. Econometrica 68 (5) (2000), 1281–1292.

[163] Rader, J., The existence of a utility function to represent preferences. Rev. Econom. Stud. 30 (1963), 229–232.

[164] Reed, M., Simon, B., Methods of modern mathematical physics. Vol. 1: Functional analysis. Academic Press, New York–London, 1972.

[165] Riedel, F., Dynamic coherent risk measures. Stoch. Proc. Appl. 112 (2) (2004), 185–200.

[166] Rockafellar, R. T., Convex analysis. Princeton University Press, Princeton, NJ, 1970.

[167] Rockafellar, R. T., Uryasev, S., Conditional value-at-risk for general loss distributions. Research report 2001-5, ISE Dept., University of Florida, 2001.

[168] Rogers, L. C. G., Equivalent martingale measures and no-arbitrage. Stochastics Stochastics Rep. 51 (1994), 41–49.

[169] Ross, S. A., The arbitrage theory of capital asset pricing. J. Econom. Theory 13 (1976), 341–360.

[170] Rothschild, M., Stiglitz, J., Increasing risk: I. A definition. J. Econom. Theory 2 (1970), 225–243.

[171] Rothschild, M., Stiglitz, J., Increasing risk: II. Its economic consequences. J. Econom. Theory 3 (1971), 66–84.

[172] Samuelson, P. A., Proof that properly anticipated prices fluctuate randomly. Industrial Management Rev. 4 (1965), 41–50.

[173] Samuelson, P. A., Rational theory of warrant pricing. Industrial Management Rev. 6 (1965), 13–31.

[174] Savage, L. J., The foundations of statistics. Wiley Publ. Stat., John Wiley and Sons, New York, 1954.

[175] Schachermayer, W., A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time. Insurance Math. Econom. 11 (1992), 249–257.

[176] Schachermayer, W., Optimal Investment in Incomplete Financial Markets. Proceedings of the first World Congress of the Bachelier Society, Paris 2000, Springer-Verlag, Berlin, 2001.

[177] Schachermayer, W., Introduction to the mathematics of financial markets. In: Lectures on probability theory and statistics (Saint-Flour, 2000) , Lecture Notes in Math. 1816, Springer-Verlag, Berlin, 107–179.

[178] Schaefer, H., Topological vector spaces. Grad. Texts in Math. 3, Springer-Verlag, New York, 1971.

[179] Schäl, M., On quadratic cost criteria for option hedging. Math. Oper. Research 19 (1994), 121–131.

[180] Schied, A., On the Neyman-Pearson problem for law-invariant risk measures and robust utility functionals. Ann. Appl. Probab. 14 (2004), 1398–1423.

[181] Schied, A., Optimal investments for robust utility functionals in complete market models. To appear in Math. Oper. Research.

[182] Schmeidler, D., Integral representation without additivity. Proc. Amer. Math. Soc. 97 (2) (1986), 255–261.

[183] Schweizer, M., Hedging of options in a general semimartingale model. Diss. ETH Zürich No. 8615 (1988).

[184] Schweizer, M., Risk-minimality and orthogonality of martingales. Stochastics Stochastics Rep. 30 (1990), 123–131.

[185] Schweizer, M., Variance-optimal hedging in discrete time. Math. Oper. Res. 20 (1995) 1–32.

[186] Schweizer, M., A Guided Tour through Quadratic Hedging Approaches. In: Option pricing, interest rates, and risk management (Jouini, E., Musiela, M., Cvitanic, J., eds.). Cambridge University Press, Cambridge, 2001, 538–574.

[187] Shiryaev, A. N., Probability. 2nd ed. Grad. Texts in Math. 95, Springer-Verlag, New York, 1995.

[188] Shiryaev, A. N., Essentials of stochastic finance. Adv. Ser. Stat. Sci. Appl. Probab. 3, World Scientific, Singapore, 1999.

[189] Snell, L., Applications of the martingale systems theorem. Trans. Amer. Math. Soc. 73 (1952), 293–312.

[190] Sprenkle, C., Warrant prices as indicators of expectations and preferences. Yale Economic Essays 1 (1964), 178–231.

[191] Strassen, V., The existence of probability measures with given marginals. Ann. Math. Stat. 36 (1965), 423–439.

[192] Staum, J., Fundamental theorems of asset pricing for good deal bounds. Math. Finance 14 (2) (2004), 141–161.

[193] Stricker, C., Arbitrage et lois de martingale. Ann. Inst. H. Poincaré Probab. Statist. 26 (1990), 451–460.

[194] Tversky, A., Kahneman, D., Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Uncertainty 5 (1992), 297–323.

[195] von Weizsäcker, H., Can one drop L^1-boundedness in Komlós' Subsequence Theorem? American Math. Monthly 111 (2004), 900–903.

[196] Wang, S., Dhaene, J., Comonotonicity, correlation order and premium principles. Insurance Math. Econom. 22 (1998), 235–242.

[197] Weber, S., Distribution-invariant dynamic risk measures. Preprint, Humboldt-Universität zu Berlin (2004).

[198] Yaari, M., The dual theory of choice under risk. Econometrica 55 (1987), 95–116.

[199] Yan, J.-A., Caractérisation d'une classe d'ensembles convexes de L^1 ou H^1. Seminaire de probabilités XIV, 1978/79 (J. Azéma and M. Yor, eds.), Lecture Notes in Math. 784, Springer-Verlag, Berlin, 1980, 220–222.

List of symbols

Index

This book is an introduction to financial mathematics. It is intended for graduate students in mathematics and for researchers working in academia and industry.

The focus on stochastic models in discrete time has two immediate benefits. First, the probabilistic machinery is simpler, and one can discuss right away some of the key problems in the theory of pricing and hedging of financial derivatives. Second, the paradigm of a complete financial market, where all derivatives admit a perfect hedge, becomes the exception rather than the rule. Thus, the need to confront the intrinsic risks arising from market incompleteness appears at a very early stage.

The first part of the book contains a study of a simple one-period model, which also serves as a building block for later developments. Topics include the characterization of arbitrage-free markets, preferences on asset profiles, an introduction to equilibrium analysis, and monetary measures of financial risk.

In the second part, the idea of dynamic hedging of contingent claims is developed in a multiperiod framework. Topics include martingale measures, pricing formulas for derivatives, American options, superhedging, and hedging strategies with minimal shortfall risk.

In this second edition major parts have been improved or entirely rewritten. Among them are those on robust representations of risk measures, arbitrage-free pricing of contingent claims, exotic derivatives in the CRR model, convergence to Black–Scholes prices, and stability under pasting with its connections to dynamically consistent coherent risk measures. Moreover, new sections have been added, including a systematic discussion of law-invariant risk measures, of concave distortions, and of the relations between risk measures and Choquet integration.

9 783110 183467

ISBN 3 11 018346 3
www.deGruyter.com