

HANDBOOKS IN MATHEMATICAL FINANCE

Option Pricing, Interest Rates and Risk Management

Edited by
Elyès Jouini
Jakša Cvitanić &
Marek Musiela

CAMBRIDGE

Option Pricing, Interest Rates and Risk Management

This handbook presents the current state of practice, method and understanding in the field of mathematical finance. Every chapter has been written by leading researchers and each starts by briefly surveying the existing results for a given topic, then discusses more recent results and, finally, points out open problems with an indication of what needs to be done in order to solve them. The primary audiences for the book are doctoral students, researchers and practitioners who already have some basic knowledge of mathematical finance. In sum, this is a comprehensive reference work for mathematical finance and will be indispensable to readers who need to find a quick introduction or reference to a specific topic, leading all the way to cutting edge material.

HANDBOOKS IN MATHEMATICAL FINANCE

*Option Pricing, Interest Rates
and Risk Management*

Edited by

E. Jouini

Université Paris – Dauphine and CREST

J. Cvitanić

University of Southern California

Marek Musiela

Paribas, London



PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014, Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

<http://www.cambridge.org>

© Cambridge University Press 2001

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2001
Reprinted 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 11/14pt. *System* L^AT_EX 2_ε [DBD]

A catalogue record of this book is available from the British Library

Library of Congress Cataloguing in Publication Data

Advances in mathematical finance / edited by E. Jouini, J. Cvitanić, Marek Musiela.

p. cm.

Includes bibliographic references and index.

ISBN 0 521 79237 1

1. Derivatives securities—Prices—Mathematical models.
2. Interest rates—Mathematical models. 3. Risk management.
4. Securities—Mathematical models. I. Jouini, E. (Elyès), 1965–
II. Cvitanić, J. (Jaksa), 1962– III. Musiela, Marek, 1950–

HG6024.A3 A38 2001

332'.01'51–dc21 00-052911

ISBN 0 521 79237 1 hardback

Contents

<i>List of Contributors</i>	page vii
<i>Introduction</i>	ix
Part one: Option Pricing: Theory and Practice	1
1 Arbitrage Theory <i>Yu. M. Kabanov</i>	3
2 Market Models with Frictions: Arbitrage and Pricing Issues <i>E. Jouini and C. Napp</i>	43
3 American Options: Symmetry Properties <i>J. Detemple</i>	67
4 Purely Discontinuous Asset Price Processes <i>D. B. Madan</i>	105
5 Latent Variable Models for Stochastic Discount Factors <i>R. Garcia and É. Renault</i>	154
6 Monte Carlo Methods for Security Pricing <i>P. Boyle, M. Broadie and P. Glasserman</i>	185
Part two: Interest Rate Modeling	239
7 A Geometric View of Interest Rate Theory <i>T. Björk</i>	241
8 Towards a Central Interest Rate Model <i>A. Brace, T. Dun and G. Barton</i>	278
9 Infinite Dimensional Diffusions, Kolmogorov Equations and Interest Rate Models <i>B. Goldys and M. Musiela</i>	314
10 Modelling of Forward Libor and Swap Rates <i>M. Rutkowski</i>	336
Part three: Risk Management and Hedging	397
11 Credit Risk Modelling: Intensity Based Approach <i>T. R. Bielecki and M. Rutkowski</i>	399
12 Towards a Theory of Volatility Trading <i>P. Carr and D. Madan</i>	458
13 Shortfall Risk in Long-Term Hedging with Short-Term Futures Contracts <i>P. Glasserman</i>	477
14 Numerical Comparison of Local Risk-Minimisation and Mean-Variance Hedging <i>D. Heath, E. Platen and M. Schweizer</i>	509

15	A Guided Tour through Quadratic Hedging Approaches	<i>M. Schweizer</i>	538
Part four: Utility Maximization			575
16	Theory of Portfolio Optimization in Markets with Frictions	<i>J. Cvitanić</i>	577
17	Bayesian Adaptive Portfolio Optimization	<i>I. Karatzas and X. Zhao</i>	632

Contributors

- G. Barton, Department of Chemical Engineering, University of Sydney, Sydney, Australia.
- T. Bielecki, Department of Mathematics, The Northeastern Illinois University, Chicago, USA.
- T. Björk, Department of Finance, Stockholm School of Economics, Box 6501, S-11383 Stockholm, Sweden.
- P. Boyle, School of Accountancy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.
- Alan Brace, FMMA and NAB, PO Box 731, Grosvenor Place, Sydney 2000, Australia.
- M. Broadie, Graduate School of Business, Columbia University, New York, NY 10027, USA.
- P. Carr, Morgan Stanley, 1585 Broadway, 6th floor, New York, NY 10036, USA.
- J. Cvitanić, Department of Mathematics, University of Southern California, 1042 West 36th Place, Los Angeles, CA 90089-1113, USA.
- J. Detemple, School of Management, Boston University, 595 Commonwealth Avenue, Boston, MA 02215, USA.
- T. Dun, Department of Chemical Engineering, University of Sydney, Sydney, Australia.
- R. Garcia, Département de Sciences Économiques, Université de Montréal, Montréal (PQ) H3C 3J7, Canada.
- P. Glasserman, Columbia Business School, Columbia University, New York, NY 10027, USA.
- B. Goldys, School of Mathematics, University of New South Wales, Sydney, 2052 NSW, Australia.
- D. Heath, University of Technology, Sydney, School of Finance & Economics, PO Box 123, Broadway, 2007 NSW, Australia.
- E. Jouini, Université Paris IX Dauphine, CEREMADE, Place du Maréchal de Lattre de Tassigny, 75775 Paris, Cedex 16, France.
- Yu. M. Kabanov, Laboratoire de Mathématiques, Université de Franche-Comté, 16 Route de Gray, F-25030 Besançon, Cedex, France.
- I. Karatzas, Departments of Mathematics and Statistics, Columbia University, New York, NY 10027, USA.
- D. Madan, College of Business and Management, University of Maryland, College Park, MD 20742, USA.

- M. Musiela, Paribas, 10 Harewood Avenue, London NW1 6AA, UK.
- C. Napp, Université Paris IX Dauphine, CEREMADE, Place du Maréchal de Lattre de Tassigny, 75775 Paris, Cedex 16, France.
- E. Platen, University of Technology, Sydney, School of Finance & Economics, PO Box 123, Broadway, 2007 NSW, Australia.
- E. Renault, Département de Sciences Économiques, Université de Montréal, Montréal (PQ) H3C 3J7, Canada.
- M. Rutkowski, Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warsaw, Poland.
- M. Schweizer, Technische Universität Berlin, Fachbereich Mathematik, Strasse des 17. Juni 136, D-10623, Berlin, Germany.
- X. Zhao, Departments of Mathematics and Statistics, Columbia University, New York, NY 10027, USA.

Introduction

This book, the final in a series of stand-alone works, is a collection of invited papers that represent the current state of research in the field of Mathematical Finance, as seen by leading researchers in the field. Some of the contributed articles survey the existing results for a given topic, some discuss and present new research, some point out open problems and future directions, while many do all of the above. While effort was made to cover most of the important topics in the field, the book is not meant to be encyclopedic in nature. The outcome was ultimately influenced by the present scientific interest of the contributors and the editors. The primary audience are researchers in academia and industry who already have some basic knowledge of the field. This book might serve as a quick introduction to a specific topic, leading to recent results and open problems. It can also serve as valuable reference material.

The first Part focuses on the theory and practice of pricing derivative securities. The paper “Arbitrage theory” by Y. Kabanov considers models where an investor, acting on a financial market with random price movements and having a given time horizon, subsequently transforms his initial endowment into a certain terminal wealth. In this framework, the author answers the following question: whether the investor has arbitrage opportunities, i.e. non-risky profits. The article examines and gives an answer to this question in different frameworks: one-step and multi-step models with finite space of possible states of the world, discrete-time models with infinite space of possible states of the world, continuous time models, semimartingale models, large financial markets and models with transaction costs. The article “Market models with frictions: arbitrage and pricing issues” by E. Jouini and C. Napp extends the previous results in two directions: first, they consider investment opportunities determined by their cash-flows instead of financial assets described by their price processes. This approach enables them to take into account classical market models as well as investment models. Second, the authors consider a wide range of possible market imperfections: transaction

costs, borrowing costs and constraints, short-selling costs and constraints, fixed and proportional transaction costs and models with defaultable numéraire. In all these cases, they characterize the no-arbitrage assumption through a unified approach and they apply these results to pricing and hedging issues.

The contribution by J. Detemple “American options: symmetry properties” surveys generalizations of the classical put–call symmetry: the value of a put option with strike price K on an underlying asset S paying dividends at rate δ in a financial market with riskless interest rate r is the same as the value of a call option with strike price S on an asset paying dividends at rate r and having initial value K , in an auxiliary financial market with interest rate δ . It is shown that the symmetry holds in a large class of models, including nonmarkovian markets with random coefficients, and even for many nonstandard American claims including barrier options, multi-asset derivatives, and occupation time derivatives. The main tool, change of numéraire technique, is also reviewed and extended to the case of dividend-paying assets. The put–call symmetry reduces the computational burden in pricing options; it provides useful insights into the economic relationship between contracts, and sometimes even helps to reduce the dimensionality of the problem, thereby making somewhat more tractable the difficult problem of evaluating American contingent claims.

The article “Monte Carlo methods for security pricing” by P. Boyle, M. Broadie and P. Glasserman, reprinted from *Journal of Economic Dynamics and Control*, is a detailed survey of simulation methods applied to numerical pricing of European, and, more recently, American options. Since European option prices can be calculated as expected values, it is natural to use Monte Carlo for computing them. However, this can often be quite slow, and this paper reviews and compares different methods used to improve the efficiency of Monte Carlo methods. So-called “variance reduction” techniques are surveyed, including control variates, antithetic variates, moment matching, importance sampling and conditional Monte Carlo methods. Next, the quasi-Monte Carlo approach is reviewed, in which, instead of random numbers, deterministic sequences are generated – so-called quasi-random numbers or low-discrepancy sequences. These are more evenly dispersed than random sequences. It is interesting that these procedures are typically based on number-theoretic methods. The paper also discusses the use of Monte Carlo methods for computing sensitivities (“Greeks”) of the option price with respect to different parameters, and the difficult problem of computing American option prices using simulation. The difficulty stems from the fact that the price of an American option is a maximum of expected values, rather than a single expected value.

In their chapter, R. Garcia and E. Renault use the concept of stochastic discount factor (SDF) or pricing kernel as a unifying principle to integrate two concepts

of latent variables, one cross-sectional, one longitudinal, in order to reduce the dimension of a statistical model specified for a multivariate time series of asset prices. In the CAPM or APT beta pricing models, the dimension reduction is cross-sectional in nature, while in time-series state-space models, dimension is reduced longitudinally by assuming conditional independence between consecutive returns given a small number of state variables. They provide this unifying analysis in the context of conditional equilibrium beta pricing as well as asset pricing with stochastic volatility, stochastic interest rates and other state variables. They address the general issue of econometric specifications of dynamic asset pricing models, which cover the modern literature on conditionally heteroskedastic factor models as well as equilibrium-based asset pricing models with an intertemporal specification of preferences and market fundamentals.

D. Madan, in his contribution “Purely discontinuous asset price processes” surveys his work with various co-authors on modeling asset prices with pure jump processes, and on pricing contingent claims in such models. It is argued that statistical analysis leads to the consideration of discontinuous asset prices models, in which the arrival rate of jumps is infinite and decreasing in the jump size. Such models are also motivated by theoretical no-arbitrage considerations, implying that the prices must be modeled as time-changed Brownian motion. If, as is argued, this time change has to be modeled as random, we are led to the class of discontinuous price processes. Being of bounded variation, these prices are also more robust relative to change of parameters than the typical diffusion models. The example of the so-called variance gamma process is presented in detail, including solutions to option pricing and optimal investment problems in such a market model. Using these solutions, the model is calibrated, which is in turn used to infer trader preferences and personalized risk neutral measures, called position measures. The paper is representative of a very active field of research, rich in theoretical and practical implications.

Part II presents different aspects of the theory and practice of interest rate modeling. Arbitrage-free movement of the forward curve is analyzed from the perspective of infinite dimensional diffusions by T. Björk in his article “A geometric view of interest rate theory”. He addresses the following questions: when is a given forward rate model consistent with a given family of forward rate curves and when can the inherently infinite dimensional forward rate process be realized by means of a finite-dimensional state space model? Necessary and sufficient conditions for consistency as well as for the existence of finite-dimensional realizations are given in terms of forward rate volatilities. That is, the forward rate model generated by a collection of volatility functions admits a finite dimensional realization if and only if the corresponding Lie algebra generated by the volatility functions and the

drift (which is also uniquely determined from the volatility functions by arbitrage considerations) is finite-dimensional in the neighbourhood of the initial condition. General consistency results are not given in this chapter, though references are made to the recent papers and the PhD thesis by D. Filipovic. Instead, the author concentrates on analysis of the Nelson–Siegel (NS) family of forward curves. It turns out that neither the Hull–White (HW) nor the Ho–Lee (HL) model is consistent with the NS family. In fact the NS manifold is too small for the HW and HL models, in the sense that if the initial curve is on the manifold, then the models will force the term structure off the manifold within an arbitrarily short period of time.

The infinite-dimensional approach is also taken in the chapter: “Infinite dimensional diffusions, Kolmogorov equations and interest rate models” by B. Goldys and M. Musiela. The main emphasis is put on differential analysis in infinite dimension. Motivation comes from the need for a better understanding of interest rate risk management issues. To be more precise let us look first at the Black–Scholes model. The lognormal diffusion process generating arbitrage free evolution of the variable of interest can also be represented by corresponding it with an infinitesimal generator. Pricing of options is identical to solving the related Kolmogorov equation. Sensitivity to the change in the stochastic variable is done by simple differentiation of the price. The situation in the interest rate area is more complex. The underlying stochastic variable is the entire forward curve. The diffusion process defining the evolution of the forward curve is infinite-dimensional. The infinitesimal generator and the corresponding Kolmogorov equation need to be defined and studied from the perspective of the sensitivity of an interest rate option to the changes in the shape of the forward curve. It turns out that one can obtain Feynman–Kac representations of solutions to such equations for a large class of terminal conditions (which include most of the treated products) and that for those the price is differentiable with respect to the initial forward curve. This is in contrast with poor smoothing properties of the associated semigroup and the fact that not all the payoffs have discounted expected values which are Fréchet differentiable. While continuous compounding associated with the continuous tenor models may ultimately lead to more unified infinite-dimensional theories of the forward curve dynamics, at the implementation level one is almost forced to work with models allowing for finite-dimensional realizations. On the other hand, simple compounding corresponding to a given discrete tenor structure has the advantage of being grounded on standard finite-dimensional semimartingale theory, which is better understood and more developed. Additionally, it represents the interest rate markets more realistically. As such, it is arguably better suited for the pricing of most Libor and swap derivatives. The canonical forward Libor and swap rate models with deterministic volatilities are by construction

finite-dimensional diffusions under any of the Libor measures (spot or forward). The explicit relationships between the measures allow for the development of exact expressions or at least of good analytic approximations to a number of options such as caps and swaptions. The chapter: “Modelling of forward Libor and swap rates” by M. Rutkowski presents an overview of recently developed methodologies related to the derivation and analysis of the arbitrage free dynamics of such market rates. The article: “Towards a central interest rate model” by A. Brace, T. Dun and G. Barton aims to expose issues related with implementation of the canonical lognormal forward Libor model. The pricing of swaptions is examined within this framework and compared to the industry standard Black swaption formula, and, by extension, to the lognormal swap rate model. Swap and swaption behaviour are investigated under arbitrary volatility and yield curve specifications. Simulation and approximation techniques are used to make comparisons in terms of observed swap rate probability distributions, swaption volatilities and prices, and swaption sensitivities defined in terms of the swap rate. Fifteen swaptions and two volatility structures are considered. Swap rates simulated under the lognormal Libor model are shown to be statistically lognormal in each case, and volatilities, prices and Greeks agree closely. Finally, the approximate delta value within the lognormal Libor model is used in a simulated delta-hedging exercise and is seen to successfully hedge Libor model swaptions. This points to the robustness of the lognormal Libor model for the following two reasons. Firstly, the exact delta of a swaption, in a lognormal Libor model, is, in fact, the vector of partial derivatives of the swaption price with respect to the underlying forward Libor rates. Secondly, the volatility of the forward swap rate under the corresponding forward swap rate measure, in the lognormal Libor model, is stochastic. Overall, in the authors’ opinion, the forward Libor model is the unifying model capable of encompassing the properties of the swap rate model and allowing for greater aggregation of risk in portfolios containing Libor and swap derivatives.

The third Part considers different types of risk in financial markets, and ways to manage and hedge exposure to risk. “Credit risk modelling: an intensity based approach” by T. Bielecki and M. Rutkowski reviews fundamental methodologies and results in the area of the intensity based default and credit risk modeling. Special care is devoted to the technical issues of the role of conditioning information in computations involving random times. The time of default is modeled via a jump process with positive jump intensity. An overview of credit-risk instruments is provided, together with market methods for pricing them. Next, the basic theory of valuation of defaultable claims is presented, and various specifications for modeling recovery value at or after the time of default are discussed. Moreover, models that account for the migration between credit-rating grades are surveyed, both in discrete-time and continuous-time. A credit-spread based HJM-type model

is presented, in which default-free and defaultable term structure is modeled. Finally, the theory is applied to the problem of valuation of some common credit derivatives.

The area of credit and default risk has been very active and popular in recent years, both in financial industry practice and in academic research. The primary purpose of the article: “Towards a theory of volatility trading” by P. Carr and D. Madan is to review three methods which recently emerged for trading realized volatility. The first method involves taking static position in options. The classic example is that of a log position in a straddle. The second method involves delta-hedging of an option. If an investor is successful in hedging away the price risk, then a prime determinant of the profit or loss from this strategy is the difference between the realized volatility and the anticipated volatility used in pricing and hedging the option. The final method reviewed for trading realized volatility involves buying or selling an over-the-counter contract whose payoff is an explicit function of volatility. The simplest example of such a volatility contract is a volatility swap. This contract pays the buyer the difference between the realized volatility and a level of volatility fixed at the outset of the contract. A secondary purpose is to uncover the link between volatility contracts and some recent ground-breaking work by Dupire and Derman, Kani, and Kamal. By restricting the set of times and price levels for which returns are used in volatility calculations, one can synthesize a contract which pays off the “local volatility”.

The contribution by P. Glasserman, “Shortfall risk in long-term hedging with short-term futures contracts” proposes and analyzes a measure of the risk of a cash shortfall in hedging a risky position over time. The measure is illustrated by comparing various hedging strategies for firm hedging a long-term commitment with short-dated future contracts. It is motivated by the infamous case of derivatives losses suffered by Metallgesellschaft Refining and Marketing. The firm had entered into long-term contracts to supply oil at fixed prices, and was hedging these commitments with short-term future contracts. While the strategy would have produced, at least theoretically, a perfect hedge at the end of the long-term contract, it led to a severe cash shortfall during the life of the contract. In a Gaussian model the theory of Gaussian extremes and large deviation approximations are used to calculate this measure, to capture qualitative features of the shortfall risk and to identify the most likely path to a shortfall under different hedging strategies. A brief summary of concepts pertinent to futures and forwards is provided in an appendix. The theory for analyzing liquidity risks is only in its infancy, and this paper indicates some possible ways for making progress in developing it.

M. Schweizer’s contribution “A guided tour through quadratic hedging approaches” gives an overview of the general theory of pricing and hedging contingent claims in incomplete markets by means of a quadratic criterion. It is

based on numerous papers by the author and his co-workers. It is an example of an abstract theory developed for very practical problems, since many models used in practice are, indeed, incomplete. The paper explains the notions of local risk-minimization, the minimal martingale measure, the variance-optimal martingale measure, mean-variance hedging, Föllmer–Schweizer decomposition, and so on. It first discusses the case in which the hedging strategies are not required to be self-financing. If the discounted price process is a local martingale, one can find a risk-minimizing strategy, which is also mean self-financing. In the general case, one can only find so-called locally risk-minimizing strategies. In the last part of the article, the mean-variance criterion is considered for those strategies that are required to be self-financing, and the connection to closedness properties of spaces of stochastic integrals is studied. Despite the significant progress that has been made on these problems over the years, and the success of complete characterization of solutions in special cases, in general, questions about how to actually construct optimal strategies remain open, and the search for those solutions is still ongoing.

The companion chapter “Numerical comparison of local risk-minimization and mean-variance hedging” by D. Heath, E. Platen and M. Schweizer focuses on the more practical aspects of the two criteria. It begins with the concrete situation of a Markovian stochastic volatility setting and there provides general comparative results on prices, hedging strategies and risks for local risk-minimization versus mean-variance hedging. A detailed analysis including numerical results is then performed for the well-known Heston and Stein/Stein stochastic volatility models. The results highlight some important quantitative differences between the two approaches and give some directions for future research.

Part IV contains papers on the optimal portfolio selection problem. The article “Theory of portfolio optimization in markets with frictions” by one of the editors (J.C.) surveys results on extending the classical Merton’s utility maximization problem in continuous-time models driven by Brownian motion, to the case of markets which are incomplete due to the presence of portfolio constraints, transaction costs, different borrowing and lending rates, and so on. The methodology employed is to first characterize the minimal cost of super-replicating a given claim in such markets, and then solve an optimization problem dual to the utility maximization problem. If the dual problem is appropriately defined, it can then be shown, using the results on super-replication, that the optimal strategy can be characterized in terms of the solution to the dual problem. Explicit results are available for many examples in the case of portfolio constraints and different borrowing and lending rates, but not in the case of transaction costs. In terms of open problems, as far as the general theory is concerned, some of these

results have not yet been fully extended to general arbitrage-free semimartingale models.

“Bayesian adaptive portfolio optimization” by I. Karatzas and X. Zhao also considers the portfolio optimization problem, but in the framework of the stock return rates being unobserved by the investor. Instead, they are modeled in a Bayesian fashion, as a random vector with a known probability distribution. The investor is assumed to observe past and present stock prices, and has to base investment decisions only on that information. The value function is obtained using both filtering/martingale and stochastic control/partial differential equation techniques. The former approach transforms the problem into one with the drift process adapted to the observation process, while the latter approach is used to show that the Hamilton–Jacobi–Bellman equation for this problem takes the form of a generalized Monge–Ampère equation, which is solved fairly explicitly. Next, it is shown that, for the logarithmic utility function, the cost of uncertainty about the unknown drift of the stock prices (relative to an investor who can observe the drift) is asymptotically negligible. The results are also extended to the case of portfolio constraints. The article is a contribution to the very lively line of research in financial economics and mathematics dealing with problems of incomplete or asymmetric information.

The editors would like to express their gratitude to the individuals who made the book possible. Thanks are above all due to all the contributors – they have worked with us with enthusiasm and efficiency, making the editorial job truly enjoyable. The project would not have been possible without the immense efforts, support and vision of David Tranah of Cambridge University Press. We are sincerely grateful for his high professionalism and constant encouragement. We are also thankful to Elsevier, for permitting us to reprint the paper by Boyle, Broadie and Glasserman in this book.

J.C., E.J. and M.M.

Part one

Option Pricing: Theory and Practice

1

Arbitrage Theory

Yu. M. Kabanov

1 Introduction

We shall consider models where an investor, acting on a financial market with random price movements and having T as his time horizon, transforms the initial endowment ξ into a certain resulting wealth; let R_T^ξ denote the set of all final wealth corresponding to possible investment strategies. The natural question is, whether the investor has arbitrage opportunities, i.e. whether he can get non-risky profits.

Let us “hide” in a “black box” the interior dynamics on the time-interval $[0, T]$ (i.e. the price process specification, market regulations, description of admissible strategies) and examine only the set R_T^ξ .

At this level of generality, the answer, as well as the hypotheses, should be formulated only in terms of properties of the sets R_T^ξ . E.g., in the simplest situation of frictionless market without constraints, R_T^0 is a linear subspace in the space L^0 of (scalar) random variables and $R_T^\xi = \xi + R_T^0$. The absence of arbitrage opportunities can be formalized by saying that the intersection of R_T^0 with the set L_+^0 of non-negative random variables contains only zero. If the underlying probability space is finite, i.e. if we assume in our model only a finite number of states of the nature, it is easy to prove that there is no arbitrage if and only if there exists an equivalent “separating” probability measure with respect to which every element of R_T^0 has zero mean. Close look at this result shows that this assertion is nothing but the Stiemke lemma [62] of 1915 which is well-known in the theory of linear inequalities and linear programming as an example of the so-called alternative (or transposition) theorems, see historical comments in [61]; notice that the earliest alternative theorem due to Gordan [21] (of 1873) can be also interpreted as a no-arbitrage criterion.

The one-step model can be generalized (or specialized, depending on the point of view) in many directions giving rise to what is called arbitrage theory. The reader should not be confused by using “general” and “special” in this context: obviously,

one-step models are particular cases of N -period models, but quite often the main difficulties in the analysis of models with a detailed (“specialized”) structure of the “black box” consist in verifying hypotheses of theorems corresponding to the one-step case. The geometric essence of these results is a separation of convex sets with a subsequent identification of the separating functional as a probability measure; the properties of the latter in connection with the price process are of particular interest.

To this date one can find in the literature dozens of models of financial markets together with a plethora of definitions of arbitrage opportunities. These models can be classified using the following scheme.

1.1 Finite probability space

Assuming only a finite number of states of the nature is popular in the literature on economics. Of course, the hypothesis is not adequate to the basic paradigm of stochastic modeling because random variables with continuous distributions cannot “live” on finite probability spaces. The advantage of working under this assumption is that a very restricted set of mathematical tools (basically, elementary finite-dimensional geometry) is required. Results obtained in this simplified setting have an important educational value and quite often may serve as the starting point for a deeper development.

1.2 General probability space

In contrast to the case of finite probability space, the straightforward separation arguments, which are the main instruments to obtain no-arbitrage criteria, fail to be applied without further topological assumptions on R_T^0 . In many particular cases, especially in the theory of continuous trading, they are not fulfilled. This circumstance led Kreps (1981) to a more sophisticated “no-arbitrage” concept, namely, that of “no free lunch” (NFL). However, certain no-arbitrage criteria are of the same form as for the models with finite probability space Ω .

1.3 Discrete-time multi-period models

Even for the case of finite probability space Ω , these models are important because they allow us to describe the intertemporal behavior of investors in financial markets, i.e. to penetrate into the structure of the “black box” using concepts of random processes. One of the most interesting features is that in the simplest model without constraints the value processes of the investor’s portfolios are martingales with respect to separating measures and the same property holds for the underlying

price process; this explains the terminology “equivalent martingale measures”. Models based on the infinite Ω posed challenging mathematical questions, e.g., whether the absence of arbitrage is still equivalent to the existence of equivalent martingale measure. For a frictionless market the affirmative answer has been given by Dalang, Morton, and Willinger (1990). Their work, together with the earlier paper of Kreps, stimulated further research in geometric functional analysis and stochastic calculus, involving rather advanced mathematics.

1.4 Continuous trading

Although the continuous-time stochastic processes were used for modeling from the very beginning of mathematical finance (one can say that they were even invented exactly for this purpose, having in mind the Bachelier thesis “Théorie de la spéculation” where Brownian motion appeared for the first time), their “golden age” began in 1973 when the famous Black–Scholes formula was published. Subsequent studies revealed the role of the uniqueness of the equivalent martingale measure for pricing of derivative securities via replication. The importance of no-arbitrage criteria seems to be overestimated in financial literature: the unfortunate alias FTAP – Fundamental Theorem of Asset (or Arbitrage) Pricing, ambitious and misleading, is still widely used. If there are many equivalent martingale measures, the idea of “pricing by replication” fails: a contingent claim may not belong to R_T^x whatever x is, or may belong to many R_T^x . In the latter case it is not clear which martingale measure can be used for pricing and this is the central problem of current studies on incomplete markets. However, as to mathematics, the no-arbitrage criteria for general semimartingale models are considered among the top achievements of the theory.

In 1980 Harrison and Pliska noticed that stochastic calculus, i.e. the integration theory for semimartingales, developed by P.-A. Meyer in a purely abstract way, is “tailor-made” for financial modeling. In 1994 Delbaen and Schachermayer confirmed this conclusion by proving that the absence of arbitrage in the class of elementary, “practically admissible” strategies implies the semimartingale property of the price process. In a series of papers they provided a profound analysis of the various concepts culminating in a result that the Kreps NFL condition (equivalent to a whole series of properties with easier economic interpretation) holds if and only if the price process is a σ -martingale under some $\tilde{P} \sim P$. There is another justification of the increasing interest in semimartingales in financial modeling: mathematical statistics sends alarming signals that in many cases empirical data for financial time series are not compatible with the hypothesis that they are generated by processes with continuous sample paths. Thus, diffusions should be viewed

only as strongly stylized models of financial data; it has been revealed that Lévy processes give much better fit.

1.5 Large financial markets

This particular group, including the so-called Arbitrage Pricing Model (or Theory), abbreviated to APM (or APT), due to Ross and Huberman (for the one-period case), has the following specific feature. In contrast with the conventional approach of describing a security market by a single probabilistic model, a sequence of stochastic bases with an increasing but always finite number of assets is considered. One can think that the agent wants to concentrate his activity on smaller portfolios because of his physical limitations but larger portfolios in this market may have better performance. The arbitrage is understood in an asymptotic sense. Its absence implies relationships between model parameters which can be verified empirically. This circumstance makes such models especially attractive. The weak side of APM is the use of the quadratic risk measure. This means that gains are punished together with losses in symmetric ways which is unrealistic. Luckily, the conclusion of APM, the Ross–Huberman boundedness condition, seems to be sufficiently “robust” with respect to the risk measure and the variation of certain model parameters.

In the recent papers [36] and [37], where the theory of large financial markets was extended to the general semimartingale framework, the concept of asymptotic arbitrage is developed for an “absolutely” risk-averse agent. In spite of a completely different approach, the absence of asymptotic arbitrage implies, for various particular models, relations similar to the Ross–Huberman condition.

1.6 Models with transaction costs

In the majority of models discussed in mathematical finance, the investor’s wealth is scalar, i.e. all positions are measured in units of a single asset (money, bond, bank account, etc.). However, in certain cases, e.g., in models with constraints and, especially, in those taking transaction costs into account, it is quite natural to consider, as the primary object, the whole vector-valued process of current positions, either in physical quantities or in units of values measured by a certain numéraire. It happens that this approach allows not only for a more detailed and realistic description of the portfolio dynamics but also opens new perspectives for further mathematical development, in particular, for an extensive use of ideas from theory of partially ordered spaces, utility theory, optimal control, and mathematical economics. Until now only a few results are available in this new branch of arbitrage theory. Recent studies [34] and [41] show that the basic concept of

arbitrage theory, that of the equivalent martingale measure, should be modified and generalized in an appropriate way. There are various approaches to the problem which will be discussed here. Notice that models with transaction costs quite often were considered as completely different from those of a frictionless market and the classical results could not be obtained as corollaries when transaction costs vanish. The modern trend in the theory is to work in the framework which covers the latter as a special case.

Arbitrage theory includes another, even more important subject, namely, hedging theorems, closely related with the no-arbitrage criteria. These results, discussed in the present survey in a sketchy way, give answers to whether a contingent claim can be replicated in an appropriate sense by a terminal value of a self-financing portfolio or whether a given initial endowment is sufficient to start a portfolio replicating the contingent claim. Other related problems such as market completeness or models with continuum securities, arising in the theory of bond markets, are not touched here.

The books [52], [57], and [29] may serve as references in convex analysis, probability, and stochastic calculus.

2 Discrete-time models

2.1 General setting

Let $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t), P)$ be a stochastic basis (i.e. filtered probability space), $t = 0, 1, \dots, T$. We assume that each σ -algebra $\mathcal{F}_t$ is complete.

We are given:

- convex cones $R_t^0 \subseteq L^0(\mathbf{R}^d, \mathcal{F}_t)$;
- closed convex cones $\mathcal{K}_t \subseteq L^0(\mathbf{R}^d, \mathcal{F}_t)$.

The notation $L^0(K_t, \mathcal{F}_t)$ is used for the set of all $\mathcal{F}_t$ -measurable random variables with values in the set K_t (or $\mathcal{F}_t$ -measurable selectors of K_t if K_t depends on ω).

The usual financial interpretation: R_t^0 is the set of portfolio values at the date t corresponding to the zero initial endowment, i.e. all imaginable results that can be obtained by the investor to the date t .

The cones $\mathcal{K}_t$ induce the partial orderings in the sets $L^0(\mathbf{R}^d, \mathcal{F}_t)$:

$$\xi \geq_t \eta \quad \Leftrightarrow \quad \xi - \eta \in \mathcal{K}_t.$$

The partial orderings $\geq_t$ allow us to compare current results.

As a rule, they are obtained by “lifting” partial orderings from $\mathbf{R}^d$ to the space of random variables.

A typical example: $\mathcal{K}_t = L^0(K, \mathcal{F}_t)$ where K is a closed cone in $\mathbf{R}^d$ (which may depend on ω and t). In particular, the “standard” ordering $\geq_t$ is induced by $K_t = \mathbf{R}_+^d$ when $\xi \geq_t \eta$ if $\xi^i \geq \eta^i$ (a.s.) for all $i \leq d$; for the case $d = 1$ it is the usual linear ordering of the real line. However, we do not exclude other partial orderings.

In the theory of frictionless market, usually, $d = 1$; for models with transaction costs d is the number of assets in the portfolio.

We define also the set $A_T^0 := R_T^0 - \mathcal{K}_T$. The elements of A_T^0 are interpreted as contingent claims which can be hedged (or super-replicated) by the terminal values of portfolios starting from zero.

The linear space $\mathcal{L}_T := \mathcal{K}_T \cap (-\mathcal{K}_T)$ describes the positions ξ such that $\xi \geq_T 0$ and $\xi \leq_T 0$, which are “financially equivalent to zero”. The comparison of results can be done modulo this equivalence, i.e. in the quotient space $L^0/\mathcal{L}_T$ equipped with the ordering induced by the proper cone $\tilde{\mathcal{K}}_T := \pi_T \mathcal{K}_T$ where $\pi_T : L^0 \rightarrow L^0/\mathcal{L}_T$ is the natural projection.

2.2 No-arbitrage criteria for finite Ω

The most intuitive formulation of the property that the market has no arbitrage opportunities for the investors without initial capital is the following:

NA. $\mathcal{K}_T \cap R_T^0 \subseteq \mathcal{L}_T$.

In the particular case when $\mathcal{K}_T$ is a proper cone we have

NA'. $\mathcal{K}_T \cap R_T^0 \subseteq \{0\}$ (with equality if R_T^0 is closed).

The first no-arbitrage criterion has the following form.

Theorem 2.1 *Let Ω be finite. Assume that R_T^0 is closed. Then **NA** holds if and only if there exists $\eta \in L^0(\mathbf{R}^d, \mathcal{F}_T)$ such that*

$$E\eta\zeta > 0 \quad \forall \zeta \in \mathcal{K}_T \setminus \mathcal{L}_T$$

and

$$E\eta\zeta \leq 0 \quad \forall \zeta \in R_T^0.$$

Because L^0 is a finite-dimensional space, this result is a reformulation of Theorem A.2 on separation of convex cones.

It is easy to verify that $\mathcal{K}_T \cap R_T^0 \subseteq \mathcal{L}_T$ if and only if $\mathcal{K}_T \cap A_T^0 \subseteq \mathcal{L}_T$. Hence, in this theorem one can replace R_T^0 by A_T^0 .

The above criterion can be classified as a result for the one-step model where T stands for “terminal”. It has important corollaries for multi-period models where the sets R_T^0 have a particular structure.

3 Multi-step models

3.1 Notations

For $X = (X_t)_{t \geq 0}$ and $Y = (Y_t)_{t \geq 0}$ we define $X_- := (X_{t-1})$ (various conventions for X_{-1} can be used), $\Delta X_t := X_t - X_{t-1}$, and, at last,

$$X \cdot Y_t := \sum_{k=0}^t X_k \Delta Y_k,$$

for the discrete-time integral. Here X and Y can be scalar or vector-valued. In the latter case sometimes we shall use the abbreviation $X \bullet Y$ for the vector process formed by the pairwise integrals of the components

$$X \bullet Y := (X^1 \cdot Y^1, \dots, X^d \cdot Y^d).$$

Though in the discrete-time case the dynamics can be expressed exclusively in terms of differences, “integral” formulae are often instructive for continuous-time extensions.

For finite Ω , if X is a predictable process (i.e. X_t is $\mathcal{F}_{t-1}$ -measurable) and Y belongs to the space $\mathcal{M}$ of martingales, then $X \cdot Y$ is also a martingale.

The product formula

$$\Delta(XY) = X \Delta Y + Y_- \Delta X$$

is obvious.

3.2 Example 1. Model of frictionless market

The model being classical, we do not give details and financial interpretations: they are widely available in many textbooks.

Let $S = (S_t)$, $t = 0, 1, \dots, T$, be a fixed n -dimensional process adapted to a discrete-time filtration $\mathbf{F} = (\mathcal{F}_t)$. Here T is a finite integer and, for simplicity, the σ -algebra $\mathcal{F}_0$ assumed to be trivial. The convention $S_{-1} = S_0$ is used. Define R_T^0 as the linear space of all scalar random variables of the form $N \cdot S_T$ where N is an n -dimensional predictable process. For $x \in \mathbf{R}$ we put $R_T^x = x + R_T^0$. We take $\mathcal{K}_0 := \mathbf{R}_+$ and $\mathcal{K}_T := L^0(\mathbf{R}_+, \mathcal{F}_T)$.

The components S^i describe the price evolution of n risky securities, N^i is the portfolio strategy which is self-financing, and V is the value process. In this specification it is tacitly assumed that there is a traded asset with the constant unit price, i.e. this asset is the *numéraire*.

Remark 3.1 One should take care that there is another specification where the numéraire is not necessarily a traded asset. A possible confusion may arise because

the formula for the value process looks similar but the integrand and the integrator are in the latter case d -dimensional processes with $d = n + 1$. The increments of a self-financing portfolio strategy are explicitly constrained by the relation

$$S_{t-1} \Delta N_t = 0.$$

If the numéraire (“cash” or “bond”) is traded, the integral with respect to the latter vanishes but, of course, holdings in “cash” are not arbitrary but defined from the above relation.

For finite Ω we have, in virtue of Theorem 2.1, that the model has no-arbitrage if and only if there is a strictly positive random variable η such that $E\eta\zeta = 0$ for all $\zeta \in R_T^0$. Without loss of generality we may assume that $E\eta = 1$ and define the probability measure $\tilde{P} = \eta P$. Clearly, $\tilde{E}\zeta = 0$ for all $\zeta \in R_T^0$ (i.e. $\tilde{E}N \cdot S_T = 0$ for all predictable N) if and only if S is a martingale. With this remark we get the Harrison–Pliska theorem:

Theorem 3.2 *Assume that Ω is finite. Then the following conditions are equivalent:*

- (a) $R_T^0 \cap L^0(\mathbf{R}_+, \mathcal{F}_T) = \{0\}$ (no-arbitrage);
- (b) *there exists a measure $\tilde{P} \sim P$ such that $S \in \mathcal{M}(\tilde{P})$.*

Let $\rho_t := d\tilde{P}_t/dP_t$ be the density corresponding to the restrictions of $\tilde{P}$ and P to $\mathcal{F}_t$. Recall that the density process $\rho = (\rho_t)$ is a martingale $\rho_t = E(\rho_T|\mathcal{F}_t)$. Since

$$S \in \mathcal{M}(\tilde{P}) \iff S\rho \in \mathcal{M}(P),$$

we can add to the conditions of the above theorem the following one:

- (b') *there is a strictly positive martingale ρ such that $\rho S \in \mathcal{M}$.*

Notice that the equivalence of (b) and (b') is a general fact which holds for arbitrary Ω and even in the continuous-time setting.

Though the property (b') can be considered simply as a reformulation of (b), it is more adapted to various extensions. The advantage of (b) is in the interpretation of $\tilde{P}$ as a “risk-neutral” probability.

3.3 Example 2. Model with transaction costs

Now we describe a discrete-time version of a multi-currency model with proportional transaction costs introduced in [34] and studied in the papers [11] and [41].

It is assumed that the components of an adapted process $S = (S_t^1, \dots, S_t^d)$, $t = 0, 1, \dots, T$, describing the dynamics of prices of certain assets, e.g., currencies quoted in a certain reference asset (say, “euro”), are strictly positive. It is

convenient to choose the scales to have $S_0^i = 1$ for all i . We do not suppose that the numéraire is a traded security.

The transaction costs coefficients are given by an adapted process $\Lambda = (\lambda^{ij})$ taking values in the set $\mathbf{M}_+^d$ of non-negative $d \times d$ -matrices with zero diagonal.

The agent's portfolio at time t can be described either by a vector of “physical” quantities $\widehat{V}_t = (\widehat{V}_t^1, \dots, \widehat{V}_t^d)$ or by a vector $V = (V_t^1, \dots, V_t^d)$ of values invested in each asset. The relation

$$\widehat{V}_t^i = V_t^i / S_t^i, \quad i \leq d,$$

is obvious. Introducing the diagonal operator

$$\phi_t(\omega) : (x^1, \dots, x^d) \mapsto (x^1 / S_t^1(\omega), \dots, x^d / S_t^d(\omega)). \quad (1)$$

we may write that

$$\widehat{V}_t = \phi_t V_t.$$

The increments of portfolio values are

$$\Delta V_t^i = \widehat{V}_t^i \Delta S_t^i + b_t^i \quad (2)$$

with

$$b_t^i = \sum_{j=1}^d \alpha_t^{ji} - \sum_{j=1}^d (1 + \lambda^{ij}) \alpha_t^{ij},$$

where $\alpha_t^{ji} \in L^0(\mathbf{R}_+, \mathcal{F}_t)$ represents the net amount transferred from the position j to the position i at the date t .

The first term in the right-hand side of (2) is due to the price increment while the second corresponds to the agent's actions (made after the revealing of new prices). Notice that these actions are charged by the amount

$$-\sum_{i=1}^d b_t^i = \sum_{i=1}^d \sum_{j=1}^d \lambda^{ij} \alpha_t^{ij}$$

diminishing the total portfolio value.

With every $\mathbf{M}_+^d$ -valued process (α_t) and any initial endowment

$$v = V_{-1} \in \mathbf{R}^d$$

we associate, using recursively the formula (2), a value process $V = (V_t)$, $t = 0, \dots, T$. The terminal values of these processes form the set R_T^v .

Remark 3.3 In the literature one can find other specifications for transaction costs coefficients. To explain the situation, let us define $\tilde{\alpha}^{ij} := (1 + \lambda^{ij}) \alpha^{ij}$. The

increment of value of the i -th position can be written as

$$b^i = \sum_{j=1}^d \mu^{ji} \tilde{\alpha}_t^{ji} - \sum_{j=1}^d \tilde{\alpha}_t^{ij},$$

where $\mu^{ji} := 1/(1 + \lambda^{ji}) \in]0, 1]$. The matrix (μ^{ij}) can be specified as the matrix of the transaction costs coefficients. In models with a traded numéraire, i.e. a non-risky asset, a mixture of both specifications is used quite often.

Before analyzing the model, we write it in a more convenient way reducing the dimension of the action space.

To this aim we define, for every (ω, t) , the convex cone

$$M_t(\omega) := \left\{ x \in \mathbf{R}^d : \exists a \in \mathbf{M}_+^d \text{ such that } x^i = \sum_{j=1}^d [(1 + \lambda_t^{ij}(\omega))a^{ij} - a^{ji}], i \leq d \right\},$$

which is a polyhedral one as it is the image of the polyhedral cone $\mathbf{M}_+^d$ under a linear mapping. Its dual positive cone

$$M_t^*(\omega) := \left\{ w \in \mathbf{R}^d : \inf_{x \in M_t(\omega)} wx \geq 0 \right\}$$

can be easily described by linear homogeneous inequalities. Specifically,

$$M_t^*(\omega) = \{w \in \mathbf{R}^d : w^j - (1 + \lambda_t^{ij}(\omega))w^i \leq 0, 1 \leq i, j \leq d\}.$$

We introduce also the solvency cone (in values)

$$K_t(\omega) := \left\{ x \in \mathbf{R}^d : \exists a \in \mathbf{M}_+^d \text{ such that } x^i + \sum_{j=1}^d [a^{ji} - (1 + \lambda_t^{ij}(\omega))a^{ij}] \geq 0, \right. \\ \left. i \leq d \right\},$$

i.e. $K_t(\omega) = M_t(\omega) + \mathbf{R}_+^d$. The negative holdings of a position vector in $K_t(\omega)$ can be liquidated (under transaction costs given by $(\lambda_t^{ij}(\omega))$) to get a position vector in $\mathbf{R}_+^d$.

Let $\mathcal{B}$ be the set of all processes $B = (B_t)$ with $\Delta B_t \in L^0(-M_t, \mathcal{F}_t)$. It is an easy exercise on measurable selection to check that ΔB_t can be represented using a certain $\mathcal{F}_t$ -measurable transfer matrix α_t . Thus, the set of portfolio process in the “value domain” coincides with the set of processes $V = V^{v, B}$, $B \in \mathcal{B}$, given by the system of linear difference equations

$$\Delta V_t^i = V_{t-1}^i \Delta Y_t^i + \Delta B_t^i, \quad V_{-1}^i = v^i, \quad (3)$$

with

$$\Delta Y_t^i = \frac{\Delta S_t^i}{S_{t-1}^i}, \quad Y_0^i = 1. \quad (4)$$

Remark 3.4 Using the notations introduced at the beginning of this section, we can rewrite these equations in the integral form

$$V = v + V_- \bullet Y + B, \quad (5)$$

with

$$Y^i = 1 + (1/S_-^i) \cdot S^i, \quad (6)$$

which remains the same also for the continuous-time version but with a different meaning of the symbols, see [34], [39].

It is easier to study no-arbitrage properties of the model working in the “physical domain” where portfolio evolves only because of the agent’s action. Indeed, the dynamics of $\widehat{V}$ is simpler:

$$\Delta V_t^i = \frac{\Delta B_t^i}{S_t^i}.$$

This equation is obvious because of its financial interpretation but one can check it formally (e.g., using the product formula).

Put $\widehat{M}_t(\omega) := \phi_t(\omega)M_t(\omega)$ and introduce the solvency cone (in physical units)

$$\widehat{K}_t(\omega) := \phi_t K_t(\omega) = \widehat{M}_t(\omega) + \mathbf{R}^d.$$

Every process $\widehat{b}$ with $\widehat{b}_t \in L^0(-\widehat{M}_t, \mathcal{F}_t)$, $0 \leq t \leq T$, defines a portfolio process $\widehat{V}$ with $\Delta V = \widehat{b}$ and the zero initial endowment. All portfolio processes (in physical units) can be obtained in this way.

The notations R_T^0 and $\widehat{R}_T^0$ are obvious.

Lemma 3.5 *The following conditions are equivalent:*

- (a) $R_T^0 \cap L^0(K_T, \mathcal{F}_T) \subseteq L^0(\partial K_T, \mathcal{F}_T)$;
- (b) $R_T^0 \cap L^0(\mathbf{R}_+^d, \mathcal{F}_T) = \{0\}$;
- (c) $\widehat{R}_T^0 \cap L^0(\mathbf{R}_+^d, \mathcal{F}_T) = \{0\}$.

Proof The equivalence of (b) and (c) is obvious. The implication (a) $\Rightarrow$ (b) holds because $\mathbf{R}_+^d \setminus \{0\}$ is a subset of $\text{int } K_T$. To prove the remaining implication (b) $\Rightarrow$ (a) we notice that if $V_T^B \in L^0(K_T, \mathcal{F}_T)$ where $B \in \mathcal{B}$ then there exists $B' \in \mathcal{B}$ such that $V_T^{B'} \in L^0(\mathbf{R}_+^d, \mathcal{F}_T)$ and $V_T^{B'}(\omega) \neq 0$ on the set $V_T^B(\omega) \notin \partial K_T(\omega)$. To construct such B' , it is sufficient to modify only ΔB_T by combining the last transfer with the liquidation of the negative positions. $\square$

In accordance with [41] we shall say that the market has *weak no-arbitrage property at the date T* ($\mathbf{NA}_T^w$) if one of the equivalent conditions of the above lemma is fulfilled. Apparently, $\mathbf{NA}_T^w$ implies $\mathbf{NA}_t^w$ for all $t \leq T$.

Lemma 3.6 Assume that Ω is finite. Then $\widehat{R}_T^0 \cap L^0(\mathbf{R}_+^d, \mathcal{F}_T) = \{0\}$ if and only if there exists a d -dimensional martingale Z with strictly positive components such that $Z_t \in L^0(\widehat{M}_t^*, \mathcal{F}_t)$.

Proof The cone $\widehat{R}_T^0$ is polyhedral. In virtue of Theorem 2.1 the first condition is equivalent to the existence of a strictly positive random variable η such that $E\eta\zeta \leq 0$ for all $\zeta \in \widehat{R}_T^0$. Let $Z_t = E(\eta|\mathcal{F}_t)$. Since $L^0(-\widehat{M}_t, \mathcal{F}_T) \subseteq \widehat{R}_T^0$, the inequality $EZ_t\zeta \geq 0$ holds for all $\zeta \in L^0(\widehat{M}_t, \mathcal{F}_t)$ implying that $Z_t \in L^0(\widehat{M}_t^*, \mathcal{F}_t)$. If the second condition of the lemma is fulfilled, we can take $\eta = Z_T$. $\square$

Let $\mathcal{D}_T$ be the set of martingales $Z = (Z_t)$ such that $\widehat{Z}_t \in L^0(K_t^*, \mathcal{F}_t)$. The following result from [41] is a simple corollary of the above criteria:

Theorem 3.7 Assume that Ω is finite. Then $\mathbf{NA}_T^w$ holds if and only if there exists a process $Z \in \mathcal{D}$ with strictly positive components.

This result contains the Harrison–Pliska theorem. Indeed, in the case where all $\lambda^{ij} = 0$, the cone $K = \tilde{K} := \{x \in \mathbf{R}^d : x^1 \geq 0\}$ and $K^* = \mathbf{R}_+ \mathbf{1}$. Thus, for $Z \in \mathcal{D}$ all components of the process $\widehat{Z}$ are equal. If, e.g., the first asset is the numéraire, then $\widehat{Z}^1 = Z^1$ is a martingale as well as the processes $S^i Z^1$, $i = 2, \dots, d$, i.e. Z^1 is a martingale density.

Remark 3.8 For models with transaction costs other types of arbitrage may be of interest. E.g., it is quite natural to consider the ordering induced by the cone $\tilde{K} := \{x \in \mathbf{R}^d : x^1 \geq 0\}$ (corresponding to the absence of transaction costs), see a criterion in [41] which can be obtained along the same lines as above.

Remark 3.9 It is easily seen that

$$\widehat{M}_t(\omega) := \left\{ y \in \mathbf{R}^d : \exists c \in \mathbf{M}_+^d \text{ such that } y^i = \sum_{j=1}^d [\pi_t^{ij}(\omega) c^{ij} - c^{ji}], \ i \leq d \right\}, \quad (7)$$

where

$$\pi_t^{ij} := (1 + \lambda_t^{ij}) S_t^j / S_t^i, \quad 1 \leq i, j \leq d. \quad (8)$$

One can start the modeling by specifying instead of the process (λ_t^{ij}) the process (π_t^{ij}) with values in the set of non-negative matrices with units on the diagonal. Defining directly the set of processes $\widehat{V}$ with $\Delta \widehat{V}_t \in L^0(-\widehat{M}_t, \mathcal{F}_t)$ and the set of “results” $\widehat{R}_T^0$, one can get Lemma 3.6 immediately. The advantage of this approach is that the existence of the reference asset (i.e. of the price process S) is not assumed and we have a model of “pure exchange”. A question arises when such a model can be reduced to a transaction costs model with a reference asset, i.e. under what

conditions on the matrix (π^{ij}) one can find a matrix (λ^{ij}) with positive entries and a vector S with strictly positive entries satisfying the relation (8).

3.4 The Dalang–Morton–Willinger theorem

Let us consider again the classical model of a frictionless market but now without any assumption on the stochastic basis.

Theorem 3.10 *The following conditions are equivalent:*

- (a) $R_T^0 \cap L^0(\mathbf{R}_+, \mathcal{F}_T) = \{0\}$ (no-arbitrage);
- (b) $A_T^0 \cap L^0(\mathbf{R}_+, \mathcal{F}_T) = \{0\}$;
- (c) $A_T^0 \cap L^0(\mathbf{R}_+, \mathcal{F}_T) = \{0\}$ and $A_T^0 = \bar{A}_T^0$, the closure in L^0 ;
- (d) $\bar{A}_T^0 \cap L^0(\mathbf{R}_+, \mathcal{F}_T) = \{0\}$;
- (e) for every probability measure $P' \sim P$ there is a measure $\tilde{P} \sim P$ such that $d\tilde{P}/dP' \leq \text{const}$ and $S \in \mathcal{M}(\tilde{P})$;
- (f) there is a probability measure $\tilde{P} \sim P$ such that $S \in \mathcal{M}(\tilde{P})$.
- (g) there is a probability measure $\tilde{P} \sim P$ such that $S \in \mathcal{M}_{loc}(\tilde{P})$.

It seems that these equivalent conditions (among many others) are the most essential ones to be collected in a single theorem. The equivalence of (a), (e), and (f) relating a “financial property” of absence of arbitrage with important “probabilistic” properties is due to Dalang, Morton, and Willinger [8]. Their approach is based on a reduction to a one-stage problem which is very simple for the case of trivial initial σ -algebra; regular conditional distributions and measurable selection theorem allows us to extend the arguments to treat the general case, see [53], [29], and [58] for other implementations of the same idea. Formally, the equivalence (a) $\Leftrightarrow$ (f) is exactly the same as the Harrison–Pliska theorem and one could think that it is just the same result under the relaxed hypothesis on Ω . In fact, such a conclusion seems to be superficial: the equivalent “functional-analytic property” (c), discovered by Schachermayer in [56], shows clearly the profound difference between these two situations. Schachermayer’s condition opens the door to an extensive use of geometric functional analysis in the discrete-time setting which was reserved previously only for continuous-time models. It is quite interesting to notice that the set R_T^0 is always closed while A_T^0 is not.

The condition (d) introduced by Stricker in [60] also gives a hint on an appropriate use of separation arguments. Specifically, the Kreps–Yan theorem (see the Appendix) can be applied to separate $A_0^T \cap L^1(P')$ from $L_+^1(P') = L^1(\mathbf{R}_+, P')$ where the measure $P' \sim P$ can be chosen arbitrarily: this freedom allows us to obtain an “equivalent separating measure” with a desired property.

Notice that the crucial implication $(b) \Rightarrow (d)$ seems to be easier to prove than $(a) \Rightarrow (c)$, see [36] where a kind of “linear algebra” with random coefficients was suggested.

The literature provides a variety of other equivalent conditions complementing the list of the above theorem. Some of them are interesting and non-trivial. A family of conditions is related with various classes of admissible strategies $\mathcal{B}$ (which is the set of all predictable process in our formulation). Since the sets R_T^0 and A_T^0 depend on this class, so does the no-arbitrage property. It happens, however, that the latter is quite “robust”: e.g., it remains the same if we consider as admissible only the strategies with non-negative value processes. The problem of admissibility is not of great importance since we assume a finite time horizon. The situation is radically different for continuous-time models where one must work out the doubling strategies which allow us to win even betting on a martingale.

Proof of Theorem 3.10 The implications $(a) \Rightarrow (b)$ and $(c) \Rightarrow (d)$ are obvious as well as the chain $(e) \Rightarrow (f) \Rightarrow (g)$.

To prove the implication $(d) \Rightarrow (e)$ we observe that the two properties are invariant under the equivalent change of measure. Thus, we may assume that $P' = P$ and, moreover, by passing to the measure $ce^{-\eta}P$ with $\eta = \sup_{t \leq T} |S_t|$, that all S_t are integrable. The set $\bar{A}_0^1 \cap L^1$ is closed in L^1 and intersects with L_+^1 only at zero. By the Kreps–Yan theorem there is a $\tilde{P}$ with $d\tilde{P}/dP \in L^\infty$ such that $\tilde{E}\xi \leq 0$ for all $\xi \in \bar{A}_0^1 \cap L^1$. Taking $\xi = \pm H_t \Delta S_t$ where H_t is bounded and $\mathcal{F}_{t-1}$ -measurable, we conclude that S is a martingale.

The implication $(g) \Rightarrow (a)$ is also easy. If $H \cdot S_t \geq 0$ for all $t \leq T$, then, by the Fatou lemma, the local $\tilde{P}$ -martingale $H \cdot S$ is a $\tilde{P}$ -supermartingale and, therefore, $\tilde{E}H \cdot S_T \leq 0$, i.e. $H \cdot S_T = 0$. In other words, there is no arbitrage in the class of strategies with non-negative value processes. This implies (a) since for any arbitrage opportunity H there is an arbitrage opportunity H' with non-negative value process. Indeed, if $P(H \cdot S_s \leq -b) > 0$ for some $s < T$ and $b > 0$, then one can take $H' = I_{[s, T] \times \{H \cdot S_s \leq -b\}} H$.

In the proof of the “difficult” implication $(b) \Rightarrow (c)$ we follow [42].

Lemma 3.11 *Let $\eta^n \in L^0(\mathbf{R}^d)$ be such that $\underline{\eta} := \liminf |\eta^n| < \infty$. Then there are $\tilde{\eta}^k \in L^0(\mathbf{R}^d)$ such that for all ω the sequence of $\tilde{\eta}^k(\omega)$ is a convergent subsequence of the sequence of $\eta^n(\omega)$.*

Proof Let $\tau_0 := 0$ and $\tau_k := \inf\{n > \tau_{k-1} : ||\eta^n| - \underline{\eta}| \leq 1/k\}$. Then $\tilde{\eta}_0^k := \eta^{\tau_k}$ is in $L^0(\mathbf{R}^d)$ and $\sup_k |\tilde{\eta}_0^k| < \infty$. Working further with the sequence of $\tilde{\eta}_0^k$ we construct, applying the above procedure to the first component, a sequence of $\tilde{\eta}_1^k$ with the convergent first component and such that for all ω the sequence of $\tilde{\eta}_1^k(\omega)$ is

a subsequence of the sequence of $\tilde{\eta}_0^n(\omega)$. Passing on each step to the newly created sequence of random variables and to the next component we arrive at a sequence with the desired properties. $\square$

To show that A_T^0 is closed we proceed by induction. Let $T = 1$. Suppose that $H_1^n \Delta S_1 - r^n \rightarrow \zeta$ a.s., where H_1^n is $\mathcal{F}_0$ -measurable and $r^n \in L_+^0$. It is sufficient to find $\mathcal{F}_0$ -measurable random variables $\tilde{H}_1^k$ convergent a.s. and $\tilde{r}^k \in L_+^0$ such that $\tilde{H}_1^k \Delta S_1 - \tilde{r}^k \rightarrow \zeta$ a.s.

Let $\Omega_i \in \mathcal{F}_0$ form a finite partition of Ω . Obviously, we may argue on each Ω_i separately as on an autonomous measure space (considering the restrictions of random variables and traces of σ -algebras).

Let $\underline{H}_1 := \liminf |H_1^n|$. On $\Omega_1 := \{\underline{H}_1 < \infty\}$ we take, using Lemma 3.11, $\mathcal{F}_0$ -measurable $\tilde{H}_1^k$ such that $\tilde{H}_1^k(\omega)$ is a convergent subsequence of $H_1^n(\omega)$ for every ω ; $\tilde{r}^k$ are defined correspondingly. Thus, if Ω_1 is of full measure, the goal is achieved.

On $\Omega_2 := \{\underline{H}_1 = \infty\}$ we put $G_1^n := H_1^n / |H_1^n|$ and $h_1^n := r_1^n / |H_1^n|$ and observe that $G_1^n \Delta S_1 - h_1^n \rightarrow 0$ a.s. By Lemma 3.11 we find $\mathcal{F}_0$ -measurable $\tilde{G}_1^k$ such that $\tilde{G}_1^k(\omega)$ is a convergent subsequence of $G_1^n(\omega)$ for every ω . Denoting the limit by $\tilde{G}_1$, we obtain that $\tilde{G}_1 \Delta S_1 = \tilde{h}_1$ where $\tilde{h}_1$ is non-negative, hence, in virtue of (b), $\tilde{G}_1 \Delta S_1 = 0$.

As $\tilde{G}_1(\omega) \neq 0$, there exists a partition of Ω_2 into d disjoint subsets $\Omega_2^i \in \mathcal{F}_0$ such that $\tilde{G}_1^i \neq 0$ on Ω_2^i . Define $\tilde{H}_1^n := H_1^n - \beta^n \tilde{G}_1$ where $\beta^n := H_1^n / \tilde{G}_1^i$ on Ω_2^i . Then $\tilde{H}_1^n \Delta S_1 = H_1^n \Delta S_1$ on Ω_2 . We repeat the procedure on each Ω_2^i with the sequence $\tilde{H}_1^n$ knowing that $\tilde{H}_1^n = 0$ for all n . Apparently, after a finite number of steps we construct the desired sequence.

Let the claim be true for $T - 1$ and let $\sum_{t=1}^T H_t^n \Delta S_t - r^n \rightarrow \zeta$ a.s., where H_t^n are $\mathcal{F}_{t-1}$ -measurable and $r^n \in L_+^0$. By the same arguments based on the elimination of non-zero components of the sequence H_1^n and using the induction hypothesis we replace H_t^n and r^n by $\tilde{H}_t^k$ and $\tilde{r}^k$ such that $\tilde{H}_1^k$ converges a.s. This means that the problem is reduced to the one with $T - 1$ steps. $\square$

4 No-arbitrage criteria in continuous time

Nowadays, in the era of electronic trading, there are no doubts that continuous-time models are much more important than their discrete-time relatives. As a theoretical tool, differential equations (eventually, stochastic) show enormous advantage with respect to difference equations. Easy to analyze, they provide very precise description of various phenomena and, quite often, allow for tractable closed-form solutions. As we mentioned already, the mathematical finance started from a continuous-time model. The unprecedented success of the Black–Scholes formula

confirmed that such models are adequate tools to describe financial market phenomena. The current trend is to go beyond the Black–Scholes world. Statistical tests for financial data reject the hypothesis that prices evolve as processes with continuous sample paths. Much better approximation can be obtained by stable or other types of Lévy processes. Apparently, semimartingales provide a natural framework for discussion of general concepts of financial theory like arbitrage and hedging problems. Though more general processes are also tried, yet a very weak form of absence of arbitrage (namely, the NFLVR-property for simple integrands) in the case of a locally bounded price process implies that it is a semimartingale, see Theorem 7.2 in [12].

4.1 No Free Lunch and separating measure

In this subsection we explain relations between the No Free Lunch (NFL) condition due to Kreps, No Free Lunch with Bounded Risk (NFLBR) due to Delbaen, and No Free Lunch with Vanishing Risk (NFLVR) introduced by Delbaen and Schachermayer (see, [48], [10], [12]).

Let us assume that in a one-step model of frictionless market admissible strategies are such that the convex cone R_T^0 (the set of final portfolio values corresponding to zero initial endowment) contains only (scalar) random variables bounded from below. As usual, let $A_T^0 := R_T^0 - L^0(\mathbf{R}_+)$. Define the set $C := A_T^0 \cap L^\infty$. We denote by $\bar{C}$, $\tilde{C}^*$, and $\bar{C}^*$ the norm closure, the union of weak* closures of denumerable subsets, and the weak* closure of C in L^∞ ; $C_+ := C \cap L_+^\infty$ etc.

The properties NA, NFLVR, NFLBR, and NFL mean that $C_+ = \{0\}$, $\bar{C}_+ = \{0\}$, $\tilde{C}_+^* = \{0\}$, and $\bar{C}_+^* = \{0\}$, respectively. Consecutive inclusions induce the hierarchy of these properties:

$$\begin{array}{ccccccc} C & \subseteq & \bar{C} & \subseteq & \tilde{C}^* & \subseteq & \bar{C}^* \\ \text{NA} & \Leftarrow & \text{NFLVR} & \Leftarrow & \text{NFLBR} & \Leftarrow & \text{NFL}. \end{array}$$

Define the ESM (Equivalent Separating Measure) property as follows: *there exists $\tilde{P} \sim P$ such that $\tilde{E}\xi \leq 0$ for all $\xi \in R_T^0$.*

The following criterion for the NFL-property was established by Kreps.

Theorem 4.1 $\text{NFL} \Leftrightarrow \text{ESM}$.

Proof ($\Leftarrow$) Let $\xi \in \bar{C}^* \cap L_+^\infty$. Since $d\tilde{P}/dP \in L^1$, there are $\xi_n \in C$ with $\tilde{E}\xi_n \rightarrow \tilde{E}\xi$. By definition, $\xi_n \leq \zeta^n$ where $\zeta^n \in R_T^0$. Thus, $\tilde{E}\xi_n \leq 0$ implying that $\tilde{E}\xi \leq 0$ and $\xi = 0$.

($\Rightarrow$) Since $\bar{C}^* \cap L_+^\infty = \{0\}$, the Kreps–Yan separation theorem given in the

Appendix provides $\tilde{P} \sim P$ such that $\tilde{E}\xi \leq 0$ for all $\xi \in C$, hence, for all $\xi \in R_T^0$. $\square$

4.2 Semimartingale model

Let $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t), P)$ be a stochastic basis, i.e. a probability space equipped with a filtration $\mathbf{F}$ satisfying the “usual conditions”. Assume for simplicity that the initial σ -algebra is trivial, the time horizon T is finite, and $\mathcal{F}_T = \mathcal{F}$.

A process $X = (X_t)_{t \in [0, T]}$ (right-continuous and with left limits) is a semimartingale if it can be represented as a sum of a local martingale and a process of bounded variation. Let $\mathcal{U}_1$ be the set of all predictable processes h taking values in the interval $[-1, 1]$. We denote by $h \cdot S$ the stochastic integral of a predictable process h with respect to a semimartingale. The definition of this integral in its full generality, especially for vector processes (necessary for financial application), is rather complicated and we send the reader to textbooks on stochastic calculus.

The linear space $\mathcal{S}$ of semimartingales starting from zero is a Fréchet space with the quasinorm

$$\mathbf{D}(X) := \sup_{h \in \mathcal{U}_1} E(1 \wedge |h \cdot X_T|)$$

which induces the Émery topology, [17].

We fix in $\mathcal{S}$ a closed convex subset $\mathcal{X}^1$ of processes $X \geq -1$ which contains 0 and satisfies the following condition: for any $X, Y \in \mathcal{X}^1$ and for any non-negative bounded predictable processes H, G with $HG = 0$ the process $Z := H \cdot X + G \cdot Y$ belongs to $\mathcal{X}^1$ if $Z \geq -1$.

Put $\mathcal{X} := \text{cone } \mathcal{X}^1$. The set $\mathcal{X}$ is interpreted as the set of value processes.

Put $R_T^0 := \{X_T : X \in \mathcal{X}\}$.

In this rather general semimartingale model we have

$$\text{NFLVR} \Leftrightarrow \text{NFLBR} \Leftrightarrow \text{NFL}$$

in virtue of the following:

Theorem 4.2 *Under NFLVR $C = \tilde{C}^*$.*

The proof of this theorem given in [34] follows closely the arguments of the Delbaen–Schachermayer paper [12]. Their setting is based on a n -dimensional price process S , the admissible strategies H are predictable $\mathbf{R}^n$ -valued processes for which stochastic integrals $H \cdot S$ are defined and bounded from below. The set $\mathcal{X}^1$ of all value process $H \cdot S \geq -1$ is closed in virtue of the Mémin theorem on closedness in $\mathcal{S}$ of the space of stochastic integrals [50]. If S is bounded then the process $H = \xi I_{[s, t]}$ is admissible for arbitrary $\xi \in L^\infty(\mathbf{R}^n, \mathcal{F}_t)$, and hence $\tilde{E}\xi(S_t - S_s) \leq 0$ for any separating measure $\tilde{P}$. In fact, there is equality

here because one can change the sign of ξ . Thus, if S is bounded then it is a martingale with respect to any separating measure $\tilde{P}$. It is an easy exercise to check that if S is locally bounded (i.e. if there exists a sequence of stopping times τ_k increasing to infinity such that the stopped processes S^{τ_k} are bounded) then S is a local martingale with respect to $\tilde{P}$. The case of arbitrary, not necessarily bounded S is of a special interest because the semimartingale model includes the classical discrete-time model as a particular case. The corresponding theorem, also due to Delbaen–Schachermayer [14], involves the notions of a σ -martingale and an equivalent σ -martingale measure.

A semimartingale S is a σ -martingale (notation: $S \in \Sigma_m$) if $G \cdot S \in \mathcal{M}_{loc}$ for some G with values in $]0, 1]$. The property $E\sigma MM$ means that there is $Q \sim P$ such that $S \in \Sigma_m(Q)$.

Theorem 4.3 *Let $\mathcal{X}^1$ be the set of stochastic integrals $H \cdot S \geq -1$. Then*

$$NFLVR \Leftrightarrow NFLBR \Leftrightarrow NFL \Leftrightarrow ESM \Leftrightarrow E\sigma MM.$$

The remaining non-trivial implication $ESM \Rightarrow E\sigma MM$ follows from

Theorem 4.4 *Let $\tilde{P}$ be a separating measure. Then for any $\varepsilon > 0$ there is $Q \sim \tilde{P}$ with $\text{Var}(\tilde{P} - Q) \leq \varepsilon$ such that S is a σ -martingale under Q .*

A brief account of the Delbaen–Schachermayer theory including a short proof of the above theorem based on the inequality for the total variation distance from [40] is given in [33].

4.3 Hedging theorem and optional decomposition

Let us consider the semimartingale model based on an n -dimensional price process S . Let C be a scalar random variable bounded from below and let

$$\Gamma := \{x \in \mathbf{R} : \exists \text{ admissible } H \text{ such that } x + H \cdot S_T \geq C\}.$$

In other words, Γ is the set of initial endowments for which one can find an admissible strategy such that the terminal value of the corresponding portfolio dominates (super-replicates) the contingent claim C . “Admissible” means that the portfolio process is bounded from below by a constant.

Obviously, if non-empty, Γ is a semi-infinite interval. The following “hedging” theorem gives its characterization.

Let $\mathcal{Q}$ be the set of probability measures $Q \sim P$ with respect to which S is a local martingale.

Theorem 4.5 Assume that $\mathcal{Q} \neq \emptyset$. Then $\Gamma = [x_*, \infty[$ where

$$x_* = \sup_{Q \in \mathcal{Q}} E_Q C.$$

This general formulation is due to Kramkov [47] who noticed that the assertion is a simple corollary of the following two results.

Theorem 4.6 Assume that $\mathcal{Q} \neq \emptyset$. Let X be a process bounded from below which is a supermartingale with respect to any $Q \in \mathcal{Q}$. Then there is an admissible strategy H and an increasing process A such that $X = X_0 + H \cdot S - A$.

The process $H \cdot S$, being bounded from below, is a local martingale with respect to every $Q \in \mathcal{Q}$ (the property that an integral with respect to a local martingale is also a local martingale if it is one-side bounded is due to Émery for the scalar case and to Ansel and Stricker [1] for the vector case). Thus, this decomposition resembles that of Doob–Meyer but it holds simultaneously for the whole set $\mathcal{Q}$; in general, it is non-unique and A may not be predictable but only adapted, hence, A , being right-continuous, is optional. This explains why the above result is usually referred to as the optional decomposition theorem. It was proved in [47] for the case where S is locally bounded; this assumption was removed in the paper [18]. The proof in [18] is probabilistic and provides an interpretation of the integrand H as the Lagrange multiplier. Alternative proofs with intensive use of functional analysis can be found in [13]. For an optional decomposition with constraints see [20], an extended discussion of the problem is given [19]. In [43] it is shown that if $P \in \mathcal{Q}$ then the subset of $\mathcal{Q}$ formed by the measures with bounded densities is dense in $\mathcal{Q}$; this result implies, in particular, that, without any hypothesis, the subset of (local) martingale measures with bounded entropy is dense in $\mathcal{Q}$.

Proposition 4.7 Assume that C is such that $\sup_{Q \in \mathcal{Q}} E_Q C < \infty$. Then there exists a process X which is a supermartingale with respect to every $Q \in \mathcal{Q}$ such that

$$X_t = \text{ess sup}_{Q \in \mathcal{Q}} E_Q(C | \mathcal{F}_t).$$

This result is due to El Karoui and Quenez [16]; its proof also can be found in [47].

Proof of Theorem 4.5 The inclusion $\Gamma \subseteq [x_*, \infty[$ is obvious: if $x + H \cdot S_T \geq C$ then $x \geq E_Q C$ for every $Q \in \mathcal{Q}$. To show the opposite inclusion we may suppose that $\sup_{Q \in \mathcal{Q}} E_Q H < \infty$ (otherwise both sets are empty). Applying the optional decomposition theorem to the process

$$X_t = \text{ess sup}_{Q \in \mathcal{Q}} E_Q(C | \mathcal{F}_t)$$

we get that $X = x_* + H \cdot S - A$. Since $x_* + H \cdot S_T \geq X_T = C$, the result follows. $\square$

4.4 Semimartingale model with transaction costs

In this model it is assumed that the price process is a semimartingale S with non-negative components. The dynamics of the value process $V = V^{v,B}$ is given by the linear stochastic equation

$$V = v + V_- \bullet Y + B$$

where $Y^i = (1/S_-^i) \cdot S^i$,

$$B^i := \sum_{j=1}^d L^{ji} - \sum_{j=1}^d (1 + \lambda^{ij}) L^{ij},$$

and L^{ij} is an increasing right-continuous process representing the accumulated net wealth “arriving” at a position i from the position j .

At this level of generality, criteria of absence of arbitrage are still not available but the paper of Jouini and Kallal [30] is an important contribution to the subject. It provides an NFL criterion for the model of stock market with a bid–ask spread where, instead of transaction costs coefficients, two processes are given, $\underline{S}$ and $\bar{S}$, describing the evolution of the selling and buying prices. It is shown that a certain (specifically formulated) NFL property holds if and only if there exist a probability measure $\tilde{P} \sim P$ and a process S whose components evolve between the corresponding components of $\underline{S}$ and $\bar{S}$ such that S is a martingale with respect to $\tilde{P}$. This result is consistent with the NA criteria for finite Ω , see [41]. Apparently, the approach of Jouini and Kallal can be easily extended to the case of currency markets. However, one should take care that the setting of [30] is that of the L^2 -theory. The limitations of the latter in the context of financial modeling are well-known; in contrast with engineering where energy constraints are welcome, they do not admit an economical interpretation. We attract the reader’s attention to the recent paper [32] of the same authors where problems of equilibrium and viability (closely related with absence of arbitrage) are discussed; see also [31] for models with short-sell constraints.

The situation with the hedging theorem is slightly better. Its first versions in [6] (for a two-asset model) and in [34] were established within the L^2 -framework. In the preprint [38] an attempt was made to work with the class of strategies for which the value process is bounded from below in the sense of partial ordering induced by the solvency cone. This class of strategies corresponds precisely to the usual definition of admissibility in the case of frictionless market. However, the result

was proved only for bounded price processes. To avoid difficulties one can look for other reasonable classes of admissible strategies. This approach was exploited in the paper [39] which contains the following hedging theorem.

It is assumed that the matrix Λ of transaction costs coefficients is constant, the first asset is the numéraire, and there exists a probability measure $\tilde{P}$ such that S is a (true) martingale with respect to $\tilde{P}$.

Let $\mathcal{B}_b$ be the class of strategies B such that the corresponding value processes are bounded from below by a price process multiplied by (negative) constants (this definition resembles that used by Sin in the frictionless case, [55]). In particular, it is admissible to keep short a finite number of units of assets.

Let $\mathcal{D}$ be the set of martingales Z such that $\hat{Z}$ takes values in K^* . Notice that $\{Z : \hat{Z} = w\rho, w \in K^*\} \subseteq \mathcal{D}$ where $\rho_t := E(d\tilde{P}/dP|\mathcal{F}_t)$. Moreover, $Z \in \mathcal{D}$ and we have $\hat{Z}^1 = Z^1$; since the transaction costs are constant, it follows from the inequalities defining K^* that $|\hat{Z}| \leq \kappa Z^1$ for a certain fixed constant κ . With these remarks it is easy to conclude that $\hat{Z}V^{v,B}$ is always a supermartingale whatever $Z \in \mathcal{D}$ and $B \in \mathcal{B}_b$ are.

Define the convex set of hedging endowments

$$\Gamma = \Gamma(\mathcal{B}_b) := \{v \in \mathbf{R}^d : \exists B \in \mathcal{B}_b \text{ such that } V_T^{v,B} \geq_K C\}$$

and the closed convex set

$$D := \{v \in \mathbf{R}^d : \hat{Z}_0 v \geq E\hat{Z}_T C \quad \forall Z \in \mathcal{D}\}.$$

Theorem 4.8 *Assume that S is a continuous process and the solvency cone K is proper. Then $\Gamma = D$.*

The “easy” inclusion $\Gamma \subseteq D$ holds in virtue of the supermartingale property of $\hat{Z}V^{v,B}$ even without extra assumptions. The proof of the opposite inclusion given in [39] is based on a bipolar theorem in the space $L^0(\mathbf{R}^d, \mathcal{F}_T)$ equipped with a partial ordering. The hypotheses of the theorem and the structure of admissible strategies are used heavily in this proof. The assumption that K is proper, i.e. the interior (of K^*) is non-empty, is essential (otherwise, Γ may not be closed). However, the assertion $\bar{\Gamma} = D$ can be established for arbitrary K . How to remove or relax the assumptions on continuity of S to make the result adequate to the hedging theorem without friction remains an open problem.

Remark 4.9 It is important to note that the set of hedging endowments depends on the chosen class of admissible strategies. Let $\mathcal{B}^0$ be the class of buy-and-hold strategies with a single revision of the portfolio, namely, at time zero when the investor enters the market. It happens that in the most popular two-asset model under transaction costs with the price dynamics given by the geometric Brownian

motion where the problem is to hedge a European call option (or, more generally, a contingent claim $C = g(S_T)$) we have $\Gamma(\mathcal{B}_b) = \Gamma(\mathcal{B}^0)$. This astonishing property was conjectured by Davis and Clark [9] and proved independently in [49] and [59], see also [7] and [2] for further generalizations. More precisely, in the mentioned papers it was shown that the investor having the initial endowment in money which is a minimal one to hedge the contingent claim C , can hedge it using buy-and-hold strategy from $\mathcal{B}^0$. In other words, the conclusion was that the point with zero ordinate on the boundary of $\Gamma(\mathcal{B}_b)$ belongs also to the boundary of a smaller set $\Gamma(\mathcal{B}^0)$. In fact, one can extend the arguments and prove that both sets coincide.

5 Large financial markets

5.1 Ross–Huberman APM

The main conclusion of the Capital Asset Pricing Model (CAPM) by Lintner and Sharp is the following:

the mean excess return on an asset is a linear function of its “beta”, a measure of risk associated with this asset.

More precisely, we have the following result. Assume for simplicity that the riskless asset pays no interest. Suppose that the return on the i -th asset has mean μ_i and variance σ_i^2 , the market portfolio return has mean μ_0 and variance σ_0^2 . Let γ_i be the correlation coefficient between the returns on the i -th asset and the market portfolio. Then $\mu_i = \mu_0\beta_i$ where $\beta_i := \gamma_i\sigma_i/\sigma_0$.

Unfortunately, the theoretical assumptions of CAPM are difficult to justify and its empirical content is dubious. One can expect that the empirical values of (β_i, μ_i) form a cloud around the so-called security market line but this phenomenon is observed only for certain data sets. The alternative approach, the Arbitrage Pricing Model (APM) suggested by Ross in [54] and placed on a solid mathematical basis by Huberman, results in a conclusion that there exists a relation between model parameters, which can be viewed as “approximately linear”, giving much better consistency with empirical data. Based on the idea of asymptotic arbitrage, it attracted considerable attention, see, e.g., [3], [4], [26], [27]; sometimes it is referred to as the Arbitrage Pricing Theory (APT). An important reference is the note by Huberman [25] who gave a rigorous definition of the asymptotic arbitrage together with a short and transparent proof of the fundamental result of Ross. The idea of Huberman is to consider a sequence of classical one-step finite-asset models instead of a single one with infinite number of securities (in the latter case an unpleasant phenomenon may arise similar to that of doubling strategies for models with infinite time horizon). When the number of assets increases to infinity, this sequence of models can be considered as a description of a large financial market.

A general specification of the n -th model M^n is as follows. We are given a stochastic basis $(\Omega^n, \mathcal{F}^n, \mathbf{F}^n, P^n)$ with a convex cone R_T^{0n} of square integrable (scalar) random variables. Assume for simplicity that the initial σ -algebra is trivial, $\mathcal{F}_T = \mathcal{F}$. Here T stands for “terminal” and can be replaced by 1. As usual, the elements of R_T^{0n} are interpreted as the terminal values of portfolios.

By definition, a sequence $\xi^n \in R_T^{0n}$ realizes an *asymptotic arbitrage opportunity* (AAO) if the following two conditions are fulfilled (E^n and D^n denote the mean and variance with respect to P^n):

- (a) $\lim_n E^n \xi^n = \infty$;
- (b) $\lim_n D^n \xi^n = \lim_n E^n (\xi^n - E^n \xi^n)^2 = 0$.

Roughly speaking, if AAO exists, then, working with large portfolios, the investor can become infinitely rich (in the mean sense) with vanishing quadratic risk.

We say that the large financial market has *NAA property* if there are no asymptotic arbitrage opportunities for any subsequence of market models $\{M^{n'}\}$.

A simple but useful remark: the NAA property remains the same if we replace (a) in the definition of AAO by the weaker property $\limsup_n E^n \xi^n > 0$ (“if one can become rich, one can become infinitely rich”).

Let ρ_n be the L^2 -distance of R_T^{0n} from the unit, i.e.

$$\rho_n := \inf_{\xi \in R_T^{0n}} E^n (\xi - 1)^2,$$

Proposition 5.1 $\text{NAA} \Leftrightarrow \liminf_n \rho_n > 0$.

Proof ($\Rightarrow$) Assume that $\liminf_n \rho_n = 0$. This means (modulo passage to a subsequence) that there are $\xi^n \in R_T^{0n}$ such that $E^n (\xi^n - 1)^2 \rightarrow 0$. It follows from the identity

$$E^n (\xi^n - 1)^2 = D^n \xi^n + (E^n \xi^n - 1)^2$$

that $D^n \xi^n \rightarrow 0$ and $E^n \xi^n \rightarrow 1$, violating NAA.

($\Leftarrow$) Assume that NAA fails. This means (modulo passage to a subsequence) that there are $\xi^n \in R_T^{0n}$, $\xi^n \neq 0$, satisfying (a) and (b). It follows that

$$E^n (\xi^n)^2 = D^n \xi^n + (E^n \xi^n)^2 \rightarrow \infty.$$

Put $\tilde{\xi}^n := \xi^n / \sqrt{E^n (\xi^n)^2}$. Then $\tilde{\xi}^n \in R_T^{0n}$,

$$D^n \tilde{\xi}^n = (1/E^n (\xi^n)^2) D^n \xi^n \rightarrow 0$$

and

$$(E^n \tilde{\xi}^n)^2 = E^n (\tilde{\xi}^n)^2 - D^n \tilde{\xi}^n = 1 - D^n \tilde{\xi}^n \rightarrow 1.$$

Thus,

$$E^n(\tilde{\xi}^n - 1)^2 = D^n \tilde{\xi}^n + (E^n \tilde{\xi}^n - 1)^2 \rightarrow 0$$

and we get a contradiction. $\square$

Suppose now that in the n -th model we are given a d -dimensional square integrable price process (S_t^n) where $t \in \{0, T\}$. In general, $d = d(n)$. Suppose that $S_0^{in} = 1$ (this is just a choice of scales).

The crucial hypothesis of the k -factor APM is that there are k common sources of randomness affecting the prices of all securities and there are also individual sources of randomness related to each security. Specifically, we suppose that

$$\Delta S_T^{in} = \mu^{in} + \sum_{j=1}^k \zeta_j^n b_j^{in} + \eta^{in}, \quad i \leq d,$$

or, in vector notation,

$$\Delta S_T^n = \mu^n + \sum_{j=1}^k \zeta_j^n b_j^n + \eta^n.$$

Here $\mu^n, b_j^n \in \mathbf{R}^d$, the scalar random variables ζ_j^n with zero means are square integrable and the d -dimensional random vector η^n with zero mean has uncorrelated components (representing randomness proper to each asset).

Assume that $D\eta^{in} \leq C$ for all $i \leq d$ and $n \in \mathbf{N}$ for a certain constant C .

A (self-financing) portfolio strategy H^n is a vector in $\mathbf{R}^d$ such that

$$H^n \mathbf{1}_d := \sum_{i=1}^d H^{in} = 0.$$

At the final date the corresponding portfolio value is

$$V_T^n = H^n \Delta S_T^n = \sum_{i=1}^d H^{i,n} \Delta S_T^{in}$$

and these random variables form the set R_T^{0n} .

Lemma 5.2 *Let $\mathcal{L}_n$ be the linear subspace in $\mathbf{R}^d$ spanned by the set $\{\mathbf{1}_d, b_j^n, j \leq k\}$ and let c^n be the projection of μ^n onto $\mathcal{L}_n^\perp$. Then*

$$\text{NAA} \quad \Rightarrow \quad \sup_n |c^n| < \infty.$$

Proof Let a_n be a real number. The vector $H^n := a_n c^n$ (being orthogonal to $\mathbf{1}_d$) is a self-financing strategy with the corresponding terminal value

$$V_T^n = a_n |c^n|^2 + a_n c^n \eta^n.$$

It follows that

$$E^n V_T^n = a_n |c^n|^2,$$

$$D^n V_T^n = a_n^2 E (c^n \eta^n)^2 = a_n^2 \sum_{i=1}^d (c^{in})^2 D^n \eta^{in} \leq C a_n^2 |c^n|^2.$$

In particular, for $a_n = |c^n|^{-3/2}$ we have an asymptotic arbitrage opportunity for any subsequence along which $|c^n|$ converges to infinity. $\square$

As is easily seen from the proof, the conditions of the lemma are equivalent if $D^n \eta^{in} \geq \varepsilon > 0$ for all i and n .

Proposition 5.3 *Assume that NAA holds. Then there exist a constant A and real-valued sequences $\{r^n\}$, $\{g_j^n\}$, $j \leq k$, such that*

$$\left| \mu^n - r^n \mathbf{1}_d - \sum_{j=1}^k g_j^n b_j^n \right|^2 := \sum_{i=1}^d \left(\mu^{in} - r^n - \sum_{j=1}^k g_j^n b_j^{in} \right)^2 \leq A.$$

The assertion is an obvious corollary of the above lemma: the vector c^n is a difference of μ^n and the projection of μ^n onto $\mathcal{L}_n$; the latter is a linear combination of the generating vectors $\mathbf{1}_d, b_1^n, \dots, b_k^n$. Of course, if the generators are not linearly independent, the coefficients $r^n, g_1^n, \dots, g_k^n$ are not uniquely defined.

The most interesting case of the APM is the “stationary” one where all random variables “live” on the same probability space and do not depend on n . All model parameters also do not depend on n except the dimension $d = n$. In other words, we are given infinite-dimensional vectors $\mu = (\mu^1, \mu^2, \dots)$, $\eta = (\eta^1, \eta^2, \dots)$, etc., and the ingredients of the n -th model, μ^n, η^n , etc., are composed of the first n coordinates of these vectors. One can think that the “real-world” market has an infinite number of securities, enumerated somehow, and the agent uses the first n of them in his portfolios. That is, the increment of the n -dimensional price process in the n -th model is

$$\Delta S_T^i = \mu^i + \sum_{j=1}^k \zeta_j b_j^i + \eta^i, \quad i \leq n.$$

Theorem 5.4 *Assume that NAA holds. Then there are constants r and g_j , $j \leq k$, such that*

$$\sum_{i=1}^{\infty} \left(\mu^i - r - \sum_{j=1}^k g_j b_j^i \right)^2 < \infty.$$

Proof Let us consider the vector space spanned by the infinite-dimensional vectors $\mathbf{1}_\infty = (1, 1, \dots)$, $b_j = (b_j^1, b_j^2, \dots)$, $j \leq k$. Without loss of generality we may assume that $\mathbf{1}_\infty, b_j, j \leq l$, is a basis in this space. There is n_0 such that for every $n \geq n_0$ the vectors formed by the first n components of the latter are linearly independent. For every $n \geq n_0$ we define the set

$$K^n := \left\{ (r, g_1, \dots, g_l, 0, \dots, 0) \in \mathbf{R}^{k+1} : \sum_{i=1}^n \left(\mu^i - r - \sum_{j=1}^k g_j b_j^i \right)^2 \leq A \right\}$$

where choosing A as in Proposition 5.3 ensures that K^n is non-empty. Clearly, K^n is closed and $K^{n+1} \subseteq K^n$. It is easily seen that K^n is bounded (otherwise we could construct a linear relation between the vectors assumed to be linearly independent). Thus, the sets K^n are compact, $\bigcap_{n \geq n_0} K^n \neq \emptyset$, and the result follows. $\square$

In the case where the numéraire is a traded security, say, the first one (i.e. $\Delta S_T^{1n} = 0$) we can take $r^n = 0$ for all n in Proposition 5.3 and $r = 0$ in Theorem 5.4. To see this, we repeat the arguments above with “truncated” price vectors and strategies, the first component being excluded. In this specification an admissible strategy is just a vector from $\mathbf{R}^{d-1}$ and the projection onto the vector with unit coordinates is not needed.

To make the relation between CAPM and APM clear, let us consider the one-factor stationary model where the numéraire is a traded security and the increments of the risky asset (enumerating from zero) are of the following structure:

$$\begin{aligned} \Delta S_T^0 &= \mu_0 + b_0 \zeta, \\ \Delta S_T^i &= \mu_i + b_i \zeta + \eta_i, \quad i \geq 1. \end{aligned}$$

where all random variables ζ and η_i are uncorrelated and have zero means. Assume that $D\eta_i \leq C$. The 0-th asset plays a particular role: all other price movements are conditionally uncorrelated given ΔS_T^0 . It can be viewed as a kind of “market portfolio” or “market index”.

If there is no asymptotic arbitrage, then there exists a constant g such that

$$\sum_{i=0}^{\infty} (\mu_i - g b_i)^2 < \infty$$

i.e. $\mu_i = g b_i + u_i$ where $u_i \rightarrow 0$. If the residual u_0 is small, then $\mu_0 \approx g b_0$. We can use the latter relation to specify g and conclude that $\mu_i \approx \mu_0 \beta_i$ (at least, for sufficiently large i) with $\beta_i := b_i/b_0$. Of course, this reasoning is far from being rigorous: the empirical data, even being in accordance with APM, may or may not follow the conclusion of CAPM.

Note that the approach of APT is based on the assumption that the agents have certain risk-preferences and in the asymptotic setting they may accept the

possibility of large losses with small probabilities; the variance is taken as an appropriate measure of risk.

A specific feature of the classical APT is that it does not deal with the problem of existence of equivalent martingale measures which is the key point of the Fundamental Theorem of Asset Pricing. For a long time these two arbitrage theories were considered as unrelated. In [35] an approach was suggested which puts together basic ideas of both of them and allows us to solve the long-standing problem of extension of APT to the continuous-time setting. A brief account of its further development is given in the next subsections.

5.2 Asymptotic arbitrage and contiguity

The theory of large financial markets contains four principal ingredients: basic concepts, functional-analytic methods, probabilistic results, and analysis of specific models. The fundamentals of this theory were established in [35] where the definitions of asymptotic arbitrage of the first and the second kind were suggested. Assuming the uniqueness of equivalent martingale measures (i.e. the completeness) for each market model, the authors proved necessary and sufficient conditions for NAA1 and NAA2 in terms of contiguity of sequences of equivalent martingale measures and objective (“historical”) probabilities. A particular model of a “large Black–Scholes market” (where the price processes are correlated geometric Brownian motions) was investigated. It was shown that the boundedness condition similar to that of Ross–Huberman can be obtained as a direct application of the Liptser–Shiryaev criteria of contiguity in terms of the Hellinger processes. The restricting uniqueness hypothesis was removed by Klein and Schachermayer (see [45], [46], and [44]). They discovered the importance of duality methods of geometric functional analysis in the context of large financial markets and found non-trivial extensions of NAA1 and NAA2 criteria for the case of incomplete market models. These criteria were complemented in [37] by new ones. In particular, it was shown that the strong asymptotic arbitrage is equivalent to the complete asymptotic separability of the historic probabilities and equivalent martingale measures. Our presentation follows the latter paper where also several modifications of classical models were analyzed and necessary and sufficient conditions for absence of asymptotic arbitrage were obtained in terms of model specifications.

In the terminology of [37], a *large financial market* is a sequence of ordinary semimartingale models of a frictionless market $\{(\mathbf{B}^n, S^n, T^n)\}$, where $\mathbf{B}^n$ is a stochastic basis with the trivial initial σ -algebra. A semimartingale price process S^n takes values in $\mathbf{R}^d$ for some $d = d(n)$. To simplify notation we shall often omit the superscript for the time horizon.

We denote by $\mathcal{Q}^n$ the set of all probability measures Q^n equivalent to P^n such that S^n is a local martingale with respect to Q^n . It is assumed that each set $\mathcal{Q}^n$ of equivalent local martingale measures is non-empty.

We define a trading strategy on $(\mathbf{B}^n, S^n, T^n)$ as a predictable process H^n with values in $\mathbf{R}^d$ such that the stochastic integral with respect to the semimartingale S^n $H^n \cdot S^n$ is well-defined on $[0, T]$.

For a trading strategy H^n and an initial endowment x^n the value process

$$V^n = V(n, x^n, H^n) := x^n + H^n \cdot S^n.$$

A sequence V^n realizes asymptotic arbitrage of the first kind (AA1) if

- (1a) $V_t^n \geq 0$ for all $t \leq T$;
- (1b) $\lim_n V_0^n = 0$ (i.e. $\lim_n x^n = 0$);
- (1c) $\lim_n P^n(V_T^n \geq 1) > 0$.

A sequence V^n realizes asymptotic arbitrage of the second kind (AA2) if

- (2a) $V_t^n \leq 1$ for all $t \leq T$;
- (2b) $\lim_n V_0^n > 0$;
- (2c) $\lim_n P^n(V_T^n \geq \varepsilon) = 0$ for any $\varepsilon > 0$.

A sequence V^n realizes strong asymptotic arbitrage of the first kind (SAA1) if

- (3a) $V_t^n \geq 0$ for all $t \leq T$;
- (3b) $\lim_n V_0^n = 0$ (i.e. $\lim_n x^n = 0$);
- (3c) $\lim_n P^n(V_T^n \geq 1) = 1$.

One can continue and give also the definition SAA2. It is easy to understand that the existence of SAA1 implies the existence of SAA2 and vice versa (provided that there are no specific constraints). So existence criteria are the same in both cases.

A large security market $\{(\mathbf{B}^n, S^n, T^n)\}$ has *no asymptotic arbitrage of the first kind* (respectively, of the *second kind*) if for any subsequence (m) there are no value processes V^m realizing asymptotic arbitrage of the first kind (respectively, of the second kind) for $\{(\mathbf{B}^m, S^m, T^m)\}$.

To formulate the results we need to extend some notions from measure theory.

Let $\mathcal{Q} = \{Q\}$ be a family of probabilities on a measurable space $(\Omega, \mathcal{F})$. Define the upper and lower envelopes of measures from $\mathcal{Q}$ as the set functions with

$$\overline{\mathbf{Q}}(A) := \sup_{Q \in \mathcal{Q}} Q(A), \quad \underline{\mathbf{Q}}(A) := \inf_{Q \in \mathcal{Q}} Q(A), \quad A \in \mathcal{F}.$$

We say that $\mathcal{Q}$ is *dominated* if any element of $\mathcal{Q}$ is absolutely continuous with respect to some fixed probability measure.

In our setting, where for every n a family $\mathcal{Q}^n$ of equivalent local martingale measures is given, we use the obvious notations $\overline{\mathbf{Q}}^n$ and $\underline{\mathbf{Q}}^n$.

Generalizing in a straightforward way the well-known notion of contiguity to set functions other than measures, we introduce the following definitions:

The sequence (P^n) is *contiguous* with respect to $(\bar{Q}^n)$ (notation: $(P^n) \triangleleft (\bar{Q}^n)$) when the implication

$$\lim_{n \rightarrow \infty} \bar{Q}^n(A^n) = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} P^n(A^n) = 0$$

holds for any sequence $A^n \in \mathcal{F}^n$, $n \geq 1$.

Obviously, $(P^n) \triangleleft (\bar{Q}^n)$ if and only if the implication

$$\lim_{n \rightarrow \infty} \sup_{Q \in \mathcal{Q}^n} E_Q g^n = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} E_{P^n} g^n = 0$$

holds for any uniformly bounded sequence g^n of positive $\mathcal{F}^n$ -measurable random variables.

A sequence (P^n) is *asymptotically separable* from $(\bar{Q}^n)$ (notation: $(P^n) \triangle (\bar{Q}^n)$) if there exists a subsequence (m) with sets $A^m \in \mathcal{F}^m$ such that

$$\lim_{m \rightarrow \infty} \bar{Q}^m(A^m) = 0, \quad \lim_{m \rightarrow \infty} P^m(A^m) = 1.$$

Proposition 5.5 *The following conditions are equivalent:*

- (a) *there is no asymptotic arbitrage of the first kind (NAA1);*
- (b) $(P^n) \triangleleft (\bar{Q}^n)$;
- (c) *there exists a sequence $R^n \in \mathcal{Q}^n$ such that $(P^n) \triangleleft (R^n)$.*

Proof (b) $\Rightarrow$ (a) Let (V^n) be a sequence of value processes realizing asymptotic arbitrage of the first kind. For any $Q \in \mathcal{Q}^n$ the process V^n is a non-negative local Q -martingale, hence a Q -supermartingale, and

$$\sup_{Q \in \mathcal{Q}^n} E_Q V_T^n \leq \sup_{Q \in \mathcal{Q}^n} E_Q V_0^n = x^n \rightarrow 0$$

by (1b). Thus,

$$\bar{Q}^n(V_T^n \geq 1) := \sup_{Q \in \mathcal{Q}^n} Q(V_T^n \geq 1) \rightarrow 0$$

and, by contiguity $(P^n) \triangleleft (\bar{Q}^n)$, we have $P^n(V_T^n \geq 1) \rightarrow 0$ in contradiction to (1c).

(a) $\Rightarrow$ (b) Assume that (P^n) is not contiguous with respect to $(\bar{Q}^n)$. Taking, if necessary, a subsequence we can find sets $\Gamma^n \in \mathcal{F}^n$ such that $\bar{Q}^n(\Gamma^n) \rightarrow 0$, $P^n(\Gamma^n) \rightarrow \gamma$ as $n \rightarrow \infty$ where $\gamma > 0$. According to Proposition 4.7 the process

$$X_t^n = \text{ess sup}_{Q \in \mathcal{Q}^n} E_Q(I_{\Gamma^n} | \mathcal{F}_t^n)$$

is a supermartingale with respect to any $Q \in \mathcal{Q}^n$. By Theorem 4.6 it admits a decomposition $X^n = X_0^n + H^n \cdot S^n - A^n$ where A^n is an increasing process. Let

us show that $V^n := X_0^n + H^n \cdot S^n$ are value processes realizing AA1. Indeed, $V^n = X^n + A^n \geq 0$,

$$V_0^n = \sup_{Q \in \mathcal{Q}^n} E_Q I_{\Gamma^n} = \overline{\mathbf{Q}}^n(\Gamma^n) \rightarrow 0,$$

and

$$\lim_n P^n(V_T^n \geq 1) \geq \lim_n P^n(X_T^n \geq 1) = \lim_n P^n(X_T^n = 1) = \lim_n P^n(\Gamma^n) = \gamma > 0.$$

(b) $\Leftrightarrow$ (c) This relation follows from the convexity of $\mathcal{Q}^n$ and a general result given below. $\square$

Proposition 5.6 *Assume that for any $n \geq 1$ we are given a probability space $(\Omega^n, \mathcal{F}^n, P^n)$ with a dominated family $\mathcal{Q}^n$ of probability measures. Then the following conditions are equivalent:*

- (a) $(P^n) \triangleleft (\overline{\mathbf{Q}}^n)$;
- (b) *there is a sequence $R^n \in \text{conv } \mathcal{Q}^n$ such that $(P^n) \triangleleft (R^n)$;*
- (c) *the following equality holds:*

$$\lim_{\alpha \downarrow 0} \liminf_{n \rightarrow \infty} \sup_{Q \in \text{conv } \mathcal{Q}^n} H(\alpha, Q, P^n) = 1,$$

where $H(\alpha, Q, P) = \int (dQ)^\alpha (dP)^{1-\alpha}$ is the Hellinger integral of order $\alpha \in]0, 1[$.

The sequence of sets of probability measures $(\mathcal{Q}^n)$ is said to be *weakly contiguous with respect to (P^n)* (notation: $(\mathcal{Q}^n) \triangleleft_w (P^n)$) if for any $\varepsilon > 0$ there are $\delta > 0$ and a sequence of measures $Q^n \in \mathcal{Q}^n$ such that for any sequence $A^n \in \mathcal{F}^n$ with the property $\limsup_n P^n(A^n) < \delta$ we have $\limsup_n Q^n(A^n) < \varepsilon$.

For the case where the sets $\mathcal{Q}^n$ are singletons containing only the measure Q^n , the relation $(\mathcal{Q}^n) \triangleleft_w (P^n)$ means simply that $(Q^n) \triangleleft (P^n)$.

Obviously, the property $(\mathcal{Q}^n) \triangleleft_w (P^n)$ can be formulated in terms of random variables:

for any $\varepsilon > 0$ there are $\delta > 0$ and a sequence of measures $Q^n \in \mathcal{Q}^n$ such that for any sequence of $\mathcal{F}^n$ -measurable random variables g^n taking values in the interval $[0, 1]$ with the property $\limsup_n E_{P^n} g^n < \delta$, we have $\limsup_n E_{Q^n} g^n < \varepsilon$.

Proposition 5.7 *The following conditions are equivalent:*

- (a) *there is no asymptotic arbitrage of the second kind (NAA2);*
- (b) $(\mathbf{Q}^n) \triangleleft (P^n)$;
- (c) $(\mathcal{Q}^n) \triangleleft_w (P^n)$.

The proof of Proposition 5.7 is similar to that of Proposition 5.5. Notice that the conditions (b) in both statements look rather symmetric in contrast to the conditions (c). In general, the condition (b) of Proposition 5.7 may hold though a sequence $Q^n \in \mathcal{Q}^n$ such that $(Q^n) \triangleleft (P^n)$ does not exist (see an example in [45]). The reason is that the set functions $\overline{\mathbf{Q}}$ and $\underline{\mathbf{Q}}$ are of a radically different nature.

The following assertion gives criteria of existence of strong asymptotic arbitrage.

Proposition 5.8 *The following conditions are equivalent:*

- (a) *there is SAAI;*
- (b) $(P^n) \triangleleft (\overline{\mathbf{Q}}^n)$;
- (c) $(\underline{\mathbf{Q}}^n) \triangleleft (P^n)$;
- (d) $(P^n) \triangleleft (Q^n)$ for any sequence $Q^n \in \mathcal{Q}^n$.

Let P and $\tilde{P}$ be two equivalent probability measures on a stochastic basis $\mathbf{B}$ and let $R := (P + \tilde{P})/2$. Let us denote by z and $\tilde{z}$ the density processes of P and $\tilde{P}$ with respect to R . For arbitrary $\alpha \in]0, 1[$ the process $Y = Y(\alpha) := z^\alpha \tilde{z}^{1-\alpha}$ is a R -supermartingale admitting the multiplicative decomposition $Y = M\mathcal{E}(-h)$ where $M = M(\alpha)$ is a local Q -martingale, $\mathcal{E}$ is the Doléan–Dade exponential, and $h = h(\alpha, P, \tilde{P})$ is an increasing predictable process, $h_0 = 0$, called the Hellinger process of order α . These Hellinger processes play an important role in criteria of absolute continuity and, more generally, contiguity of probability measures, see [28] for details.

In the abstract setting of Proposition 5.6 when the probability spaces are equipped with filtrations (i.e. they are stochastic bases) we have the following results which are helpful in analysis of particular models arising in mathematical finance.

Theorem 5.9 *The following conditions are equivalent:*

- (a) $(P^n) \triangleleft (\overline{\mathbf{Q}}^n)$;
- (b) for all $\varepsilon > 0$

$$\lim_{\alpha \downarrow 0} \limsup_{n \rightarrow \infty} \inf_{Q \in \text{conv } \mathcal{Q}^n} P^n(h_\infty(\alpha, Q, P^n) \geq \varepsilon) = 0.$$

Theorem 5.10 *Assume that the family $\mathcal{Q}^n$ is convex and dominated for any n . Then the following conditions are equivalent:*

- (a) $(\underline{\mathbf{Q}}^n) \triangleleft (P^n)$;
- (b) for all $\varepsilon > 0$

$$\lim_{\alpha \downarrow 0} \limsup_{n \rightarrow \infty} \inf_{Q \in \mathcal{Q}^n} Q(h_\infty(\alpha, P^n, Q) \geq \varepsilon) = 0.$$

The concept of contiguity is useful in relation with an important question whether the option prices calculated in “approximating” models converge to the “true” option price, see [24] and [58].

5.3 A large BS-market

Let $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t), P)$ be a stochastic basis with a countable set of independent one-dimensional Wiener processes w^i , $i \in \mathbf{Z}_+$, $\mathbf{w}^n = (w^0, \dots, w^n)$, and let $\mathbf{F}^n = (\mathcal{F}_t^n)$ be a filtration generated by $\mathbf{w}^n$. For simplicity, assume that T is fixed.

The behavior of the stock prices is described by the following stochastic differential equations:

$$dX_t^0 = \mu_0 X_t^0 dt + \sigma_0 X_t^0 dw_t^0,$$

$$dX_t^i = \mu_i X_t^i dt + \sigma_i X_t^i (\gamma_i dw_t^0 + \bar{\gamma}_i dw_t^i), \quad i \in \mathbf{N},$$

with (deterministic strictly positive) initial points X_0^i . Here γ_i is a function taking values in $[0, 1]$ and $\gamma_i^2 + \bar{\gamma}_i^2 = 1$. We assume that $\mu_i, \sigma_i \in L^2[0, T]$ and $\sigma_i > 0$.

Notice that the process ξ^i with

$$d\xi_t^i = \gamma_i dw_t^0 + \bar{\gamma}_i dw_t^i, \quad \xi_0^i = 0,$$

is a Wiener process. Thus, in the case of constant coefficients price processes are geometric Brownian motions as in the classical case of Black and Scholes. The model is designed to reflect the fact that in the market there are two different types of randomness: the first type is proper to each stock while the second one originates from some common source and it is accumulated in a “stock index” (or “market portfolio”) whose evolution is described by the first equation. Set

$$\beta_i := \frac{\gamma_i \sigma_i}{\sigma_0} = \frac{\gamma_i \sigma_i \sigma_0}{\sigma_0^2}.$$

In the case of deterministic coefficients, β_i is a well-known measure of risk which is the covariance between the return on the asset with number i and the return on the index, divided by the variance of the return on the index.

Let $\mathbf{b}_n(t) := (b_0(t), b_1(t), \dots, b_n(t))$ where

$$b_0 := -\frac{\mu_0}{\sigma_0}, \quad b_i := \frac{\beta_i \mu_0 - \mu_i}{\sigma_i \bar{\gamma}_i}.$$

Assume that for every n

$$\int_0^T |\mathbf{b}_n(t)|^2 dt < \infty.$$

We consider the stochastic basis $\mathbf{B}^n = (\Omega, \mathcal{F}, \mathbf{F}^n = (\mathcal{F}_t^n)_{t \leq T}, P^n)$ with the $(n+1)$ -dimensional semimartingale $S^n := (X_t^0, X_t^1, \dots, X_t^n)$ and $P^n := P|_{\mathcal{F}_T^n}$. The

sequence $\{(\mathbf{B}^n, S^n, T)\}$ is a large security market. In our case each $(\mathbf{B}^n, S^n, T)$ is a model of a complete market and the set $\mathcal{Q}^n$ is a singleton which consists of the measure $Q^n = Z_T(\mathbf{b}_n)P^n$ where

$$Z_T(\mathbf{b}_n) := \exp \left\{ \int_0^T (\mathbf{b}_n(t), d\mathbf{w}_t^n) - \frac{1}{2} \int_0^T |\mathbf{b}_n(t)|^2 dt \right\}.$$

The Hellinger process has an explicit expression

$$h(\alpha, Q^n, P^n) = \frac{\alpha(1-\alpha)}{2} \int_0^T \left[\left(\frac{\mu_0}{\sigma_0} \right)^2 + \sum_{i=1}^n \left(\frac{\mu_i - \beta_i \mu_0}{\sigma_i \bar{\gamma}_i} \right)^2 \right] ds.$$

As a corollary of Theorem 5.9 we have

Proposition 5.11 *The condition NAA1 holds if and only if*

$$\int_0^T \left[\left(\frac{\mu_0}{\sigma_0} \right)^2 + \sum_{i=1}^n \left(\frac{\mu_i - \beta_i \mu_0}{\sigma_i \bar{\gamma}_i} \right)^2 \right] ds < \infty.$$

In fact, in this model both conditions NAA1 and NAA2 hold simultaneously.

In the particular case of constant coefficients, finite T , and $0 < c \leq \sigma_i \bar{\gamma}_i \leq C$ we get that the property NAA1 holds if and only if

$$\sum_{i=1}^{\infty} (\mu_i - \beta_i \mu_0)^2 < \infty,$$

i.e. the Huberman–Ross boundedness is fulfilled.

5.4 One-factor APM revisited

We consider the “stationary” one-factor model of the following specific structure (cf. with the model given at the end of Subsection 5.1). Let $(\epsilon_i)_{i \geq 0}$ be independent random variables given on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in a finite interval $[-N, N]$, $E\epsilon_i = 0$, $E\epsilon_i^2 = 1$. At time zero all asset prices $S_0^i = 1$ and

$$\begin{aligned} \Delta S_T^0 &= 1 + \mu_0 + \sigma_0 \epsilon_0, \\ \Delta S_T^i &= 1 + \mu_i + \sigma_i (\gamma_i \epsilon_0 + \bar{\gamma}_i \epsilon_i), \quad i \geq 1. \end{aligned}$$

The coefficients here are deterministic, $\sigma_i > 0$, $\bar{\gamma}_i > 0$ and $\gamma_i^2 + \bar{\gamma}_i^2 = 1$. The asset with number zero is interpreted as a market portfolio, γ_i is the correlation coefficient between the rate of return for the market portfolio and the rate of return for the asset with number i .

For $n \geq 0$ we consider the stochastic basis $\mathbf{B}^n = (\Omega, \mathcal{F}^n, \mathbf{F}^n = (\mathcal{F}_t^n)_{t \in [0,1]}, P^n)$ with the $(n+1)$ -dimensional random process $\mathbf{S}^n := (S_t^0, S_t^1, \dots, S_t^n)_{t \in [0,1]}$ where

$\mathcal{F}_0^n$ is the trivial σ -algebra, $\mathcal{F}_1^n = \mathcal{F}^n := \sigma\{\epsilon_0, \dots, \epsilon_n\}$, and $P^n = P|_{\mathcal{F}^n}$. According to our definition, the sequence $\mathbf{M} = \{(\mathbf{B}^n, \mathbf{S}^n, 1)\}$ is a large security market.

Let $\beta_i := \gamma_i \sigma_i / \sigma_0$,

$$b_0 := -\frac{\mu_0}{\sigma_0}, \quad b_i := \frac{\mu_0 \beta_i - \mu_i}{\sigma_i \bar{\gamma}_i}, \quad i \geq 1.$$

It is convenient to rewrite the price increments as follows:

$$\begin{aligned} \Delta S_T^0 &= 1 + \sigma_0(\epsilon_0 - b_0), \\ \Delta S_T^i &= 1 + \sigma_i \gamma_i(\epsilon_0 - b_0) + \sigma_i \bar{\gamma}_i(\epsilon_i - b_i), \quad i \geq 1. \end{aligned}$$

The set $\mathcal{Q}^n$ of equivalent martingale measures for $\mathbf{S}^n$ has a very simple description: $Q \in \mathcal{Q}^n$ iff $Q \sim P^n$ and

$$E_Q(\epsilon_i - b_i) = 0, \quad 0 \leq i \leq n,$$

i.e. the b_i are mean values of ϵ_i under Q . Obviously, $\mathcal{Q}^n \neq \emptyset$ iff $P(\epsilon_i > b_i) > 0$ and $P(\epsilon_i < b_i) > 0$ for all $i \leq n$.

As usual, we assume that $\mathcal{Q}^n \neq \emptyset$ for all n ; this implies, in particular, that $|b_i| < N$.

Let F_i be the distribution function of ϵ_i . Put

$$\underline{s}_i := \inf\{t : F_i(t) > 0\}, \quad \bar{s}_i := \inf\{t : F_i(t) = 1\},$$

$\underline{d}_i := b_i - \underline{s}_i$, $\bar{d}_i := \bar{s}_i - b_i$, and $d_i := \underline{d}_i \wedge \bar{d}_i$. In other words, d_i is the distance from b_i to the end points of the interval $[\underline{s}_i, \bar{s}_i]$.

Proposition 5.12 *The following assertions hold:*

- (a) $\inf_i d_i = 0 \Leftrightarrow \text{SAA} \Leftrightarrow (P^n) \triangle (\overline{\mathbf{Q}}^n)$,
- (b) $\inf_i d_i > 0 \Leftrightarrow \text{NAA1} \Leftrightarrow (P^n) \triangleleft (\overline{\mathbf{Q}}^n)$,
- (c) $\limsup_i |b_i| = 0 \Leftrightarrow \text{NAA2} \Leftrightarrow (\underline{\mathbf{Q}}^n) \triangleleft (P^n)$.

The hypothesis that the distributions of ϵ_i have finite support is important: it excludes the case where the value of every non-trivial portfolio is negative with positive probability. For the proof of this result, we send the reader to the original paper [37].

Appendix: Facts from convex analysis

1 By definition, a subset K in $\mathbf{R}^n$ (or in a linear space X) is a *cone* if it is convex and stable under multiplication by the non-negative constants. It defines the partial ordering:

$$x \geq_K y \quad \Leftrightarrow \quad x - y \in K;$$

in particular, $x \geq_K 0$ means that $x \in K$.

A closed cone K is *proper* if the linear space $F := K \cap (-K) = \{0\}$, i.e. if the relations $x \geq_K$ and $x \leq_K 0$ imply that $x = 0$.

Let K be a closed cone and let $\pi : \mathbf{R}^n \rightarrow \mathbf{R}^n / F$ be the canonical mapping onto the quotient space. Then πK is a proper closed cone.

For a set C we denote by $\text{cone } C$ the set of all conic combinations of elements of C . If C is convex then $\text{cone } C = \bigcup_{\lambda \geq 0} \lambda C$.

Let K be a cone. Its *dual positive cone*

$$K^* := \{z \in \mathbf{R}^n : zx \geq 0 \forall x \in K\}$$

is closed (the *dual cone* K° is defined using the opposite inequality, i.e. $K^\circ = -K^*$); K is closed if and only if $K = K^{**}$.

We use the notations $\text{int } K$ for the interior of K and $\text{ri } K$ for the relative interior (i.e. the interior in $K - K$, the linear subspace generated by K).

A closed cone K in the Euclidean space $\mathbf{R}^n$ is proper if and only if there exists a compact convex set C such that $0 \notin C$ and $K = \text{cone } C$. One can take as C the convex hull of the intersection of K with the unit sphere $\{x \in \mathbf{R}^n : |x| = 1\}$.

A closed cone K is proper if and only if $\text{int } K^* \neq \emptyset$.

We have

$$\text{ri } K^* = \{w : wx > 0 \forall x \in K, x \neq 0\};$$

in particular, if K is proper then

$$\text{int } K^* = \{w : wx > 0 \forall x \in K, x \neq 0\}.$$

By definition, the cone K is *polyhedral* if it is the intersection of a finite number of half-spaces $\{x : p_i x \geq 0\}$, $p_i \in \mathbf{R}^n$, $i = 1, \dots, N$.

The Farkas–Minkowski–Weyl theorem:

a cone is polyhedral if and only if it is finitely generated.

The following result is a direct generalization of the Stiemke lemma.

Lemma A.1 *Let K and R be closed cones in $\mathbf{R}^n$. Assume that K is proper. Then*

$$R \cap K = \{0\} \quad \Leftrightarrow \quad (-R^*) \cap \text{int } K^* \neq \emptyset.$$

Proof ($\Leftarrow$) The existence of w such that $wx \leq 0$ for all $x \in R$ and $wy > 0$ for all $y \in K \setminus \{0\}$ obviously implies that R and $K \setminus \{0\}$ are disjoint.

($\Rightarrow$) Let C be a convex compact set such that $0 \notin C$ and $K = \text{cone } C$. By the separation theorem (for the case where one set is closed and another is compact)

there is a non-zero $z \in \mathbf{R}^n$ such that

$$\sup_{x \in R} zx < \inf_{y \in C} zy.$$

Since R is a cone, the left-hand side of this inequality is zero, hence $z \in -R^*$ and, also, $zy > 0$ for all $y \in C$. The latter property implies that $zy > 0$ for $z \in K$, $z \neq 0$, and we have $z \in \text{int } K$. $\square$

In the classical Stiemke lemma $K = \mathbf{R}_+^n$ and $R = \{y \in \mathbf{R}^n : y = Bx, x \in \mathbf{R}^d\}$ where B is a linear mapping. Usually, it is formulated as the alternative:

either there is $x \in \mathbf{R}^d$ such that $Bx \geq_K 0$ and $Bx \neq 0$ or there is $y \in \mathbf{R}^n$ with strictly positive components such that $B^*y = 0$.

Lemma A.1 can be slightly generalized.

Let J_1 be the natural projection of $\mathbf{R}^n$ onto $\mathbf{R}^n/F$.

Theorem A.2 *Let K and R be closed cones in $\mathbf{R}^n$. Assume that the cone πR is closed. Then*

$$R \cap K \subseteq F \quad \Leftrightarrow \quad (-R^*) \cap \text{ri } K^* \neq \emptyset.$$

Proof It is easy to see that $\pi(R \cap K) = \pi R \cap \pi K$ and, hence,

$$R \cap K \subseteq F \quad \Leftrightarrow \quad \pi R \cap \pi K = \{0\}.$$

By Lemma A.1

$$\pi R \cap \pi K = \{0\} \quad \Leftrightarrow \quad (-\pi R)^* \cap \text{int } (\pi K)^* \neq \emptyset.$$

Since $(\pi R)^* = \pi^{*-1} R^*$ and $\text{int } (\pi K)^* = \pi^{*-1}(\text{ri } K^*)$, the condition in the right-hand side can be written as

$$\pi^{*-1}((-R^*) \cap \text{ri } K^*) \neq \emptyset$$

or, equivalently,

$$(-R^*) \cap \text{ri } K^* \cap \text{Im } \pi^* \neq \emptyset.$$

But $\text{Im } \pi^* = (K \cap (-K))^* = K^* - K^* \supseteq \text{ri } K^*$ and we get the result. $\square$

Notice that if R is polyhedral then πR is also polyhedral, hence closed.

2 The following result is referred to as the Kreps–Yan theorem, see [48], [63], [5]. It holds for arbitrary $p \in [1, \infty]$, $p^{-1} + q^{-1} = 1$, but the cases $p = 1$ and $p = \infty$ are the most important.

Theorem A.3 Let $\mathcal{C}$ be a convex cone in L^p closed in $\sigma\{L^p, L^q\}$, containing $-L_+^p$ and such that $\mathcal{C} \cap L_+^p = \{0\}$. Then there is a $\tilde{P} \sim P$ with $d\tilde{P}/dP \in L^q$ such that $\tilde{E}\xi \leq 0$ for all $\xi \in \mathcal{C}$.

Proof By the Hahn–Banach theorem any non-zero $x \in L_+^p := L^p(\mathbf{R}_+, \mathcal{F})$ can be separated from $\mathcal{C}$: there is a $z_x \in L^q$ such that $Ez_x x > 0$ and $Ez_x \xi \leq 0$ for all $\xi \in \mathcal{C}$. Since $\mathcal{C} \supseteq -L_+^p$, the latter property yields that $z_x \geq 0$; we may assume $\|z_x\|_q = 1$. By the Halmos–Savage lemma the dominated family $\{P_x = z_x P : x \in L_+^p, x \neq 0\}$ contains a countable equivalent family $\{P_{x_i}\}$. But then $z := \sum 2^{-i} z_{x_i} > 0$ and we can take $\tilde{P} := zP$. $\square$

Recall that the Halmos–Savage lemma, though important, is, in fact, very simple. It suffices to prove its claim for the case of a convex family (in our situation we even have this property). A family $\{P_{x_i}\}$ such that the sequence $I_{\{z_{x_i} > 0\}}$ increases to $\text{ess sup } I_{\{z_x > 0\}}$ (existing because of convexity) meets the requirement.

The above theorem has the following “purely geometric” version, [5].

Theorem A.4 Suppose J and K are non-empty convex cones in a separable Banach space X such that $J \cap \overline{K - J} = \{0\}$. Then there is a continuous linear functional z such that $zx > 0 \forall x \in J$ and $zx \leq 0 \forall x \in K$.

The first step of the proof is the same as of the previous theorem: the separation of single points allows us to construct the set of $\{z_x \in X', x \in K\}$ with unit norms. The second step is to select a countable weak* dense subset. This can be done because the separability of X implies that the weak*-topology on the unit ball of X' (always weak* compact) is metrizable. For the Lebesgue spaces the separability means that the σ -algebra is countably generated. Specific properties of these spaces allow us, by means of the Halmos–Savage lemma, to avoid such an unpleasant assumption on the σ -algebra.

References

- [1] Ansel, J.-P. and Stricker, C. (1994), Couverture des actifs contingents. *Ann. Inst. Henri Poincaré* **30**, 2, 303–15.
- [2] Bouchard-Denize, B. and Touzi, N. (2001), Explicit solution of the multivariate super-replication problem under transaction costs. Preprint.
- [3] Chamberlain, G. (1983), Funds, factors, and diversification in arbitrage pricing models. *Econometrica* **51**, 5, 1305–23.
- [4] Chamberlain, G. and Rothschild, M. (1983), Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51**, 5, 1281–304.
- [5] Clark, S.A. (1992), The valuation problem in arbitrage price theory. *J. Math. Economics* **22**, 463–78.
- [6] Cvitanic, J. and Karatzas, I. (1996), Hedging and portfolio optimization under transaction costs: a martingale approach. *Mathematical Finance* **6**, 2, 133–65.

- [7] Cvitanić, J., Pham, H. and Touzi, N. (1999), A closed form solution to the problem of super-replication under transaction costs. *Finance and Stochastics* **3**, 1, 35–54.
- [8] Dalang, R.C., Morton, A. and Willinger, W. (1990), Equivalent martingale measures and no-arbitrage in stochastic securities market model. *Stochastics and Stochastic Reports* **29**, 185–201.
- [9] Davis, M.H.A. and Clark, J.M.C. (1994), A note on super-replicating strategies. *Philos. Trans. Roy. Soc. London A* **347**, 485–94.
- [10] Delbaen, F. (1992), Representing martingale measures when asset prices are continuous and bounded. *Mathematical Finance* **2**, 107–30.
- [11] Delbaen, F., Kabanov, Yu.M and Valkeila, S. (2001), Hedging under transaction costs in currency markets: a discrete-time model. *Mathematical Finance*. To appear.
- [12] Delbaen, F. and Schachermayer, W. (1994), A general version of the fundamental theorem of asset pricing. *Math. Annalen* **300**, 463–520.
- [13] Delbaen, F. and Schachermayer, W. (1999), A compactness principle for bounded sequence of martingales with applications. *Proceedings of the Seminar of Stochastic Analysis, Random Fields and Applications*, 1999.
- [14] Delbaen, F. and Schachermayer, W. (1998), The fundamental theorem of asset pricing for unbounded stochastic processes. *Math. Annalen* **312**, 215–50.
- [15] Dellacherie, C. and Meyer, P.-A. *Probabilités et Potentiel*. Hermann, Paris, 1980.
- [16] El Karoui, N. and Quenez, M.-C. (1995), Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal on Control and Optimization* **33**, 1, 27–66.
- [17] Émery, M. (1979), Une topologie sur l'espace de semimartingales. Séminaire de Probabilités XIII. *Lect. Notes Math.*, **721**, 260–80.
- [18] Föllmer, H. and Kabanov, Yu.M. (1998), Optional decomposition and Lagrange multipliers. *Finance and Stochastics* **2**, 1, 69–81.
- [19] Föllmer, H. and Kabanov, Yu.M. (1996), Optional decomposition theorems in discrete time. *Atti del convegno in onore di Oliviero Lessi, Padova, 25–26 marzo 1996*, 47–68.
- [20] Föllmer, H. and Kramkov, D.O. (1997), Optional decomposition theorem under constraints. *Probability Theory and Related Fields* **109**, 1, 1–25.
- [21] Gordan, P. (1873), Über di Auflösung linearer Gleichungen mit reelen Koeffizienten. *Math. Annalen* **6**, 23–8.
- [22] Hall, P. and Heyde, C.C. *Martingale Limit Theory and Its Applications*. Academic Press, New York, 1980.
- [23] Harrison, M. and Pliska, S. (1981), Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* **11**, 215–60.
- [24] Hubalek, F. and Schachermayer, W. (1998), When does convergence of asset price processes imply convergence of option prices? *Mathematical Finance* **8**, 4, 215–33.
- [25] Huberman, G. (1982), A simple approach to arbitrage pricing theory. *Journal of Economic Theory* **28**, 1, 183–91.
- [26] Ingersoll, J.E., Jr. (1984), Some results in the theory of arbitrage pricing. *Journal of Finance* **39**, 1021–39.
- [27] Ingersoll, J.E., Jr. *Theory of Financial Decision Making*. Rowman and Littlefield, 1989.
- [28] Jacod, J. and Shiryaev, A.N. *Limit Theorems for Stochastic Processes*. Springer, Berlin–Heidelberg–New York, 1987.
- [29] Jacod, J. and Shiryaev, A.N. (1998), Local martingales and the fundamental asset pricing theorem in the discrete-time case. *Finance and Stochastics* **2**, 3, 259–73.
- [30] Jouini, E. and Kallal, H. (1995), Martingales and arbitrage in securities markets with

transaction costs. *J. Economic Theory* **66**, 178–97.

- [31] Jouini, E. and Kallal, H. (1995), Arbitrage in securities markets with short sale constraints. *Mathematical Finance* **5**, 3, 197–232.
- [32] Jouini, E. and Kallal, H. (1999), Viability and equilibrium in securities markets with frictions. *Mathematical Finance* **9**, 3, 275–92.
- [33] Kabanov, Yu.M. On the FTAP of Kreps–Delbaen–Schachermayer. *Statistics and Control of Random Processes. The Liptser Festschrift. Proceedings of Steklov Mathematical Institute Seminar*, World Scientific, 1997, 191–203.
- [34] Kabanov, Yu.M. (1999), Hedging and liquidation under transaction costs in currency markets. *Finance and Stochastics* **3**, 2, 237–48.
- [35] Kabanov, Yu. M. and Kramkov, D.O. (1994), Large financial markets: asymptotic arbitrage and contiguity. *Probability Theory and its Applications* **39**, 1, 222–9.
- [36] Kabanov, Yu.M. and Kramkov, D.O. (1994), No-arbitrage and equivalent martingale measure: an elementary proof of the Harrison–Pliska theorem. *Probability Theory and Its Applications*, **39**, 3, 523–7.
- [37] Kabanov, Yu.M. and Kramkov, D.O. (1998), Asymptotic arbitrage in large financial markets. *Finance and Stochastics* **2**, 2, 143–72.
- [38] Kabanov, Yu.M. and Last, G. (2001a), Hedging in a model with transaction costs. Preprint.
- [39] Kabanov, Yu.M. and Last, G. (2001b), Hedging under transaction costs in currency markets: a continuous-time model. *Mathematical Finance*. To appear.
- [40] Kabanov, Yu.M., Liptser, R.Sh. and Shirayev, A.N. (1981), On the variation distance for probability measures defined on a filtered space. *Probability Theory and Related Fields* **71**, 19–36.
- [41] Kabanov, Yu.M. and Stricker, Ch. (2001a), The Harrison–Pliska arbitrage pricing theorem under transaction costs. *J. Math. Econ.* To appear.
- [42] Kabanov, Yu.M. and Stricker Ch. (2001b), A teachers’ note on no-arbitrage criteria. Séminaire de Probabilités. To appear.
- [43] Kabanov, Yu.M., Stricker, Ch. (2001c), On equivalent martingale measures with bounded densities. Séminaire de Probabilité. To appear.
- [44] Klein, I. (2001), A fundamental theorem of asset pricing for large financial markets. Preprint.
- [45] Klein, I. and Schachermayer, W. (1996), Asymptotic arbitrage in non-complete large financial markets. *Probability Theory and its Applications* **41**, 4, 927–34.
- [46] Klein, I. and Schachermayer, W. (1996), A quantitative and a dual version of the Halmos–Savage theorem with applications to mathematical finance. *Annals of Probability* **24**, 2, 867–81.
- [47] Kramkov, D.O. (1996), Optional decomposition of supermartingales and hedging in incomplete security markets. *Probability Theory and Related Fields* **105**, 4, 459–79.
- [48] Kreps, D.M. (1981), Arbitrage and equilibrium in economies with infinitely many commodities. *J. Math. Economics* **8**, 15–35.
- [49] Levental, S. and Skorohod, A.V. (1997), On the possibility of hedging options in the presence of transaction costs. *The Annals of Applied Probability* **7**, 410–43.
- [50] Mémin, J. (1980), Espace de semimartingales et changement de probabilité. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.*, **52**, 9–39.
- [51] Pshenychnyi, B.N. *Convex Analysis and Extremal Problems*. Nauka, Moscow, 1980 (in Russian).
- [52] Rockafellar, R.T. *Convex Analysis*. Princeton University Press, Princeton, 1970.
- [53] Rogers, L.C.G. (1994), Equivalent martingale measures and no-arbitrage. *Stochastic and Stochastics Reports* **51**, 41–51.

- [54] Ross, S.A. (1976), The arbitrage theory of asset pricing. *Journal of Economic Theory* **13**, 1, 341–60.
- [55] Sin, C.A. Strictly local martingales and hedge ratios on stochastic volatility models. PhD-dissertation, Cornell University, 1996.
- [56] Schachermayer, W. (1992), A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time. *Insurance: Mathematics and Economics* **11**, 249–57.
- [57] Shiryaev, A.N. *Probability*. Springer, Berlin–Heidelberg–New York, 1984.
- [58] Shiryaev, A.N. *Essentials of Stochastic Finance*. World Scientific, Singapore, 1999.
- [59] Soner, H.M., Shreve, S.E. and Cvitanić, J. (1995), There is no non-trivial hedging portfolio for option pricing with transaction costs. *The Annals of Applied Probability* **5**, 327–55.
- [60] Stricker, Ch. (1990), Arbitrage et lois de martingale. *Annales de l'Institut Henri Poincaré. Probabilité et Statistiques* **26**, 3, 451–60.
- [61] Schrijver, A. *Theory of Linear and Integer Programming*. Wiley, 1986.
- [62] Stiemke, E. (1915), Über positive Lösungen homogener linearer Gleichungen. *Math. Annalen* **76**, 340–2.
- [63] Yan, J.A. (1980), Caractérisation d'une classe d'ensembles convexes de L^1 et H^1 . Séminaire de Probabilités XIV. *Lect. Notes Math.*, **784**, 260–80.

2

Market Models with Frictions: Arbitrage and Pricing Issues

Elyès Jouini and Clotilde Napp

1 Introduction

The Fundamental Theorem of Asset Pricing, which originates in the Arrow–Debreu model (Debreu (1959)) and is further formalized in (among others) Harrison and Kreps (1979), Kreps (1981), Harrison and Pliska (1981), Duffie and Huang (1986), Dybvig and Ross (1987), Dalang, Morton and Willinger (1989), Back and Pliska (1990), Stricker (1990), Delbaen (1992), Lakner (1993) and Delbaen and Schachermayer (1994, 1998), asserts that the absence of free lunch in a frictionless (and complete) securities market model is equivalent to the existence of an equivalent martingale measure for the normalized securities price processes. The only arbitrage free and viable pricing rule on the set of contingent claims, which is a linear space, is then equal to the expected value with respect to the unique equivalent martingale measure.

In this chapter, we study some foundational issues in the theory of asset pricing in general models with flows as well as in securities market models with frictions. We consider financial models, where any investment opportunity is described by the cash flow that it generates. For instance, in such models, the investment opportunity, which consists, in a perfect financial model, of buying at time t_1 one unit of a risky asset, whose price process is given by $(S_t)_{t \geq 0}$, and selling at time t_2 with $t_1 \leq t_2$ the unit bought, is described by the process $(\Phi_t)_{t \geq 0}$ which is null outside $\{t_1, t_2\}$ and which satisfies $\Phi_{t_1} = -S_{t_1}$ and $\Phi_{t_2} = S_{t_2}$.

Sections 2 and 3 deal with a convex cone framework, i.e. a framework where the set of all available investments consists of a convex cone. A large class of imperfect market models, that we shall denote by $\mathfrak{J}$, can fit in this framework: models with imperfections on the numéraire like no borrowing or different borrowing and lending rates, models with dividends, short-sale constraints, convex cone constraints, proportional transaction costs.

Section 2 is devoted to the characterization of the no-free-lunch assumption first in a general convex cone framework with flows, then in all the models with imperfections belonging to $\mathfrak{J}$, and is taken from Jouini and Napp (2001) and Napp (2000). We consider first a quite general model; the investment opportunities are not specifically related to the buying and selling of securities on a financial market. The time horizon is not supposed to be finite. The framework is the one of continuous time. We don't assume that there exists a numéraire, enabling investors to transfer money from one date to another, and even if such possibilities exist, we do not assume that the lending rate is equal to the borrowing rate or that we have possibilities to borrow. It is proved that the absence of free lunch in a general convex cone framework with flows is essentially equivalent to the existence of a discount process such that the "net present value" of any investment opportunity is nonpositive. This result is then applied to obtain the Fundamental Theorem of Asset Pricing for all cases of market imperfections in $\mathfrak{J}$. In each case, we find that there is no free lunch if and only if a given specific convex set of discount processes is nonempty. For instance, in the case with short-sale constraints, we find that the absence of free lunch is equivalent to the existence of a discount process such that the discounted price process of any security that cannot be sold short (resp. that can only be sold short) is a supermartingale (resp. a submartingale).

Section 3 is devoted to pricing issues first in a general convex cone framework with flows, then in all the models with imperfections belonging to $\mathfrak{J}$, and is taken from Napp (2000). Section 3.1 is in the spirit of Harrison and Kreps (1979); we generalize existing results by considering general investment flows, and by taking almost any kind of imperfection into account. We consider a "primitive" market, consisting of a certain set of investment opportunities and we want to give a price to an additional contingent flow by using arbitrage considerations. More precisely, we define an admissible price for an additional contingent flow Φ as a price which is compatible with the assumption of no-arbitrage (or no free lunch) in the "full" market consisting of our primitive market and Φ . For a general contingent flow, we obtain an interval of admissible prices, which is given by the "net present value" of the flow under all admissible discount processes. We then apply this result to obtain arbitrage intervals for the price of contingent claims in market models with frictions in $\mathfrak{J}$. Section 3.2 is devoted to the characterization of the obtained arbitrage bounds in terms of superreplication cost. We start by defining in a general model with flows the so-called superreplication cost, which essentially corresponds to the minimum initial wealth needed to cover all future contingent flows. We show that for any contingent flow, it is equal to the upper bound of the arbitrage interval.

The notion of superreplication cost was first introduced by Kreps (1981), for classical contingent claims and in the context of incomplete markets (with no

other imperfection). In a diffusion framework, and still with no other imperfection than incompleteness, El Karoui and Quenez (1995) obtain a dual formulation for the superreplication price; in Delbaen (1992) and Delbaen and Schachermayer (1994), this result is obtained in a more general framework. In the spirit of Kreps (1981), Jouini and Kallal (1995a,b) take into account the cases of proportional transaction costs and short sale constraints. For transaction costs, the problem was first introduced by Bensaid *et al.* (1992), who show that in a binomial model with transaction costs, perfect replication is not optimal. Cvitanić and Karatzas (1996) give, in a diffusion framework, a dual formulation for the superreplication price. Delbaen, Kabanov and Valkeila (2001) and Kabanov (1999) generalize this result to the multivariate case, in discrete as well as in continuous time, and with a semimartingale price process. For convex constraints, and still in a diffusion framework, the dual formulation is obtained in Cvitanić and Karatzas (1993). In a more general framework, the result is obtained in Föllmer and Kramkov (1997).

Section 4 deals with economies with fixed transactions costs, which do not fall in the preceding framework, since the set of all available investments is not a convex cone. It is adapted from Jouini, Kallal and Napp (2000). We first obtain a characterization of the no-free-lunch assumption in a general model with flows. We find that the assumption of no-arbitrage is essentially equivalent to the existence of a family of nonnegative “discount processes” such that the net present value of any available investment is nonnegative. Then we apply this result to a securities market model where investors are submitted to both fixed and proportional transaction costs. In that case, the nonnegative discount processes can be interpreted as absolutely continuous martingale measures. Finally, we study pricing issues in securities market models with fixed transaction costs. We adopt an axiomatic approach. We define admissible pricing rules on the set of attainable contingent claims as the price functionals that are arbitrage free and are lower than or equal to the superreplication cost. Indeed, no rational agent would pay more than its superreplication cost for a contingent claim since there is a cheaper way to achieve at least the same payoff using a trading strategy. We then show that the only admissible pricing rules on the set of attainable contingent claims are those that are equal to the sum of an expected value with respect to any absolutely continuous martingale measure and of a bounded fixed cost functional.

2 The Fundamental Theorem of Asset Pricing

We start by describing our general model with flows in a convex cone framework, and in such a model, we characterize the assumption of no free lunch. Then we apply this result to all cases of market imperfections belonging to the class $\mathfrak{I}$.

2.1 In a general “convex cone model” with flows

We adopt the framework of Jouini and Napp (2001), Napp (2000) or Jouini *et al.* (2000, Section 1). We introduce a few notations.

For a filtered probability space $(\Omega, F, (F_t)_{t \geq 0}, P)$, define the measure space $(\hat{\Omega}, \hat{F}, \hat{\mu})$ as the direct sum of the probability spaces (Ω, F_t, P) , i.e. $\hat{\Omega}$ is the disjoint union of continuum many copies $(\Omega_t)_{t \geq 0}$ of Ω , $\hat{F}$ is the sigma-algebra of sets $\hat{A} \subseteq \hat{\Omega}$ such that $\hat{A} \cap \Omega_t \in F_t$, for each $t \geq 0$, and $\hat{\mu}$ induces on each $(\Omega_t, \hat{F}|_{\Omega_t})$ the original probability measure P . We then may represent the Banach space $X \equiv L^1(\hat{\Omega}, \hat{F}, \hat{\mu})$ as the space of all families $\Phi = (\Phi_t)_{t \geq 0}$ such that $\Phi_t \in L^1(\Omega, F_t, P)$ and

$$\|\Phi\|_{L^1(\hat{\Omega}, \hat{F}, \hat{\mu})} = \sum_{t \geq 0} \|\Phi_t\|_{L^1(\Omega, F_t, P)} < \infty.$$

The finiteness of the above sum implies in particular that $\Phi_t = 0$ for all but countably many t in R_+ . The dual space of X may be represented as $Y \equiv L^\infty(\hat{\Omega}, \hat{F}, \hat{\mu})$, which is defined as the space of all families $g = (g_t)_{t \geq 0}$ such that $g_t \in L^\infty(\Omega, F_t, P)$ and

$$\|g\|_{L^\infty(\hat{\Omega}, \hat{F}, \hat{\mu})} = \sup_{t \geq 0} \|g_t\|_{L^\infty(\Omega, F_t, P)} < \infty.$$

The scalar product is defined by $\langle \Phi, g \rangle_{X,Y} = \sum_{t \geq 0} \langle \Phi_t, g_t \rangle$. Elements of X and Y are defined up to a modification.

Let $\check{X} \equiv L^1(\hat{\Omega}_0, \hat{F}_0, \hat{\mu}_0)$, where $(\hat{\Omega}_0, \hat{F}_0, \hat{\mu}_0)$ is the direct sum of the probability spaces $(\Omega, (F_t)_{t > 0}, P)$. Then $\check{Y} \equiv L^\infty(\hat{\Omega}_0, \hat{F}_0, \hat{\mu}_0)$ denotes the dual space of $\check{X}$. For $x, y \in X$ or Y (resp. $\check{X}$ or $\check{Y}$), we write $x \geq y$ if for all $t \geq 0$ (resp. $t > 0$), $x_t \geq y_t$ a.s. P . For all subset Z of $X, Y, \check{X}$ or $\check{Y}$, we denote by Z_+ (resp. Z_-) the set of $x \in Z$ such that $x \geq 0$ (resp. $x \leq 0$).

We consider a model in which agents face investment opportunities described by their cash flows. A probability space (Ω, F, P) is specified and fixed. The set Ω represents all possible states of the world. An information structure, which describes how information is revealed to investors, is given by a filtration $(F_t)_{t \geq 0}$ satisfying the “usual conditions” and such that $F_0 = \{\emptyset, \Omega\}$. We consider investments of the following form:

Definition 2.1 *An investment is a process $\Phi = (\Phi_t)_{t \geq 0} \in X$.*

For each $t \geq 0$, the random variable Φ_t corresponds to the cash flow generated at time t by the investment Φ ; if $\Phi_t(\omega) = k$, this means that the investor receives k at date t if k is nonnegative and pays $-k$ at date t if k is nonpositive. An arbitrage opportunity is as usual a possibility to find an investment that yields a positive gain in some circumstances without a countervailing threat of loss in

other circumstances. In our framework, an arbitrage opportunity would consist of a nonnegative nonnull available investment.

We consider a convex cone J of available investments: this amounts to saying that an investor has a right to subscribe to (a finite number of) different investment plans and that he can decide at the starting date of any investment opportunity which amount of this particular investment he wants to buy. We are led to consider convex cones in order to take into account the fact that investors are not necessarily able to sell an investment plan (consider for instance the case of short sale constraints or transaction costs). In order to obtain the Fundamental Theorem of Asset Pricing in this context, we make the additional assumption that there is in the convex cone J some possibility of transferring some money. More precisely, we introduce the following assumption.

Assumption A: there exists a sequence $d = (d_n)_{n \geq 0}$ such that for all $t^* \geq 0$, for all B_{t^*} in F_{t^*} of positive probability, there exists Φ in J such that $\Phi_t = 0 \ \forall t < t^*$, $\Phi_{t^*} = 0$ outside B_{t^*} , $\Phi_t \geq 0 \ \forall t > t^*$ and $\exists d_n \in d$, $P[\Phi_{d_n} > 0] > 0$.

In words, this means that there exists a sequence of trading dates such that, for every date and for every event at that date, there exists an investment plan in our set of available investments that starts at that date and in that event, that can take any value at that date and in that event, but that is nonnegative after that date and nonnull at one date belonging to the above mentioned sequence of dates. This assumption is not too restrictive. See Jouini and Napp (2001) for more details on this assumption.

We don't specify the elements of J so far. The assumption of no-arbitrage for J can be written $J \cap X_+ = \{0\}$ or equivalently $(J - X_+) \cap X_+ = \{0\}$. A free lunch denoting the possibility of getting arbitrarily close to an arbitrage opportunity, we introduce the following definition.

Definition 2.2 *There is no free lunch for J if and only if $\overline{J - X_+} \cap X_+ = \{0\}$, where the bar denotes the closure for the norm topology in X .*

We now characterize the absence of free lunch. Notice that since we do not necessarily have the opportunity to transfer money from one time to another, we cannot consider “net gains” anymore, and we have to get an analog of the Kreps–Yan theorem (Yan (1980), Kreps (1981)) in a more complex space than the classical $L^1(\Omega, F, P)$ for a probability (or sigma-finite measure) space (Ω, F, P) . In our general context with investments in X , we obtain the following Fundamental Theorem of Asset Pricing.

Theorem 2.3 *Under Assumption A, there is no free lunch for J if and only if there exists a positive process $g = (g_t)_{t \geq 0}$ in Y such that $g|_J \leq 0$.*

Note that positive means here that g seen as a linear functional on X is positive, or equivalently that for all t , $g_t > 0$ a.s. P . Since for all $\Phi \in J$, $\langle \Phi, g \rangle_{X,Y} = E \left[\sum_{t \geq 0} g_t \Phi_t \right]$, Theorem 2.3 means that the absence of free lunch (for J) is essentially equivalent to the existence of a discount process under which the “net present value” of any available investment (in J) is nonpositive. We shall denote by $\mathcal{G}^J$ the set of all “admissible discount processes”, i.e. $\mathcal{G}^J \equiv \{g \in Y, g > 0, g|_J \leq 0\}$. If there is no free lunch, then according to Theorem 2.3, $\mathcal{G}^J$ is non-void.

2.2 Application to the characterization of the no-free-lunch assumption in all cases of market imperfections in $\mathfrak{I}$

As our investment opportunities are supposed to be very general, it is shown in Jouini and Napp (1998) that most market models involving imperfections can fit in the model for a specific convex cone of investments J satisfying *Assumption A*. This is the case for the following set (that we shall denote by $\mathfrak{I}$) of imperfect market models: models with imperfections concerning the numéraire (no borrowing, different borrowing and lending rates), models with dividends, short-sale constraints, convex cone constraints, proportional transaction costs. Let us see how for instance Theorem 2.3, obtained in a general setting, can be applied to the case of short sale constraints. As in Jouini and Kallal (1995b), we consider a model of financial market where two sorts of securities can be traded. Short selling the first type of securities is not allowed, i.e. they can only be held in nonnegative amounts, whereas the second type of securities can only be held in nonpositive amounts. The model includes situations where holding negative amounts of a security is possible but costly as well as situations where some (or all) securities are not subject to any constraints, since we may include a security twice in the model, in the first and in the second set of securities. For $1 \leq k \leq n$ (resp. $n+1 \leq k \leq N$), we denote by S^k the price process of the security k that can only be held in nonnegative (resp. nonpositive) amounts. We assume that for $k \in \{1, \dots, N\}$, S_t^k belongs to $L^1(\Omega, \mathcal{F}_t, P)$ for all t , and that $S^1 \equiv 1$ (i.e. there are lending opportunities). For all $t_1 \leq t_2$, for all bounded nonnegative $\mathcal{F}_{t_1}$ -measurable real-valued random variables θ , we let $\Phi^{(k; \theta, t_1, t_2)}$ denote the process given by $\Phi_t^{(k; \theta, t_1, t_2)} = -\theta S_{t_1}^k 1_{t=t_1} + \theta S_{t_2}^k 1_{t=t_2}$ for $1 \leq k \leq n$ and $\Phi_t^{(k; \theta, t_1, t_2)} = \theta S_{t_1}^k 1_{t=t_1} - \theta S_{t_2}^k 1_{t=t_2}$ for $n+1 \leq k \leq N$. We assume that the set J_S is the convex cone generated by all these investments. Then J_S satisfies *Assumption A* and by an immediate application of Theorem 2.3, we get that there is no free lunch for J_S , or equivalently that there is no free lunch in a model with short sale constraints, if and only if the set $\mathcal{G}^{J_S}$ is nonempty, where $\mathcal{G}^{J_S}$ denotes the set of positive processes $g \in Y$ such that for all securities k that cannot

be sold short (i.e. $k \leq n$), gS^k is a supermartingale and for all securities k that can only be sold short (i.e. $n + 1 \leq k$), gS^k is a submartingale.

We adopt in Jouini and Napp (2001) a similar approach for all other market imperfections in $\mathfrak{I}$. Each time, we introduce a specific set of available investments corresponding to the considered imperfection, we apply Theorem 2.3 and obtain more or less directly a specific characterization of the no-free-lunch condition in these imperfect market models. In each case, we find that there is no free lunch if and only if a given specific convex set of discount processes¹ is nonempty.

2.3 A few remarks and extensions

- In Jouini *et al.* (2000), we adopt a new topology on X for the definition of a free lunch. The idea is to weaken the topology on X ; to motivate this idea, recall that we have considered the norm topology on $L^1(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mu})$ so that its dual equals $L^\infty(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mu})$. Considering the elements $g = (g_t)_{t \in R_+} \in L^\infty(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mu})$ as functions on $\Omega \times R_+$ note that, for fixed $\omega \in \Omega$, the function $t \mapsto g_t(\omega)$ does not obey any continuity or measurability requirements (apart from being uniformly bounded). The space $Y = L^\infty(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mu})$ seems too big for a useful economic interpretation and should be replaced by a space Y of more regular processes, e.g., the adapted bounded processes $(y_t)_{t \in R_+}$ which almost surely have càd (right continuous) or càg (left continuous) or continuous trajectories. This leads us to consider the space $X = L^1(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mu})$ in duality with the space Y proposed above and to equip X with a topology τ compatible with the dual pair $\langle X, Y \rangle$. We prove in Jouini *et al.* (2000) that in this setting we do have a positive result of Yan type, hence a characterization of the no-free-lunch assumption, without *Assumption A*; more precisely, we prove that for all closed convex cones in X such that $C \supseteq X_-$, if $C \cap X_+ = \{0\}$, then we can find a strictly positive linear functional $y \in Y_{++}$, such that $y|_C \leq 0$.
- Still in Jouini *et al.* (2000), we generalize the framework of Section 2.1, by considering a space of investments given by a space of measures. More precisely, we take X given by $\mathcal{M}(R_+ \times \Omega, \mathcal{O})$, the space of equivalence classes of finite measures μ on the optional sigma-algebra $\mathcal{O}$, modulo the measures supported by evanescent sets. Note that this enables us to model in X continuous time payment streams (which may or may not be absolutely continuous with respect to Lebesgue-measure). We obtain a characterization of the no-free-lunch assumption in such a context.
- We study in Napp (2000) the links between the extremality or the uniqueness of the “admissible discount process” given by the absence of free lunch and the

¹ See Section 4 for a description of this set in the transaction costs case.

completeness of the market, in the case where the convex cone J of available investments is a linear subspace of X . Similar results have been obtained in Jacod (1979), Harrison and Pliska (1981), Delbaen (1992) and Delbaen and Schachermayer (1994).

3 Arbitrage intervals and superreplication cost

Now that we have characterized the absence of free lunch, we shall turn to pricing issues, still in the framework of Section 2.

3.1 Arbitrage intervals

We start with the general framework with a convex cone of available flows. We adopt the approach of Harrison and Kreps (1979). We assume that we are faced with a so-called primitive financial market consisting of a convex cone C of available investment opportunities satisfying *Assumption A*. We suppose that there is no free lunch in the primitive market or equivalently that there is no free lunch for C , so that according to Theorem 2.3, the set $\mathcal{G}^C$ is nonempty. In addition to this primitive market, we consider a contingent flow in the form of some process $\check{\Phi} = (\Phi_t)_{t>0} \in \check{X}$. The aim of this subsection is to give a “fair” price to this additional contingent flow by only using arbitrage considerations.

We say that $(-\Phi_0)$ is a fair (buying) price for $\check{\Phi} \in \check{X}$ if there is no free lunch in the so-called full market consisting of the convex cone C' generated in X by C and $\Phi \equiv (\Phi_t)_{t>0}$. These values of $(-\Phi_0)$ can be seen as the price to pay at date 0 in order to have access to the flows Φ_t at each date $t > 0$, in a way that is compatible with the no-free-lunch condition. For all $\check{\Phi} \in \check{X}$, let $l_{\check{\Phi}} \equiv \inf_{g \in \mathcal{G}^C} \left\langle \check{\Phi}, \left(\frac{g_t}{g_0} \right)_{t>0} \right\rangle_{\check{X}, \check{Y}}$ and $u_{\check{\Phi}} \equiv \sup_{g \in \mathcal{G}^C} \left\langle \check{\Phi}, \left(\frac{g_t}{g_0} \right)_{t>0} \right\rangle_{\check{X}, \check{Y}}$. For simplicity of notation, we shall indifferently write $\left\langle \check{\Phi}, \left(\frac{g_t}{g_0} \right)_{t>0} \right\rangle_{\check{X}, \check{Y}}$ or $\left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$.

Lemma 3.1 *A price $(-\Phi_0)$ is a fair price for $\check{\Phi}$ if and only if there exists $g \in \mathcal{G}^C$, $(-\Phi_0) \geq \left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$. Any fair price $(-\Phi_0)$ satisfies $(-\Phi_0) \geq l_{\check{\Phi}}$. Conversely, any price $(-\Phi_0) > l_{\check{\Phi}}$ is a fair price for $\check{\Phi}$.*

We have obtained a lower bound on the value of any fair (buying) price. Any fair buying price for a contingent flow is a price that is greater than or equal to the net present value of the flow with respect to some admissible discount process. In a natural way, a fair selling price for $\check{\Phi} \in \check{X}$ is the opposite of a fair (buying) price for $-\check{\Phi} \equiv (-\check{\Phi}_t)_{t>0}$. By applying Lemma 3.1 to $-\check{\Phi}$, we get that any fair selling

price for $\check{\Phi}$ satisfies $(-\Phi)_0 \leq u_{\check{\Phi}}$ and that, conversely, any price $(-\Phi)_0 < u_{\check{\Phi}}$ is a fair selling price for $\check{\Phi}$. Notice that if $\check{\Phi}$ can be bought and sold, then by arbitrage considerations, its buying price necessarily lies above its selling price.

We say that $(-\Phi_0)$ is a fair buying–selling price for $\check{\Phi} \in \check{X}$ if there is no free lunch in the market consisting of the convex cone generated in X by C , Φ and $-\Phi$. It corresponds to the price at which $\check{\Phi}$ can be bought and sold without generating any free lunch.

Corollary 3.2 *A price $(-\Phi_0)$ is a fair buying–selling price for $\check{\Phi}$ if and only if there exists $g \in \mathcal{G}^C$, $(-\Phi_0) = \left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$. Any fair buying–selling price $(-\Phi_0)$ belongs to $[l_{\check{\Phi}}, u_{\check{\Phi}}]$. Conversely, if $l_{\check{\Phi}} = u_{\check{\Phi}}$, then there is a unique fair buying–selling price equal to $l_{\check{\Phi}}$, and if $l_{\check{\Phi}} < u_{\check{\Phi}}$, then any price $(-\Phi_0) \in]l_{\check{\Phi}}, u_{\check{\Phi}}[$ is a fair buying–selling price for $\check{\Phi}$.*

If $\mathcal{G}^C$ is reduced to a singleton, then there exists a unique fair buying–selling price for any $\check{\Phi} \in \check{X}$. If $\mathcal{G}^C$ is not reduced to a singleton, we only obtain arbitrage intervals for the price of contingent flows. For any contingent flow which can be bought and sold, its arbitrage interval consists of its net present value under all admissible discount processes in $\mathcal{G}^C$.

We can now apply these results for the pricing of contingent claims in any market model in $\mathcal{J}$. Let $T \in R_+^*$. A contingent claim will denote any random variable H in $L^1(\Omega, F_T, P)$, corresponding to the payoff at date² T . We want to give a fair price to a contingent claim H by only using arbitrage considerations. We still assume that we are faced with a so-called primitive financial market consisting of a convex cone C of available investment opportunities satisfying *Assumption A* and such that the set $\mathcal{G}^C$ is nonempty. In addition to this primitive market, we assume that investors have access to the contingent claim H so that the set of all available investment opportunities consists of the convex cone C' generated by C and the contingent flow $\Phi^H \in X$ given by $\Phi_T^H = H$ and $\Phi_t^H = 0$ for all $t \notin \{0, T\}$. We say that $(-\Phi_0^H)$ is a fair (buying) price for H if it is a fair price for $(\Phi_t^H)_{t>0} \in \check{X}$.

By applying Lemma 3.1 to the investment opportunity $(\Phi_t^H)_{t>0}$ in $\check{X}$, we immediately get the following result.

Corollary 3.3 *Any fair buying price $(-\Phi_0^H)$ for a contingent claim H satisfies $(-\Phi_0^H) \geq \inf_{g \in \mathcal{G}^C} E \left[\frac{g_T}{g_0} H \right]$. Any fair selling price for H satisfies $\Phi_0^{-H} \leq$*

² Notice that contingent claims whose payoffs belong to $\check{X}$, without necessarily being related to a unique date T , also fall in our framework.

$\sup_{g \in \mathcal{G}^C} E \left[\frac{g_T}{g_0} H \right]$. If H can be bought and sold at the same price, then $(-\Phi_0^H) \in \left[\inf_{g \in \mathcal{G}^C} E \left[\frac{g_T}{g_0} H \right], \sup_{g \in \mathcal{G}^C} E \left[\frac{g_T}{g_0} H \right] \right]$.

We are now able to use the specific characterization of the set $\mathcal{G}^C$ obtained in the different imperfect market models in $\mathfrak{I}$ (see Jouini and Napp (2001)) to obtain in each case specific arbitrage bounds. We state the result with short sale constraints, i.e. in the case where, with the notations of Section 2, C is given by J_S .

Corollary 3.4 *If there are short sale constraints, the buying price for any contingent claim H is greater than or equal to $\inf_{g \in \mathcal{G}^{J_S}} E \left[\frac{g_T}{g_0} H \right]$, and if there is a selling price for H , it is smaller than or equal to $\sup_{g \in \mathcal{G}^{J_S}} E \left[\frac{g_T}{g_0} H \right]$.*

We shall now pin down these arbitrage intervals, through the use of the superreplication cost.

3.2 Arbitrage bounds and superreplication cost

The aim of this subsection is to show that the upper bound of the arbitrage interval, in a general context with flows as well as in market models with frictions in $\mathfrak{I}$, is given by the so-called superreplication cost; for a contingent flow $x \in \check{X}$, this cost corresponds to the minimum initial wealth needed to obtain, through available investments, at least as much as the flow x . This notion was originally introduced by Kreps (1981) for classical contingent claims in the context of incomplete markets (with no other imperfection). All available investments still consist of a convex cone and we consider the set M of contingent flows in $\check{X}$ that agents can “dominate” by using available investment opportunities,

$$M \equiv \{x \in \check{X}, \exists \Phi \in J, \Phi_t \geq x_t \forall t > 0\}.$$

In words, M is the set of flows m for which there exists an available investment (in J), which is unambiguously better than m after the initial date. We now introduce on M the notion of superreplication cost.

Definition 3.5 *For all $m \in M$, the superreplication cost of m is denoted by $\bar{\pi}(m)$ and given by*

$$\bar{\pi}(m) \equiv \inf \left\{ \liminf (-\Phi_0^n); \Phi_t^n \geq m_t^n \forall t > 0, (\Phi^n, m^n) \in J \times M, m^n \rightarrow_{\check{X}} m \right\}.$$

The superreplication cost represents the infimum wealth necessary to subscribe to an investment opportunity which will provide us with at least as much as a flow arbitrarily close to m . Like in Jouini and Kallal (1995a) for the case of proportional transaction costs, we start by describing the set M and the functional $\bar{\pi}$.

Lemma 3.6 *The set M is a convex cone. If there is no free lunch for J , the price functional $\bar{\pi}$ is a sublinear³ lower semi continuous⁴ functional which takes values in R .*

We are now in a position to obtain a dual representation formula for the upper bound of the arbitrage intervals.

Proposition 3.7 *If there is no free lunch for J , then for all $m \in M$, $\bar{\pi}(m) = \sup_{g \in \mathcal{G}^J} \left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$.*

This means that the superreplication cost of a contingent flow is equal to the supremum of its expected value with respect to all admissible discount processes, which coincides with the upper bound of the arbitrage interval. If we now consider some $m \in M$ such that $-m \in M$, a symmetric argument yields $-\bar{\pi}(-m) = \inf_{g \in \mathcal{G}^J} \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$, or $[-\bar{\pi}(-m), \bar{\pi}(m)] = cl \left\{ \left\langle \frac{g}{g_0}, m \right\rangle; g \in \mathcal{G}^J \right\}$, so that the bounds of the arbitrage intervals, in the general context with flows as well as for contingent claims in imperfect market models (belonging to $\mathfrak{I}$), are completely characterized in terms of superreplication cost.

Note that for some authors, the “true” superreplication cost is given on M by $\pi(m) = \inf \{(-\Phi_0); \Phi \in J, \Phi_t \geq m_t \forall t > 0\}$. It is proved in Napp (2000) that under the assumption of no-free-lunch, $\bar{\pi}$ is the largest lower semi continuous functional lying below π . Besides, we investigate when the upper bound of the arbitrage interval is effectively given by the “true” superreplication price π , in other words, when $\bar{\pi} = \pi$. We get the equality when π is l.s.c. or each time that for every scalar λ , the set of contingent flows that can be dominated by an available investment opportunity with initial value smaller than or equal to λ is closed. More generally, we consider some specific market models in $\mathfrak{I}$ for which more simple expressions for $\bar{\pi}$ can be obtained: discrete models as well as models with short sale constraints and imperfections on the numéraire if we assume that asset prices are continuous. Notice however that the approach with $\bar{\pi}$ has enabled us to characterize the arbitrage bounds in a general framework.

3.3 A few remarks and extensions

In Napp (2000), we adopt an axiomatic approach. Like in Harrison and Pliska (1981) and more recently Jouini (2000) for the case of proportional transaction

³ That is, for all m_1, m_2 in M and all $\lambda \in R_+$, we have $\bar{\pi}(m_1 + m_2) \leq \bar{\pi}(m_1) + \bar{\pi}(m_2)$ and $\bar{\pi}(\lambda m_1) = \lambda \bar{\pi}(m_1)$.

⁴ That is, such that $\{(m, \lambda) \in M \times R; \bar{\pi}(m) \leq \lambda\}$ is closed in $M \times R$, or equivalently such that $\{m \in M; \bar{\pi}(m) \leq \lambda\}$ is closed in M for all $\lambda \in R$, or equivalently such that $\liminf_n \{\bar{\pi}(m_n)\} \geq \bar{\pi}(m)$ whenever the sequence $(m_n) \subset M$ converges to $m \in M$.

costs, and Koehl and Pham (2000) for convex constraints, we start from a certain number of axioms that a price functional, defined on the set of contingent flows, must satisfy in order to be admissible. These axioms are linked not only to arbitrage but also equilibrium considerations. We obtain a dual characterization of all admissible functionals. A similar axiomatic approach will be adopted in Section 4 for models with fixed transaction costs.

We also study issues related to the viability (a notion introduced by Harrison and Kreps (1979)), or equivalently to the compatibility with an equilibrium, of the pricing rules we have found. We emphasize that all results obtained for a general contingent flow can be applied to contingent claims in securities market models with frictions belonging to $\mathfrak{J}$.

4 Models with fixed transaction costs

We consider in this section financial models where the available investment flows are subject to fixed transaction costs.

4.1 The characterization of the no-free-lunch assumption in a general model with fixed costs

We introduce a few notations. We denote by $\mathcal{S}^f$ the collection of stopping times of $(F_t)_{t \geq 0}$ taking a finite number of values in R_+ . For any $\tau \in \mathcal{S}^f$, we denote by $\mathcal{S}_\tau^f$ the class of stopping times ν in $\mathcal{S}^f$ with $\tau \leq \nu$ a.s.

Definition 4.1 *An investment consists of*

1. *an initial stopping time τ in $\mathcal{S}^f$*
2. *a starting event B in F_τ*
3. *an $(F_t)_{t \geq 0}$ -adapted process $\Phi = (\Phi_t)_{t \geq 0}$ such that Φ is null outside B , and there exists a finite set of stopping times $\tau = \tau_1^\Phi \leq \dots \leq \tau_{N_\Phi}^\Phi$ in $\mathcal{S}_\tau^f$ for which $\Phi_t = 0$ for all $t \notin \{\tau_l^\Phi\}_{l \in \{1, \dots, N_\Phi\}}$ and for all l , $\Phi_{\tau_l^\Phi} \in L^1(\Omega, F_{\tau_l^\Phi}, P)$.*

We shall call the process Φ the investment process. The starting stopping time and event can correspond to the stopping time and event at which one investor may subscribe to the investment opportunity. The investment process corresponds to the associated cash flow.

We still consider a convex cone $\mathcal{I}$ of available investment processes and for all pairs $(\tau, B) \in \mathcal{S}^f \times F_\tau$, we let $\mathcal{I}^{\tau, B}$ (resp. $\mathcal{J}^{\tau, B}$) denote the set of all available investment processes associated with investments with starting stopping time τ and starting event B (resp. starting after τ and B , i.e. $\mathcal{J}^{\tau, B} = \bigcup_{\substack{\nu \geq \tau \\ B' \subseteq B}} \mathcal{I}^{\nu, B'}$).

We assume that we can transfer wealth from one date to another, i.e. that, for all stopping times τ_1, τ_2 in $\mathcal{S}^f$ and for all random variables θ in $L^1(\Omega, F_{\tau_1 \wedge \tau_2}, P)$, the process denoted by $\Phi^{(0; \theta, \tau_1, \tau_2)}$ and given by $\Phi_t^{(0; \theta, \tau_1, \tau_2)} = -\theta 1_{t=\tau_1} + \theta 1_{t=\tau_2}$ with starting stopping time $\tau_1 \wedge \tau_2$ and starting event equal to $\{\theta \neq 0\}$ belongs to the set $\mathcal{I}$ of all available investment processes. We shall denote by Γ the set of such transfers, i.e. the convex cone generated by all these investment processes.

We assume that it is not costless to subscribe to an investment, i.e. that there are “fixed costs” associated with any investment plan. More precisely, we associate with each investment (τ, B, Φ) a nonnegative cost process $c^{(\tau, B, \Phi)} = (c_t^{(\tau, B, \Phi)})_{t \geq 0}$; when there is no ambiguity, we shall sometimes write c^Φ instead of $c^{(\tau, B, \Phi)}$. The assumptions we make on the fixed costs are the following: we assume first that the cost process is $(F_t)_{t \geq 0}$ -adapted, which means that investors know at time t the past and current values of the fixed cost but nothing more. We assume that the cost process $c^{(\tau, B, \Phi)}$ is null before the stopping time τ , outside the event B , and outside a finite number of stopping times in $\mathcal{S}^f$. Besides, we assume that there is no fixed cost associated with the transferring of wealth from one date to another, i.e. for all $\Phi \in \mathcal{I}$, for all $\Psi \in \Gamma$, we have $c^\Phi = c^{\Phi + \Psi}$. Moreover, the total cost associated with any investment opportunity is supposed to be bounded, i.e. there exists a positive real number C such that $\sum_{t \geq 0} c_t^\Phi \leq C$ for all $\Phi \in \mathcal{I}$, which can be interpreted as the investors’ refusal to pay more than a certain given amount for fixed costs: this explains why we call these costs fixed costs as opposed to proportional costs. Finally, the fixed costs incurred at the initial stopping time must be “positive”, i.e. for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, there exists a positive real number $\varepsilon^{\tau, B}$, such that all investment processes $\Phi \in \mathcal{I}^{\tau, B}$ with $\Phi \notin \Gamma$ satisfy $c_t^\Phi \geq \varepsilon^{\tau, B}$ on B .

According to these assumptions, the fixed costs can be interpreted as information costs, opportunity costs, time costs, etc. In a financial market model, they can correspond to fixed brokerage fees. They can account for a sort of cost of accessing⁵ the available investments or more generally for frictions of all kinds.

As usual, an arbitrage opportunity is an investment plan that yields a positive gain in some circumstance, without a countervailing threat of loss in other circumstances and a free lunch is a possibility of getting arbitrarily close to an arbitrage opportunity.

Definition 4.2 *An arbitrage opportunity is an available investment (τ, B, Φ) with Φ in $\mathcal{I}$ such that $\Phi_t - c_t^\Phi \geq 0$ for all $t \geq 0$, and there exists a date for which it is nonnull.*

⁵ This “cost of accessing the investment opportunities” can be understood in a general sense: it can be a fee (such as an investment tax), or the cost of setting up an office.

For all pairs $(\tau, B) \in \mathcal{S}^f \times F_\tau$, we let $\mathcal{A}^{\tau, B}$ denote the set of all nonnegative investment processes u such that $u_\tau > \varepsilon''$ on B for some positive constant ε'' and we obtain the following characterization of the absence of arbitrage opportunity in our model.

Lemma 4.3 *There is no arbitrage opportunity if and only if for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, we have $\mathcal{I}^{\tau, B} \cap \mathcal{A}^{\tau, B} = \emptyset$.*

Using the same notations as for the definition of an arbitrage opportunity, we now introduce the notion of free lunch. We shall consider the set $\mathcal{I}$ as a subset of $L^1(\hat{\Omega}, \hat{F}, \hat{\mu})$, considered in Section 2.1, and adopt the norm topology on this space.

Definition 4.4 *There is a free lunch if and only if there exist a pair $(\tau, B) \in \mathcal{S}^f \times F_\tau$ for which $\overline{\mathcal{I}^{\tau, B}} - L_+^1(\hat{\Omega}, \hat{F}, \hat{\mu}) \cap \mathcal{A}^{\tau, B} \neq \emptyset$, where the bar denotes the closure in $L^1(\hat{\Omega}, \hat{F}, \hat{\mu})$.*

See Jouini, Kallal and Napp (2000) for an interpretation of the definition of a free lunch in a securities market model with fixed transaction costs. Notice that the assumption of no-free-lunch in such a model is less restrictive than in the without-fixed-cost otherwise identical model. We now obtain the main result.

Theorem 4.5 *There is no free lunch if and only if for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, there exists an absolutely continuous probability measure $P^{\tau, B}$ with bounded density such that $P^{\tau, B}(B) = 1$ and for every investment process Φ in $\mathcal{J}^{\tau, B}$, $E^{P^{\tau, B}}[\sum_{t \geq 0} \Phi_t] \leq 0$.*

This means that the absence of free lunch in our model with fixed trading costs is equivalent to the existence of a family of absolutely continuous probability measures under which the net present value of any available investment is nonpositive.

4.2 Application to securities market models with both fixed and proportional costs

We consider an economy where agents can trade a finite number of securities and we assume that these securities are subject to bid–ask spreads: at each date, there is not a unique price for a security but an ask price, at which investors can buy the security and a bid price, at which they can sell the security. Notice that this model includes situations where there is a unique price process Z and where the proportional transaction cost remains constant over time, i.e. situations where at each time t , investors must pay $Z_t(1 + c)$ for some positive constant c to buy the security and receive $Z_t(1 - c)$ when selling it.

More precisely, we consider $(n + 1)$ securities and for each security k for $0 \leq k \leq n$, we let $(Z_t^k)_{t \geq 0}$ and $(Z_t'^k)_{t \geq 0}$ denote respectively the ask and bid price process. We assume that the $(n + 1)$ -dimensional processes Z and Z' are right-continuous and of class $\mathcal{D}^f$, i.e. that the families $\{Z_\tau\}_{\tau \in \mathcal{S}^f}$ and $\{Z'_\tau\}_{\tau \in \mathcal{S}^f}$ are uniformly integrable.

For each k in $\{0, \dots, n\}$, for all stopping times τ_1 and τ_2 in $\mathcal{S}^f$, for all nonnegative real-valued bounded random variables θ in $F_{\tau_1 \wedge \tau_2}$, we let $\Phi^{(k; \theta, \tau_1, \tau_2)}$ denote the process given by

$$\Phi_t^{(k; \theta, \tau_1, \tau_2)} = \theta \left[-Z_{\tau_1}^k 1_{t=\tau_1} + Z_{\tau_2}^k 1_{t=\tau_2} \right]$$

and we assume that the set $\mathcal{I}$ of all available investment processes consists of the convex cone generated by all the processes $\Phi^{(k; \theta, \tau_1, \tau_2)}$. This means that all available investment opportunities are related to the buying and selling of the $(n + 1)$ securities, at some stopping times and in random quantities. We still assume that we can transfer wealth without friction, i.e. we set for all t , $Z_t^0 = Z_t'^0 = 1$.

Like in the previous section, we assume that there are fixed costs associated with these investment opportunities. The assumptions made on the fixed costs remain the same as above but their interpretation in this specific setting can be made more accurately.

First, if an investor doesn't trade in the risky securities at time t , then he doesn't pay any additional cost; but in order to buy at stopping time τ a "portfolio" Θ_τ , he must pay $\Theta_\tau \cdot Z_\tau + c_\tau^\Theta$, where c_τ^Θ denotes the fixed cost to be paid by the investor at stopping time τ when following the strategy Θ . The fixed cost can depend upon the strategy followed by the investors: for instance at the same date and event, it can be different according to what the investor has done before that date and event; this means equivalently that the fixed costs to be paid are not necessarily the same for all investors.

Second, the aggregated fixed costs are bounded independently of the chosen strategy and independently of the considered investor, or in other words we assume that there exists a positive real number C such that for all strategies Θ , $\sum_{t \geq 0} c_t^\Theta \leq C$. This means in particular that the fixed costs to be paid at some date t are bounded independently of the amount traded, which explains why we call them fixed costs as opposed to proportional costs.

Finally, we assume that at the first time an investor trades, he incurs a positive fixed cost, which is to be interpreted as a cost of accessing the market.

We get the following characterization of the absence of free lunch in a model with proportional and fixed transaction costs.

Theorem 4.6 *There is no free lunch in our model with fixed and proportional transaction costs if and only if for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, there exists an absolutely*

continuous probability measure $P^{\tau, B}$ with bounded density such that $P^{\tau, B}(B) = 1$ and some process $S^{\tau, B}$ satisfying

$$\begin{aligned} Z'_t 1_{B \cap \{\tau \leq t\}} &\leq S_t^{\tau, B} 1_{B \cap \{\tau \leq t\}} \leq Z_t 1_{B \cap \{\tau \leq t\}} \\ E^{P^{\tau, B}}[S_{t \vee \tau}^{\tau, B} | F_{s \vee \tau}] &= S_{s \vee \tau}^{\tau, B} \quad \text{for } t \geq s. \end{aligned}$$

This means that for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$ there exists an absolutely continuous probability measure $P^{\tau, B}$ that transforms some price process $S^{\tau, B}$ lying after τ and on B between the discounted bid and ask price processes into a martingale from the stopping time τ and event B . In the case where there is no proportional transaction cost, i.e. if $Z = Z'$, we find that the absence of free lunch in a securities market model with fixed transaction costs is equivalent to the existence of a family of absolutely continuous martingale measures. Our characterization of the no-free-lunch assumption is then weaker than the classical one, and leads to a larger class of arbitrage-free models.

4.3 Pricing issues in securities market models with fixed transaction costs

The framework is the same as in the previous section except that in order to concentrate on the fixed costs, we assume that $Z = Z'$, in other words there is no proportional transaction cost. As in Section 3, we consider a finite time horizon T , and a contingent claim H to consumption at the terminal date T is a random variable belonging to $L^1(\Omega, F_T, P)$. A contingent claim H is said to be attainable (in the model without fixed cost) if there exists some available investment process Φ in $\mathcal{I}^{0, \Omega}$ such that $\Phi_t = 0$ for all $t \in]0, T[$ and $\Phi_T = H$. Note that the set M of all attainable contingent claims is a linear space. We shall now define and characterize pricing rules p on M that are admissible. As in Section 3, we introduce the definition of the superreplication price of H , $\pi_c(H)$, in our framework with fixed costs

$$\begin{aligned} \pi_c(H) \equiv \inf \{ & -\Phi_0 + c_0^\Phi, \Phi \in \mathcal{I}^{0, \Omega}, \Phi_t - c_t^\Phi \geq 0 \text{ for all } t \in]0, T[, \\ & \Phi_T \geq H + c_T^\Phi \} \end{aligned}$$

Definition 4.7 An admissible pricing rule on M is a functional p defined on M , such that

1. p induces no arbitrage, i.e., it is not possible to find processes $\Phi^1, \dots, \Phi^n$ in $\mathcal{I}^{0, \Omega}$, such that $\Phi_t^i = 0$ for all $t \in]0, T[$ and for which $\sum_{i=1}^n p(\Phi_T^i) \leq 0$, $\sum_{i=1}^n \Phi_T^i \geq 0$ and one of the two is nonnull.
2. $p(H) \leq \pi_c(H)$.

Part 1 is the usual no-arbitrage condition. Part 2 says that an admissible price for the contingent claim H must be smaller than its superreplication price: if it is possible to obtain a payoff at least equal to H at a cost $\pi_c(H)$, then no rational agent (who prefers more to less) will accept to pay more than $\pi_c(H)$ for the contingent claim H .

The following proposition characterizes the admissible pricing rules on M through the use of the absolutely continuous martingale measures obtained in Theorem 4.6.

Proposition 4.8 *Under the assumption of no-free-lunch, any admissible pricing rule p on M can be written as*

$$p(H) = E^{P^*}[H] + c(H) \quad \text{for all } H \text{ in } M$$

where P^* is any absolutely continuous martingale measure and c is a bounded functional defined on M .

If we assume that for a large enough scalar λ , we have $p(\lambda x) < \lambda[p(x)]$, then the fixed cost functional is nonnegative; moreover, if we assume that there exists $\varepsilon > 0$, such that for a large enough λ , $p(\lambda x) < \lambda[p(x) - \varepsilon]$, then the fixed cost is greater than or equal to this positive constant ε .

Notice that Proposition 4.8 implies that $p(\lambda H)/\lambda \rightarrow_{\lambda \rightarrow \infty} E^{P^*}[H]$ for any attainable contingent claim H , where P^* is any absolutely continuous martingale measure. This means that the unit price of any attainable contingent claim H is equal to $E^{P^*}[H]$ in the limit of large quantities. In particular, in a Black–Scholes-like model with fixed costs, the unique asymptotic price for any contingent claim is given by the usual Black–Scholes price.

Appendix A

Proof of Theorem 2.3 The proof is adapted from Yan (1980). It is very similar to the one in Jouini and Napp (2001), where *Assumption A* is also made. Let $x \in \overline{J - X_+} \cap X_+$, $x = \lim_n x_n$, where for all n , $x_n \leq \Phi_n$, $\Phi_n \in J$. Then, since g is nonnegative and $g|_J \leq 0$, for all n , $\langle x_n, g \rangle_{X,Y} \leq \langle \Phi_n, g \rangle_{X,Y} \leq 0$. This implies $\langle x, g \rangle_{X,Y} \leq 0$, hence $x = 0$.

Conversely, if $\overline{J - X_+} \cap X_+ = \{0\}$, then for all $x \neq 0$, belonging to X_+ , the Hahn–Banach Separation Theorem yields the existence of $g \neq 0$, belonging to Y such that $g|_{J-X_+} \leq 0 < \langle x, g \rangle_{X,Y}$. It is easy to check that g is nonnegative. Let G^J denote the nonempty set of all nonnegative $g \in Y$, $g|_J \leq 0$.

We start by proving that for all dates t , there exists a process $g^t \in G^J$, such that $g_t^t > 0$ P a.s. Let $\mathcal{S}^t$ be the family of equivalence classes of subsets of Ω formed

by the supports of the g_t for all g in G^J . By applying the Separation Theorem to the element x of X_+ such that $x_t = 1$, $x_s = 0$, $\forall s \neq t$, we get that the family $\mathcal{S}'$ is not reduced to the empty set. It is easy to see that the family $\mathcal{S}'$ is closed under countable unions. Hence there is g^t in G^J such that $\mathcal{S}' \equiv \{g^t > 0\}$ satisfies

$$P(\mathcal{S}') = \sup \{P(\mathcal{S}) ; \mathcal{S} \in \mathcal{S}'\}.$$

We necessarily have $P(\mathcal{S}') = 1$; indeed, if $P(\mathcal{S}') < 1$, then we can apply the Separation Theorem to x such that $x_t = 1_{(\Omega - \mathcal{S}')}$, $x_s = 0$, $\forall s \neq t$ and get the existence of $g'' \in G^J$, $\langle x, g'' \rangle_{X,Y} > 0$. Then $\{g_t^t + g_t'' > 0\}$ would be an element of $\mathcal{S}'$, with P -measure strictly greater than $\mathcal{S}'$: a contradiction.

Now we show that there exists $g \in G^J$ such that $g_{d_n} > 0$ almost surely for all $d_n \in d$, where d is the sequence introduced in *Assumption A*. We consider the process g such that for all $t \geq 0$, $g_t = \sum_{n \geq 0} a_n g_t^{d_n}$, where $(a_n)_{n \geq 0}$ is a sequence of positive scalars such that $\sum_{n \geq 0} a_n \|g^{d_n}\|_Y < \infty$. We find that g belongs to G^J and satisfies $g_{d_n} > 0$ almost surely for all $d_n \in d$.

It remains to show that for all t , $g_t > 0$ P a.s. Assume that for some T outside the set of dates $\{d_n ; n \in N\}$ we have just considered, the event $B_T \equiv \{g_T = 0\}$ has positive P -probability; according to *Assumption A*, we know that there exists $\Phi \in J$ such that $\Phi_T = 0$ outside B_T , $\Phi_t = 0 \quad \forall t < T$, $\Phi_t \geq 0 \quad \forall t > T$ and $\exists d_n \in d$, $P[\Phi_{d_n} > 0] > 0$. For this particular investment $\Phi \in J$, we would have $\langle \Phi, g \rangle_{X,Y} \geq E[\Phi_{d_n} g_{d_n}] > 0$: a contradiction. $\square$

Proof of Lemma 3.1 Since C' satisfies *Assumption A*, and C' is the convex cone generated in X by C and $\Phi \equiv (\Phi_t)_{t \geq 0}$, a price $(-\Phi_0)$ is a fair price for $\check{\Phi}$ if and only if there exists $g \in \mathcal{G}^C$ satisfying $E[\sum_{t \geq 0} g_t \Phi_t] \leq 0$ or, using the strict positivity of g , $(-\Phi_0) \geq \left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$. $\square$

Proof of Corollary 3.2 Since $\left\{ \frac{g}{g_0}, g \in \mathcal{G}^C \right\}$ is a convex set, if $\left\langle \check{\Phi}, \frac{g^1}{g_0^1} \right\rangle_{\check{X}, \check{Y}} \leq -\Phi_0 \leq \left\langle \check{\Phi}, \frac{g^2}{g_0^2} \right\rangle_{\check{X}, \check{Y}}$ for $g^1, g^2 \in \mathcal{G}^C$, then there exists $g \in \mathcal{G}^C$, $g_0 = 1$, such that $-\Phi_0 = \left\langle \check{\Phi}, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$. $\square$

Proof of Corollary 3.3 Immediate using Lemma 3.1. $\square$

Proof of Corollary 3.4 Immediate applying Corollary 3.3. $\square$

Proof of Lemma 3.6 The proof is adapted from Kreps (1981) and Jouini and Kallal (1995a). We shall repeatedly use the fact (F) that by a standard diagonalization

procedure, there exists a sequence (Φ^n, m^n) , $\Phi^n \geq m^n \rightarrow_{\check{X}} m$, for which $\bar{\pi}(m) = \lim_n (-\Phi_0^n)$.

By definition, for all $m \in M$, $\bar{\pi}(m) < \infty$. If there is no free lunch, for all $g \in \mathcal{G}^J$, we have $\bar{\pi}(m) \geq \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$ for all $m \in M$; indeed, assume that there exists a sequence (Φ^n, m^n) in $\check{J} \times M$ such that $\Phi_t^n \geq m_t^n \forall t > 0$, $m^n \rightarrow_{\check{X}} m$, then for all $g \in \mathcal{G}^J$, $-\Phi_0^n \geq \left\langle m^n, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}} \rightarrow_n \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$, so that using (F), $\bar{\pi}(m) \geq \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$. In particular, this implies that for all $m \in M$, $\bar{\pi}(m) > -\infty$ and for all $m \neq 0$ belonging to $\check{X}_+ \cap M$, $\bar{\pi}(m) > 0$.

Since J is a convex cone, it is easy to see that M is also a convex cone. Using (F), it is immediate that $\bar{\pi}$ is such that for all m_1, m_2 in M and all $\lambda \in R_+^*$, we have $\bar{\pi}(m_1 + m_2) \leq \bar{\pi}(m_1) + \bar{\pi}(m_2)$ and $\bar{\pi}(\lambda m_1) = \lambda \bar{\pi}(m_1)$. By definition of $\bar{\pi}$, we have $\bar{\pi}(0) \leq 0$; we have seen that for all $g \in \mathcal{G}^J$, $\bar{\pi}(m) \geq \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$ for all $m \in M$, thus $\bar{\pi}(0) = 0$.

Let us show that $\bar{\pi}$ is l.s.c. Let $\lambda \in R$ and (m^n) be a sequence in M converging to $m \in M$ such that $\bar{\pi}(m^n) \leq \lambda$ for all $n \geq 0$. Then, using (F), for all $n \geq 0$, there exists (Φ^n, m^{*n}) in $J \times M$, such that $\|m^n - m^{*n}\|_{\check{X}} \leq 1/n$, $\Phi_t^n \geq m_t^{*n} \forall t > 0$ and $-\Phi_0^n \leq \lambda + 1/n$. Since m^{*n} converges to m , we must then have $\bar{\pi}(m) \leq \lambda$ and the set $\{m \in M; \bar{\pi}(m) \leq \lambda\}$ is closed. $\square$

Proof of Proposition 3.7 We show that $(M, \bar{\pi})$ satisfies the assumptions of Corollary B.2 in Appendix B. If there is no free lunch, $\bar{\pi}$ is an l.s.c. functional on the convex cone M (Lemma 3.6). By definition of M and $\bar{\pi}$, we have $\check{X}_- \subseteq M$ and $\bar{\pi} \leq 0$ on $\check{X}_-$. Since there is no free lunch for J , $\mathcal{G}^J \neq \emptyset$ and for all $g \in \mathcal{G}^J$, $\bar{\pi}(m) \geq \left\langle m, \frac{g}{g_0} \right\rangle_{\check{X}, \check{Y}}$, hence there exists a positive continuous linear functional on $\check{X}$, whose restriction to M lies below $\bar{\pi}$. We can apply Corollary B.2, and we obtain that for all $m \in M$, $\bar{\pi}(m) = \sup\{l(m), l \in \check{Y}, l > 0, l|_M \leq \bar{\pi}\}$. It is then easy to verify that a positive $l \in \check{Y}$ satisfies $l|_M \leq \bar{\pi}$ if and only if it is of the form $l = \left(\frac{g_t}{g_0} \right)_{t>0}$ for some $g \in \mathcal{G}^J$. Indeed, we have seen in the proof of Lemma 3.6 that any $g \in \mathcal{G}^J$, $g_0 = 1$ satisfies $g|_M \leq \bar{\pi}$; conversely, if $l|_M \leq \bar{\pi}$, then for all $\Phi \in J$, $E[\sum_{t>0} l_t \Phi_t] \leq -\Phi_0$ and letting $l_0 = 1$, $(l_t)_{t \geq 0}|_J \leq 0$. $\square$

Proof of Lemma 4.3 If there is an arbitrage opportunity, then there exists an available investment (τ, B, Φ) for which $\Phi_t - c_t^\Phi \geq 0$ for all $t \geq 0$, hence $\Phi_\tau \geq c_\tau^\Phi \geq \varepsilon^{\tau, B}$ on B and $\Phi_t \geq 0$ for all $t \geq 0$, so that $\Phi \in \mathcal{I}^{\tau, B} \cap \mathcal{A}^{\tau, B}$.

Conversely, suppose that there exists $\Phi \in \mathcal{I}^{\tau, B} \cap \mathcal{A}^{\tau, B}$. Then there exists $\varepsilon^\Phi \in R_+^*$ such that $\Phi_\tau \geq \varepsilon^\Phi$. The investment process $\lambda \Phi$ with λ such that $\lambda \varepsilon^\Phi \geq$

C enables us to get enough at the initial stopping time to cover, through wealth transfer, present and future transaction costs. $\square$

Proof of Theorem 4.5 Using Lemma 4.3, it is easy to see that there is no free lunch if and only if for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, $\overline{K^{\tau,B} - L_+^1} \cap A^B = \emptyset$, where $K^{\tau,B} \equiv \{\sum_{t \geq 0} \Phi_t; \Phi \in \mathcal{J}^{\tau,B}\}$, $A^B \equiv \{f \in L^1; \exists \varepsilon > 0, f \geq \varepsilon \text{ on } B\}$ and the bar denotes the closure in $L^1(\Omega, R)$. Assume first the existence of a family of absolutely continuous probability measures like in the theorem. Let u belong to $\overline{K^{\tau,B} - L_+^1}$. Then there exist sequences $(u_n)_{n \geq 0}$ and $(m_n)_{n \geq 0}$ such that $u_n \leq m_n$, $m_n \in K^{\tau,B}$ and $u_n \xrightarrow{L^1} u$. Since $E^{P^{\tau,B}}[m_n] \leq 0$, we have $E^{P^{\tau,B}}[u_n] \leq 0$ and since $P^{\tau,B}$ has bounded density, we have $E^{P^{\tau,B}}[u_n] \xrightarrow{n \rightarrow \infty} E^{P^{\tau,B}}[u]$. Then $E^{P^{\tau,B}}[u] \leq 0$ and it is not possible to have $u \geq \varepsilon$ on B for some positive real number ε .

Conversely, assume now that for all (τ, B) in $\mathcal{S}^f \times F_\tau$, we have $\overline{K^{\tau,B} - L_+^1} \cap A^B = \emptyset$. Since $\mathcal{J}^{\tau,B}$ is a convex cone, the set $K^{\tau,B}$ is also a convex cone and we can apply a strict separation theorem in L^1 to the closed convex cone $\overline{K^{\tau,B} - L_+^1}$ and $\{1_B\}$ to find $g^{\tau,B}$ in L^∞ and two real numbers α and β with $\alpha < \beta$ such that $g^{\tau,B}|_{\overline{K^{\tau,B} - L_+^1}} \leq \alpha < \beta < \langle 1_B, g^{\tau,B} \rangle$. It is easy to see that $g^{\tau,B} \geq 0$, that we can take $\alpha = 0$, that $g^{\tau,B} \neq 0$ on B and that $g^{\tau,B}|_{K^{\tau,B}} \leq 0$. Letting then $P^{\tau,B}$ be given by $dP^{\tau,B}/dP \equiv \frac{1_B g^{\tau,B}}{E[1_B g^{\tau,B}]}$, we get the result wanted. $\square$

Proof of Theorem 4.6 Assume first that there exist a family of probability measures and an associated family of price processes like in the theorem. Then, according to the proof of Theorem 4.5, and adopting the same notations, we only need to prove that for all $(\tau, B) \in \mathcal{S}^f \times F_\tau$, for all random variables u in $K^{\tau,B}$, $E^{P^{\tau,B}}[u] \leq 0$. Using the specific form of $K^{\tau,B}$, we are reduced to proving that $E^{P^{\tau,B}}[\theta(Z_{\tau_2}^k - Z_{\tau_1}^k)] \leq 0$ for all $\tau_1, \tau_2 \in \mathcal{S}_\tau^f$, $k \in \{1, \dots, n\}$ and $\theta \in L^\infty(\Omega, F_{\tau_1 \wedge \tau_2}, P)$. For such θ , we have

$$E^{P^{\tau,B}}[\theta(Z_{\tau_2}^k - Z_{\tau_1}^k)] \leq E^{P^{\tau,B}}\left[\theta E^{P^{\tau,B}}\left[(S_{\tau_2}^{\tau,B})^k - (S_{\tau_1}^{\tau,B})^k \mid F_{\tau_1 \wedge \tau_2}\right]\right].$$

By the optional sampling theorem (see e.g. Karatzas and Shreve (1988)), we obtain that

$$E^{P^{\tau,B}}\left[(S_{\tau_2}^{\tau,B})^k \mid F_{\tau_1 \wedge \tau_2}\right] = (S_{\tau_1 \wedge \tau_2}^{\tau,B})^k = E^{P^{\tau,B}}\left[(S_{\tau_1}^{\tau,B})^k \mid F_{\tau_1 \wedge \tau_2}\right].$$

For the converse implication, we assume that there is no free lunch, so we know from Theorem 4.5 that for all (τ, B) in $\mathcal{S}^f \times F_\tau$, there exists an absolutely continuous probability measure $P^{\tau,B}$ with bounded density such that $P^{\tau,B}(B) = 1$ and for all $\Phi \in \mathcal{J}^{\tau,B}$, $E^{P^{\tau,B}}[\sum_{t \geq 0} \Phi_t] \leq 0$. For all $k \in \{1, \dots, n\}$, for any stopping

times τ_1 and τ_2 in $\mathcal{S}_\tau^f$ and for all A in $F_{\tau_1 \wedge \tau_2}$, the investment process $\Phi^{(k; 1_A, \tau_1, \tau_2)} \in \mathcal{J}^{\tau, B}$ and we get that $E^{P^{\tau, B}} [-Z_{\tau_1}^k + Z_{\tau_2}^k \mid F_{\tau_1 \wedge \tau_2}] \leq 0$, thus

$$E^{P^{\tau, B}} [Z_{\tau_2}^k \mid F_{\tau_1 \wedge \tau_2}] \leq E^{P^{\tau, B}} [Z_{\tau_1}^k \mid F_{\tau_1 \wedge \tau_2}]. \quad (\text{A.1})$$

For all $\nu \in \mathcal{S}_\tau^f$, we consider the two n -dimensional families $(\tilde{Z}'_\nu)_{\nu \in \mathcal{S}_\tau^f}$ and $(\tilde{Z}_\nu)_{\nu \in \mathcal{S}_\tau^f}$ given by

$$\begin{aligned} \tilde{Z}'_\nu &= \text{ess sup}_{\kappa \in \mathcal{S}_\nu^f} E^{P^{\tau, B}} [Z'_\kappa \mid F_\nu] \\ \tilde{Z}_\nu &= \text{ess inf}_{\kappa \in \mathcal{S}_\nu^f} E^{P^{\tau, B}} [Z_\kappa \mid F_\nu]. \end{aligned}$$

In words, $\tilde{Z}'_\nu$ is the supremum of the conditional expected value of the proceeds from the strategies that consist of going short in the security k (and investing the proceeds in security 0) after the stopping time ν . The random variable $\tilde{Z}_\nu$ is defined symmetrically.

It is a standard result in optimal stopping that for all κ in $\mathcal{S}_\nu^f$

$$\begin{aligned} E^{P^{\tau, B}} [\tilde{Z}'_\kappa \mid F_\nu] &\leq \tilde{Z}'_\nu \\ E^{P^{\tau, B}} [\tilde{Z}_\kappa \mid F_\nu] &\geq \tilde{Z}_\nu. \end{aligned}$$

Now, taking $\nu \equiv s \vee \tau$ and $\kappa \equiv t \vee \tau$ for all (s, t) for which $s \leq t$, we obtain that the process $(\tilde{Z}'_{t \vee \tau})_{t \geq 0}$ is a $P^{\tau, B}$ -supermartingale for $(F_{t \vee \tau})_{t \geq 0}$ and that the process $(\tilde{Z}_{t \vee \tau})_{t \geq 0}$ is a $P^{\tau, B}$ -submartingale for $(F_{t \vee \tau})_{t \geq 0}$. Using inequality (A.1), we have $\tilde{Z}'_{t \vee \tau} \leq \tilde{Z}_{t \vee \tau}$. Now, using Lemma 3 in Jouini and Kallal (1995b) or Proposition 2.6 in Choulli and Stricker (1997), we get that there is a process $S^{\tau, B}$ lying between $(\tilde{Z}_{t \vee \tau})_{t \geq 0}$ and $(\tilde{Z}'_{t \vee \tau})_{t \geq 0}$ on B , which is a $P^{\tau, B}$ -martingale for $(F_{t \vee \tau})_{t \geq 0}$.

By definition, we have $Z' \leq \tilde{Z}'$ and $\tilde{Z} \leq Z$ after τ and on B , so that after τ and on B , $Z' \leq \tilde{Z}' \leq \tilde{Z} \leq Z$. The process $S^{\tau, B}$ is then automatically between Z' and Z , after τ and on B , which completes the proof. $\square$

Proof of Proposition 4.8 We have assumed that there is no arbitrage in the primitive market, so that if Φ and Ψ in $\mathcal{I}^{0, \Omega}$ are such that for all $t \in]0, T]$, $\Phi_t = \Psi_t$, then $\Phi_0 = \Psi_0$. We define on M a linear functional l given by $l(\Phi_T) = \Phi_0$. Now it is easy to see that for all H in M ,

$$\lim_{\lambda \rightarrow +\infty} \frac{\pi_c(\lambda H)}{\lambda} = \lim_{\lambda \rightarrow +\infty} \frac{-\pi_c(-\lambda H)}{\lambda} = l(H).$$

Since there is no arbitrage, we must have $p(H) \geq -p(-H)$ so that

$$-\pi_c(-H) \leq -p(-H) \leq p(H) \leq \pi_c(H),$$

and the price functional p can be written as the sum of a continuous linear functional and a fixed cost, i.e., for all H , $p(H) = l(H) + c(H)$ where $c(\lambda H)/\lambda \rightarrow_{\lambda \rightarrow \infty} 0$. Notice that $c(H) \equiv p(H) - l(H) \leq \pi_c(H) - l(H) \leq C$.

Consequently, in the absence of free lunch, the fair price $p(H)$ associated with any attainable contingent claim H is given by

$$p(H) = E^{P^*}(H) + c(H)$$

where P^* is any absolutely continuous martingale measure. $\square$

Appendix B

Lemma B.1 Any l.s.c. sublinear functional s on a convex cone $K \subseteq \check{X}$ can be written as the supremum over all continuous linear functionals on $\check{X}$, whose restriction to K lies below s , i.e. for all $k \in K$, $s(k) = \sup_{\substack{l \in \check{Y} \\ l|_K \leq s}} l(k)$.

Proof We adapt the proof of the Fenchel–Moreau Theorem. Let

$$t(k) \equiv \sup \left\{ l(k), l \in \check{Y}, l|_K \leq s \right\}.$$

It is immediate that for all $k \in K$, $s(k) \geq t(k)$. Suppose that there exists $k_0 \in K$, such that $t(k_0) < s(k_0)$. Let $A \equiv \{(z, \lambda) \in K \times R, s(z) \leq \lambda\}$. Since s is sublinear, A is a convex cone. Then the closure of A in $\check{X} \times R$, denoted by $\bar{A}$, is a closed convex cone. Since s is l.s.c., $(k_0, t(k_0)) \notin \bar{A}$. By the Hahn–Banach Separation Theorem, there exists a continuous linear functional φ defined on $\check{X} \times R$ and $\alpha \in R$ such that

$$\varphi(k_0, t(k_0)) < \alpha \leq \varphi(z, \lambda) \text{ for all } (z, \lambda) \in \bar{A}. \quad (\text{B.1})$$

The set $\bar{A}$ being a cone, we can take $\alpha = 0$. Hence there exist a continuous linear functional φ_1 on $\check{X}$ and $\beta \in R$ for which $\varphi_1(k_0) + \beta[t(k_0)] < 0 \leq \varphi_1(z) + \beta\lambda$ for all $(z, \lambda) \in \bar{A}$. By taking $z \in D(s)$, i.e. z such that $s(z) < \infty$, and $\lambda = n \rightarrow \infty$ in the preceding inequality, we see that $\beta \geq 0$.

Consider first the case $s \geq 0$. Let $\varepsilon \in R_+^*$. Noting that by definition of A , for all $z \in D(s)$, $(z, s(z)) \in A$, we get $\varphi_1(z) + (\beta + \varepsilon)s(z) \geq 0$. This implies that the continuous linear functional $-\frac{1}{(\beta + \varepsilon)}\varphi_1$ lies below s on K , and by definition of t , $t(k_0) \geq -\frac{1}{(\beta + \varepsilon)}\varphi_1(k_0)$. This leads to $\varphi_1(k_0) + (\beta + \varepsilon)t(k_0) \geq 0$ for all $\varepsilon > 0$, which contradicts (B.1).

For a general s , consider the functional $\bar{s} \equiv s - f_0$, where f_0 is some continuous linear functional lying below s on K (the condition $D(s) \neq \emptyset$ ensures its existence). The functional $\bar{s}$ is a nonnegative l.s.c. sublinear functional on K

such that $D(\bar{s}) \neq \emptyset$. The first part of the proof may be applied and we know that $\bar{t}(k) \equiv \sup\{l(k), l \in \check{Y}, l|_K \leq \bar{s}\} = \bar{s}(k)$. It is clear that $\bar{t} = t - f_0$, hence $s = t$ on K . $\square$

Corollary B.2 *With the same notations as in Lemma B.1, if $K \supseteq \check{X}_-$ and $s \leq 0$ on $\check{X}_-$, then for all $k \in K$, $s(k) = \sup\{l(k), l \in \check{Y}_+, l|_K \leq s\}$. Moreover, if there exists $f \in \check{Y}$, $f > 0$, $f|_K \leq s$, then $s(k) = \sup\{l(k), l \in \check{Y}, l > 0, l|_K \leq s\}$.*

Proof Let $l \in \check{Y}$, $l|_K \leq s$. If $K \supseteq \check{X}_-$ and $s \leq 0$ on $\check{X}_-$, then for all $x \in \check{X}_-$, $\langle x, l \rangle_{\check{X}, \check{Y}} \leq 0$, which means that $l \in \check{Y}_+$. Now, suppose that $L \equiv \{f \in \check{Y}, f > 0, f|_K \leq s\} \neq \emptyset$. Let $f \in L$. For all $l \in \check{Y}_+, l|_K \leq s$, $(\frac{1}{n}f + (1 - \frac{1}{n})l)$ is a sequence of elements of L , and for all $k \in K$, $\langle k, \frac{1}{n}f + (1 - \frac{1}{n})l \rangle \rightarrow_n \langle k, l \rangle$. $\square$

References

- Adler, I. and Gale, D. (1997), Arbitrage and growth rate for riskless investments in a stationary economy *Math. Fin.* **2**, 73–81.
- Back, K. and Pliska, S.R. (1990), On the fundamental theorem of asset pricing with an infinite state space *J. Math. Econ.*, **20**, 1–18.
- Bensaid, B., Lesne, J.-P., Pagès, H. and Scheinkman, J. (1992), Derivative asset pricing with transaction costs *Math. Fin.* **2**, 63–86.
- Choulli, T. and Stricker, C. (1997), *Séparation d'une sur- et d'une sousmartingale par une martingale*. Thèse de T. Choulli. Université de Franche-Comté.
- Cvitanić, J. and Karatzas, I. (1993), Hedging contingent claims with constrained portfolios *Ann. App. Prob.* **3**(3), 652–81.
- Cvitanić, J. and Karatzas, I. (1996), Hedging and portfolio optimization under transaction costs: a martingale approach *Math. Fin.* **6**, 133–66.
- Dalang, R.C., Morton, A. and Willinger, W. (1989), Equivalent martingale measures and no arbitrage in stochastic securities market models *Stochastics and Stochastic Rep.* **29**, 185–202.
- Debreu, G. (1959), *Theory of Value*. Wiley, New York.
- Delbaen, F. (1992), Representing martingale measures when asset prices are continuous and bounded *Math. Fin.* **2**, 107–30.
- Delbaen, F., Kabanov, Y. and Valkeila, E. (2001), Hedging under transaction costs in currency markets: a discrete-time model. To appear in *Math. Fin.*
- Delbaen, F. and Schachermayer, W. (1994), A general version of the fundamental theorem of asset pricing *Math. Ann.* **300**, 463–520.
- Delbaen, F. and Schachermayer, W. (1998), The fundamental theorem of asset pricing for unbounded stochastic processes. *Math. Ann.* **312**, 215–50.
- Duffie, D. and Huang, C. (1986), Multiperiod security markets with differential information: martingales and resolution times *J. Math. Econ.* **15**, 283–303.
- Dybvig, P. and Ross, S. (1987), Arbitrage, in: Eatwell, J., Milgate, M. and Newman, P., eds., *The New Palgrave: A Dictionary of Economics*, vol. 1. Macmillan, London, 100–6.
- El Karoui, N. and Quenez, M.-C. (1995), Dynamic programming and pricing of contingent claims in an incomplete market *SIAM J. Control and Optimization* **33**, 29–66.

- Föllmer, H. and Kramkov, K. (1997), Optional decomposition under constraints *Prob. Theory Relat. Fields* **109**, 1–25.
- Harrison, M. and Kreps, D. (1979), Martingales and arbitrage in multiperiod security markets *J. Econ. Theory* **20**, 381–408.
- Harrison, M. and Pliska, S. (1981), Martingales and stochastic integrals in the theory of continuous trading *Stochastic Processes Appl.* **11**, 215–60.
- Jacod, J. (1979), *Calcul Stochastique et Problèmes de Martingales*. Springer, Berlin.
- Jouini, E. (2000), Price functionals with bid–ask spreads. An axiomatic approach. *J. Math. Econ.* **34**, 547–58.
- Jouini, E. and Kallal, H. (1995a), Martingales and arbitrage in securities markets with transaction costs *J. Econ. Theory* **66**, 178–97.
- Jouini, E. and Kallal, H. (1995b), Arbitrage in securities markets with short-sales constraints *Math. Fin.* **5**, 197–232.
- Jouini, E. and Kallal, H. (1999), Viability and equilibrium in securities markets with frictions *Math. Fin.* **9**(3), 275–92.
- Jouini, E., Kallal, H. and Napp, C. (2000), Arbitrage and viability in securities markets with fixed transaction costs. To appear in *J. Math. Econ.*
- Jouini, E. and Napp, C. (2001), Arbitrage and investment opportunities. To appear in *Finance and Stochastics*.
- Jouini, E., Napp, C. and Schachermayer, W. (2000), Arbitrage and state price deflators in a general intertemporal framework. Preprint.
- Kabanov, Y. (1999), Hedging and liquidation under transaction costs in currency markets *Finance and Stochastics* **3**(2), 237–48.
- Karatzas, I. and Shreve, S. (1988), *Brownian Motion and Stochastic Calculus*, (Graduate Texts in Mathematics, Vol. 113), Springer-Verlag, Berlin.
- Koehl, P.-F. and Pham, H. (2000), Sublinear price functionals under portfolio constraints *J. Math. Econ.* **33**(3), 339–51.
- Kreps, D. (1981), Arbitrage and equilibrium in economies with infinitely many commodities *J. Math. Econ.* **8**, 15–35.
- Lakner, P. (1993), Martingale measures for a class of right-continuous processes *Math. Fin.* **3**(1), 43–53.
- Napp, C. (2000), Pricing issues with investment flows. Applications to market models with frictions. To appear in *J. Math. Econ.*
- Schachermayer, W. (1994), Martingale measures for discrete time processes with infinite horizon *Math. Fin.* **4**, 25–55.
- Stricker, C (1990), Arbitrage et lois de martingale. *Ann. Inst. Henri Poincaré*, vol. **26**, 451–60.
- Yan, J.A. (1980), Caractérisation d’une classe d’ensembles convexes de L^1 ou H^1 . Sémin. de Probabilités. *Lecture notes in Mathematics* XIV 784, 220–2

3

American Options: Symmetry Properties

Jérôme Detemple

1 Introduction

Put–call symmetry (PCS) holds when the price of a put option can be deduced from the price of a call option by relabeling its arguments. For instance, in the context of the standard financial market model with constant coefficients the value of an American put equals the value of an American call with strike price S , maturity date T , in a financial market with interest rate δ and in which the underlying asset price pays dividends at the rate r . This result was originally demonstrated by McDonald and Schroder (1990, 1998) using a binomial approximation of the lognormal model and by Bjerk Sund and Stensland (1993) in the continuous time model using PDE methods; it is a version of the international put–call equivalence (Grabbe (1983)).

Put–call symmetry is a useful property of options since it reduces the computational burden in implementations of the model. Indeed, a consequence of the property is that the same numerical algorithm can be used to price put and call options and to determine their associated optimal exercise policy. Another benefit is that it reduces the dimensionality of the pricing problem for some payoff functions. Examples include exchange options and quanto options. PCS also provides useful insights about the economic relationship between contracts. Puts and calls, forward prices and discount bonds, exchange options and standard options are simple examples of derivatives that are closely connected by symmetry relations.

Some intuition for PCS is based on the properties of the normal distribution. Indeed, in the model with constant coefficients the distribution of the terminal stock price is lognormal. Symmetry of the put and call option payoff function combined with the symmetry of the normal distribution then suggest that the put and call values can be deduced from each other by interchanging the arguments of the pricing functions. This can be verified directly from the valuation formulas for standard European and American options. As demonstrated by Gao, Huang and Subrahmanyam (2000) it is also true for European and American barrier options,

such as down and out call and up and out put options, in the model with constant coefficients.

Since option values depend only on the volatility of the underlying asset price it seems reasonable to conjecture that PCS will hold in diffusion models in which the drift is an arbitrary function of the asset price but the volatility is a symmetric function of the price. This intuition is exploited by Carr and Chesney (1994) who show that PCS indeed extends to such a setting. Since alternative assumptions about the behavior of the underlying asset price destroy the symmetry of the terminal price distribution it would appear that the property cannot hold in more general contexts. Somewhat surprisingly, Schroder (1999), relying on a change of numeraire introduced by Geman, El Karoui and Rochet (1995), is able to show that the result holds in very general environments including models with stochastic coefficients and discontinuous underlying asset price processes.¹

This chapter surveys the latest results in the field and provides further extensions. Our basic market structure is one in which the underlying asset price follows an Itô process with progressively measurable coefficients (including the dividend rate) and the interest rate is an adapted stochastic process. We show that a version of PCS holds under these general market conditions. One feature behind the property is the homogeneity of degree one of the put and call payoff functions with respect to the stock price and the exercise price. For such payoffs the standard symmetry property of prices follows from a simple change of measure which amounts to taking the asset price as numeraire.

The identification of the change of numeraire as a central feature underlying the standard PCS property permits the extension of the result to more complex contracts which involve liquidation provisions. A *random maturity* option is an option (put or call) which is automatically liquidated at a prespecified random time and, in such an event, pays a prespecified random cash flow. A typical example is a down and out put option with barrier L . This option expires automatically if the underlying asset price hits the level L (null liquidation payoff), but pays off $(K - S)^+$ if exercised prior to expiration. Put–call symmetry for random maturity options states that the value of an American put with strike price K , maturity date T , automatic liquidation time τ_l and liquidation payoff H_{τ_l} equals the value of an American call with strike S , maturity date T , automatic liquidation time τ_l^* and liquidation payoff $H_{\tau_l}^*$ in an auxiliary financial market with interest rate δ and in which the underlying asset price pays dividends at the rate r and has initial value K . The liquidation characteristics τ_l^* and $H_{\tau_l}^*$ of the equivalent call can be expressed in terms of the put specifications K , τ_l and H_{τ_l} and the initial value of the underlying

¹ Symmetry results in general market environments are also reported in Kholodnyi and Price (1998). Their proofs are based on no-arbitrage arguments and use operator theory and group theory notions.

asset S . For a down and out put option with barrier L which has characteristics

$$\tau_L = \inf\{t \in [0, T] : S_t = L\} \text{ and } H_{\tau_L} = 0$$

the equivalent up and out call has characteristics

$$\tau_L^* = \tau_{L^*} = \inf\left\{t \in [0, T] : S_t^* = L^* \equiv \frac{KS}{L}\right\} \text{ and } H_{\tau_L^*}^* = 0,$$

where S^* denotes the price of the underlying asset in the auxiliary financial market.

Contingent claims which are written on multiple assets also exhibit symmetry properties when their payoff is homogeneous of degree one. In fact the same change of measure argument as in the one asset case identifies classes of contracts which are related by symmetry and therefore can be priced off each other. In particular, for contracts on two underlying assets, we show that American call max-options are symmetric to American options to exchange the maximum of an asset and cash against another asset, that American exchange options are symmetric to standard call or put options (on a single underlying asset) and that American capped exchange options with proportional cap are symmetric to both capped call options with constant caps and capped put options with proportional caps. In all of these relationships the symmetric contract is valued in an auxiliary financial market with suitably adjusted interest rate and underlying asset prices.

We then discuss extensions of the property to a class of contracts analyzed recently in the literature, namely occupation time derivatives. These contracts, typically, depend on the amount of time spent by the underlying asset price in certain prespecified regions of the state space. Examples of such path-dependent contracts are Parisian and cumulative barrier options (Chesney, Jeanblanc-Picqué and Yor (1997)), step options (Linetsky (1999)) and quantile options (Miura (1992)). More general payoffs based on the occupation time of a constant set, above or below a barrier, are discussed in Hugonnier (1998). While the literature has focused exclusively on European-style contracts in the context of models with geometric Brownian motion price processes, we consider American-style occupation time derivatives in models with Itô price processes. We also allow for occupation times of random sets. We show that occupation time derivatives with homogeneous payoff functions satisfy a symmetry property in which the symmetric contract depends on the occupation time of a suitably adjusted random set. Extensions to multiasset occupation time derivatives are also presented.

Symmetry-like properties also hold when the contract under consideration is homogeneous of degree $\nu \neq 1$. In this instance the interest rate in the auxiliary economy depends on the coefficient ν , the interest rate in the original economy and the dividend rate and volatility coefficients of the numeraire asset in the original

economy. The dividend rates of other assets in the new numeraire are also suitably adjusted.

Since symmetry properties reflect the passage to a new numeraire asset it is of interest to examine the replicability of attainable payoffs under changes of numeraire. For the case of nondividend paying assets Geman, El Karoui and Rochet (1995) have established that contingent claims that are attainable in one numeraire are also attainable in any other numeraire and that the replicating portfolios are the same. We show that these results extend to the case of dividend-paying assets. This demonstrates that any symmetric contract can indeed be attained in the appropriate auxiliary economy with new numeraire and that its price satisfies the usual representation formula involving the pricing measure and the interest rate that characterize the auxiliary economy.

The second section reviews the property in the context of the standard model with constant coefficients. In Section 3 PCS is extended to a financial market model with Brownian filtration and stochastic opportunity set. The markovian model with diffusion price process (and general volatility structure) is examined as a subcase of the general model. Extensions to random maturity options, multiasset contingent claims, occupation time derivatives and payoffs that are homogeneous of degree ν are carried out in Sections 4–7. Questions pertaining to changes of numeraire, replicating portfolios and representation of asset prices are examined in Section 8. Concluding remarks are formulated last.

2 Put–call symmetry in the standard model

We consider the standard financial market model with constant coefficients (constant opportunity set). The underlying asset price, S , follows a geometric Brownian motion process

$$dS_t = S_t[(r - \delta)dt + \sigma d\tilde{z}_t], t \in [0, T]; S_0 \text{ given} \quad (1)$$

where the coefficients (r, δ, σ) are constant. Here r represents the interest rate, δ the dividend rate and σ the volatility of the asset price. The asset price process (1) is represented under the equivalent martingale measure Q : the process $\tilde{z}$ is a Q -Brownian motion.

In this complete financial market it is well known that the price of any contingent claim can be obtained by a no-arbitrage argument. In particular the value of a European call option with strike price K and maturity date T is given by the Black

and Scholes (1973) formula

$$c(S_t, K, r, \delta, t) = S_t e^{-\delta(T-t)} N(d(S_t, K, r, \delta, T-t)) - K e^{-r(T-t)} N(d(S_t, K, r, \delta, T-t) - \sigma \sqrt{T-t}) \quad (2)$$

where

$$d(S, K, r, \delta, T-t) = \frac{\log(S/K) + (r - \delta + \frac{1}{2}\sigma^2)(T-t)}{\sigma \sqrt{T-t}}. \quad (3)$$

Similarly the value of a European put with the same characteristics (K, T) is

$$p(S_t, K, r, \delta, t) = K e^{-r(T-t)} N(-d(S_t, K, r, \delta, T-t) + \sigma \sqrt{T-t}) - S_t e^{-\delta(T-t)} N(-d(S_t, K, r, \delta, T-t)). \quad (4)$$

Comparison of these two formulas leads to the following symmetry property:

Theorem 1 (European PCS) *Consider European put and call options with identical characteristics K and T written on an asset with price S given by (1). Let $p(S, K, r, \delta, t)$ and $c(S, K, r, \delta, t)$ denote the respective price functions. Then*

$$p(S, K, r, \delta, t) = c(K, S, \delta, r, t). \quad (5)$$

Proof of Theorem 1 Substituting (K, S, δ, r) for (S, K, r, δ) in (2) and using

$$\begin{aligned} d(K, S, \delta, r, T-t) &= \frac{\log(K/S) + (\delta - r + \frac{1}{2}\sigma^2)(T-t)}{\sigma \sqrt{T-t}} \\ &= -\frac{\log(S/K) + (r - \delta + \frac{1}{2}\sigma^2)(T-t)}{\sigma \sqrt{T-t}} + \sigma \sqrt{T-t} \\ &= -d(S, K, r, \delta, T-t) + \sigma \sqrt{T-t} \end{aligned} \quad (6)$$

gives the desired result. $\square$

This result shows that the put value in the financial market under consideration is the same as the value of a call option with strike price S and maturity date T in an economy with interest rate δ and in which the underlying asset price follows a geometric Brownian motion process with dividend rate r , volatility σ and initial value K , under the risk neutral measure.

This symmetry property between the value of puts and calls is even more striking when we consider American options. For these contracts (Kim (1990), Jacka (1991) and Carr, Jarrow and Myneni (1992)) have shown that the value of a call has the early exercise premium representation (EEP)

$$C(S_t, K, r, \delta, t, B^c(\cdot)) = c(S_t, K, r, \delta, t) + \pi(S_t, K, r, \delta, t, B^c(\cdot)) \quad (7)$$

where $C(S, K, r, \delta, t, B^c(\cdot))$ is the value of the American call, $c(S, K, r, \delta, t)$ represents the value of the European call in (2) and $\pi(S, K, r, \delta, t, B^c(\cdot))$ is the early exercise premium

$$\pi(S_t, K, r, \delta, t, B^c(\cdot)) = \int_t^T \phi(S_t, K, r, \delta, v - t, B_v^c) dv \quad (8)$$

with

$$\begin{aligned} \phi(S_t, K, r, \delta, v - t, B_v^c) = & \delta S_t e^{-\delta(v-t)} N(d(S_t, B_v^c, r, \delta, v - t)) \\ & - r K e^{-r(v-t)} N(d(S_t, B_v^c, r, \delta, v - t) - \sigma \sqrt{v - t}). \end{aligned} \quad (9)$$

The exercise boundary $B^c(\cdot)$ of the call option solves the recursive integral equation

$$B_t^c - K = C(B_t^c, K, r, \delta, t, B^c(\cdot)) \quad (10)$$

subject to the boundary condition $B_T^c = \max(K, \frac{r}{\delta} K)$. Let $B^c(K, r, \delta, t)$ denote the solution.

The EEP representation for the American put can be obtained by following the same approach as for the call. Alternatively the put value can be deduced from the call formula by appealing to the following result (McDonald and Schroder (1998)).

Theorem 2 (American PCS) *Consider American put and call options with identical characteristics K and T written on an asset with price S given by (1). Let $P(S, K, r, \delta, t, B^p(\cdot))$ and $C(S, K, r, \delta, t, B^c(\cdot))$ denote the respective price functions and $B^p(K, r, \delta, \cdot)$ and $B^c(S, r, \delta, \cdot)$ the corresponding immediate exercise boundaries. Then*

$$P(S, K, r, \delta, t, B^p(K, r, \delta, \cdot)) = C(K, S, \delta, r, t, B^c(S, \delta, r, \cdot)) \quad (11)$$

and for all $t \in [0, T]$

$$B^p(K, r, \delta, t) = \frac{SK}{B^c(S, \delta, r, t)}. \quad (12)$$

This result can again be demonstrated by substitution along the lines of the proof of Theorem 1. A more elegant approach relies on a change of measure detailed in the next section.

Hence, even for American options the value of a put is the same as the value of a call with strike S , maturity date T , in an economy with interest rate δ and in which the underlying asset price, under the risk neutral measure, follows a geometric

Brownian motion process with dividend rate r , volatility σ and initial value K . Furthermore the exercise boundary for the American put equals the inverse of the exercise boundary for the American call with characteristics (S, δ, r) multiplied by the product SK .

Some intuition for this result rests on the properties of normal distributions. In models with constant coefficients (r, δ, σ) the value of put and call options can be expressed in terms of the cumulative normal distribution. Combining the symmetry of the normal distribution with the symmetry of the put and call payoffs leads to the relationship between the option values and the exercise boundaries.

A priori this intuition may suggest that the property does not extend beyond the financial market model with constant coefficients. As we show next this conjecture turns out to be incorrect.

3 Put–call symmetry with Itô price processes

In this section we demonstrate that a version of PCS holds under fairly general financial market conditions. The key to the approach is the adoption of the stock as a new numeraire. Changes of numeraire have been discussed thoroughly in the literature, in particular in Geman, El Karoui and Rochet (1995). The extension of options' symmetry properties to general uncertainty structures based on this change of numeraire is due to Schroder (1999). This section considers a special case of Schroder, namely a market with Brownian filtration.

Suppose we have an economy with finite time period $[0, T]$, a complete probability space $(\Omega, \mathcal{F}, P)$ and a filtration $\mathcal{F}_{(\cdot)}$. A Brownian motion process z is defined on $(\Omega, \mathcal{F})$ and takes values in $\mathbb{R}$. The filtration is the natural filtration generated by z and $\mathcal{F}_T = \mathcal{F}$.

The financial market has a stochastic opportunity set and nonmarkovian price dynamics. The underlying asset price follows the Itô process,

$$dS_t = S_t[(r_t - \delta_t)dt + \sigma_t d\tilde{z}_t], t \in [0, T]; S_0 \text{ given} \quad (13)$$

under the Q -measure. The interest rate r , the dividend rate δ and the volatility coefficient σ are progressively measurable and bounded processes of the Brownian filtration $\mathcal{F}_{(\cdot)}$ generated by the underlying Brownian motion process z . The process $\tilde{z}$ is a Q -Brownian motion.

At various stages of the analysis we will also be led to consider an alternative financial market with interest rate δ , in which the underlying asset price S^* satisfies

$$dS_t^* = S_t^*[(\delta_t - r_t)dt + \sigma_t dz_t^*], t \in [0, T]; S_0^* \text{ given} \quad (14)$$

under some risk neutral measure Q^* . In this market the asset has dividend rate r and volatility coefficient σ . The process z^* is a Brownian motion under the pricing measure Q^* . Both z^* and Q^* will be specified further as we proceed.

We first state a relationship between the values of European puts and calls in the general financial market model under consideration.

Theorem 3 (Generalized European PCS) *Consider a European put option with characteristics K and T written on an asset with price S given by (13) in the market with stochastic interest rate r . Let $p(S, K, r, \delta; \mathcal{F}_t)$ denote the put price process. Then*

$$p(S_t, K, r, \delta; \mathcal{F}_t) = c(S_t^*, S, \delta, r; \mathcal{F}_t) \quad (15)$$

where $c(S_t^*, S, \delta, r; \mathcal{F}_t)$ is value of a call with strike price $S = S_t$ and maturity date T in a financial market with interest rate δ and in which the underlying asset price follows the Itô process (14) for $v \in [t, T]$ with initial value $S_t^* = K$ and with z^* defined by

$$dz_v^* = -d\tilde{z}_v + \sigma_v dv \quad (16)$$

for $v \in [0, T]$, with $z_0^* = 0$.

This result extends the PCS property of the previous section to nonmarkovian economies with Itô price processes and progressively measurable interest rates. The key behind this general equivalence is a change of measure, detailed in the proof, which converts a put option in the original economy into a call option with symmetric characteristics in the auxiliary economy. Note that the equivalence is obtained by switching (S, K, r, δ) to (S^*, S, δ, r) , but keeping the trajectories of the Brownian motion the same, i.e. the filtration which is used to compute the value of the call in the auxiliary financial market is the one generated by the original Brownian motion z . Thus information is preserved across economies. In effect the change of measure creates a new asset whose price is the inverse of the original asset price adjusted by a multiplicative factor which depends only on the initial conditions. As we shall see below in the context of diffusion models the change of measure is instrumental in proving the symmetry property without placing restrictions on the volatility coefficient.

Proof of Theorem 3 In the original financial market the value $p_t \equiv p(S_t, K, r, \delta; \mathcal{F}_t)$ of the put option with characteristics (K, T) has the (present value) representation

$$p_t = \tilde{E} \left[\exp \left(- \int_t^T r_v dv \right) \left(K - S_t \exp \left(\int_t^T \alpha_v dv + \int_t^T \sigma_v d\tilde{z}_v \right) \right)^+ \mid \mathcal{F}_t \right]$$

where $\alpha \equiv r - \delta - \frac{1}{2}\sigma^2$ and the expectation is taken relative to the equivalent martingale measure Q . Simple manipulations show that the right hand side of this equation equals

$$\begin{aligned} \tilde{E} \left[\exp \left(- \int_t^T \left(\delta_v + \frac{1}{2} \sigma_v^2 \right) dv + \int_t^T \sigma_v d\tilde{z}_v \right) \right. \\ \left. \times \left(K \exp \left(- \int_t^T \alpha_v dv - \int_t^T \sigma_v d\tilde{z}_v \right) - S_t \right)^+ \mid \mathcal{F}_t \right]. \end{aligned}$$

Consider the new measure

$$dQ^* = \exp \left(- \frac{1}{2} \int_0^T \sigma_v^2 dv + \int_0^T \sigma_v d\tilde{z}_v \right) dQ \quad (17)$$

which is equivalent to Q . Girsanov's Theorem (1960) implies that the process

$$dz_v^* = -d\tilde{z}_v + \sigma_v dv \quad (18)$$

is a Q^* -Brownian motion. Substituting (18) in the put pricing formula and passing to the Q^* -measure yields

$$\begin{aligned} p_t = E^* \left[\exp \left(- \int_t^T \delta_v dv \right) \left(K \exp \left(\int_t^T \left(\delta_v - r_v - \frac{1}{2} \sigma_v^2 \right) dv \right. \right. \right. \\ \left. \left. \left. + \int_t^T \sigma_v dz_v^* \right) - S_t \right)^+ \mid \mathcal{F}_t \right]. \end{aligned} \quad (19)$$

But the right hand side is the value of a call option with strike $S = S_t$, maturity date T in an economy with interest rate δ , asset price with dividend rate r and initial value $S_t^* = K$, and pricing measure Q^* . $\square$

An even stronger version of the preceding result is obtained if the coefficients of the model are adapted to the subfiltration generated by the process z^* . Let $\mathcal{F}_{(\cdot)}^*$ denote the filtration generated by this Q^* -Brownian motion process.

Corollary 4 Suppose that the coefficients (r, δ, σ) are adapted to the filtration $\mathcal{F}_{(\cdot)}^*$. Then

$$p(S_t, K, r, \delta; \mathcal{F}_t) = c(S_t^*, S, \delta, r; \mathcal{F}_t^*)$$

where $c(S_t^*, S, \delta, r; \mathcal{F}_t^*)$ is value of a call with strike price $S = S_t$ and maturity date T in a financial market with information filtration $\mathcal{F}_{(\cdot)}^*$ generated by the Q^* -Brownian motion process (16), interest rate δ and in which the underlying asset price follows the Itô process (14) with initial value $S_t^* = K$.

In the context of this corollary part of the information embedded in the original information filtration generated by the Brownian motion z may be irrelevant for pricing the put option. Since all the coefficients are adapted to the subfiltration generated by z^* this is the only information which matters in computing the expectation under Q^* in (19).

Remark 5 *Note that the standard European PCS in the model with constant coefficients is a special case of this corollary. Indeed in this setting direct integration over z^* leads to the call value in the auxiliary economy and the put value in the original economy.*

Let us now consider the case of American options. For these contracts early exercise, prior to the maturity date T , is under the control of the holder. At any time prior to the optimal exercise time the put value $P_t \equiv P(S_t, K, r, \delta; \mathcal{F}_t)$ in the original economy is (see Bensoussan (1984) and Karatzas (1988))

$$P_t = \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) \left(K - S_t \exp \left(\int_t^\tau (r_v - \delta_v - \frac{1}{2} \sigma_v^2) dv + \int_t^\tau \sigma_v d\tilde{z}_v \right) \right)^+ \mid \mathcal{F}_t \right]$$

where $\mathcal{S}_{t,T}$ denotes the set of stopping times of the filtration $\mathcal{F}_{(\cdot)}$ with values in $[t, T]$. Using the same arguments as in the proof of Theorem 3 we can write

$$P_t = \sup_{\tau \in \mathcal{S}_{t,T}} E^* \left[\exp \left(- \int_t^\tau \delta_v dv \right) \left(K \exp \left(\int_t^\tau (\delta_v - r_v - \frac{1}{2} \sigma_v^2) dv + \int_t^\tau \sigma_v dz_v^* \right) - S_t \right)^+ \mid \mathcal{F}_t \right]$$

where the expectation is relative to the equivalent measure Q^* and conditional on the information $\mathcal{F}_t$. Since the change of measure performed does not affect the set of stopping times over which the holder optimizes the following result holds.

Theorem 6 (Generalized American PCS) *Consider an American put option with characteristics K and T written on an asset with price S given by (13) in the market with stochastic interest rate r . Let $P(S, K, r, \delta; \mathcal{F}_t)$ denote the American put price process and $\tau^p(K, r, \delta)$ the optimal exercise time. Then, prior to exercise, the put price is*

$$P(S_t, K, r, \delta; \mathcal{F}_t) = C(S_t^*, S, \delta, r; \mathcal{F}_t) \quad (20)$$

where $C(S_t^*, S, \delta, r; \mathcal{F}_t)$ is the value of an American call with strike price $S = S_t$ and maturity date T in a financial market with interest rate δ and in which the

underlying asset price follows the Itô process (14) with initial value $S_t^* = K$ and with z^* defined by (16). The optimal exercise time for the put option is

$$\tau^p(S, K, r, \delta) = \tau^c(K, S, \delta, r) \quad (21)$$

where $\tau^c(K, S, \delta, r)$ denotes the optimal exercise time for the call option.

Remark 7 Consider the model with constant coefficients (r, δ, σ) . In this setting the optimal exercise time for the call option in the auxiliary financial market is

$$\tau^c(K, S, \delta, r) = \inf \left\{ t \in [0, T] : K \exp \left(\left(\delta - r - \frac{1}{2} \sigma^2 \right) t + \sigma z_t^* \right) = B^c(S, \delta, r, t) \right\}.$$

On the other hand the optimal exercise time for the put option in the original financial market is

$$\tau^p(S, K, r, \delta) = \inf \left\{ t \in [0, T] : S \exp \left(\left(r - \delta - \frac{1}{2} \sigma^2 \right) t + \sigma \tilde{z}_t \right) = B^p(K, r, \delta, t) \right\}$$

where $B^p(K, r, \delta, t)$ is the put exercise boundary. Using the definition of z^* in (16) we conclude immediately that

$$B^p(K, r, \delta, t) = \frac{SK}{B^c(S, \delta, r, t)}.$$

3.1 Diffusion financial market models

Suppose that the stock price satisfies the stochastic differential equation

$$dS_t = S_t[(r(S_t, t) - \delta(S_t, t))dt + \sigma(S_t, t)d\tilde{z}_t], t \in [0, T]; S_0 \text{ given} \quad (22)$$

under the Q -measure. In this market the interest rate r may depend on the stock price and along with the other coefficients of (22) satisfies appropriate Lipschitz and growth conditions for the existence of a unique strong solution (see Karatzas and Shreve (1988)). We assume that the solution is continuous relative to the initial conditions.

Since this markovian financial market is a special case of the general model of the previous section PCS holds. However, in the model under consideration the exercise regions of options have a simple structure which leads to a clear comparison between the put and the call exercise policies.

Define the discount factor

$$R_{t,s} = \exp \left(- \int_t^s r(S_v, v) dv \right)$$

for $t, s \in [0, T]$ and the Q -martingale

$$M_{t,s} \equiv \exp\left(-\frac{1}{2} \int_t^s \sigma(S_v, v)^2 dv + \int_t^s \sigma(S_v, v) d\tilde{z}_v\right)$$

for $t, s \in [0, T]$, $s \geq t$.

Consider an American call option and let $\mathcal{E}$ denote the exercise set. Continuity of the strong solution of (22) relative to the initial conditions implies that the option price is continuous and that the exercise region is a closed set. Thus we can meaningfully define its boundary B^c .² Let $\mathcal{E}(t)$ denote the t -section of the exercise region. The EEP representation for a call option with strike K and maturity date T is

$$C(S_t, K, r, \delta, t, B^c(\cdot)) = c(S_t, K, r, \delta, t) + \pi(S_t, K, r, \delta, t, B^c(\cdot)) \quad (23)$$

where $C(S, K, r, \delta, t, B^c(\cdot))$ is the value of the American call, $c(S, K, r, \delta, t)$ represents the value of the European call

$$c(S_t, K, r, \delta, t) = \tilde{E}\left[\left(S_t \exp\left(-\int_t^T \delta(S_v, v) dv\right) M_{t,T} - K R_{t,T}\right)^+ \mid S_t\right] \quad (24)$$

and $\pi_t \equiv \pi(S_t, K, r, \delta, t, B^c(\cdot))$ is the early exercise premium

$$\begin{aligned} \pi_t = \tilde{E}\left[\int_t^T \left(\delta(S_v, v) S_t \exp\left(-\int_t^s \delta(S_v, v) dv\right) M_{t,s} \right. \right. \\ \left. \left. - r(S_s, s) K R_{t,s}\right) 1_{\{S_s \in \mathcal{E}(s)\}} ds \mid S_t\right]. \end{aligned} \quad (25)$$

In these expressions dependence on r and δ is meant to represent dependence on the functional form of $r(\cdot)$ and $\delta(\cdot)$. The boundary $B^c(\cdot)$ of the exercise set for the call option solves the recursive integral equation

$$B_t^c - K = C(B_t^c, K, r, \delta, t, B^c(\cdot)) \quad (26)$$

subject to the boundary condition $B_T^c = \max(K, (r(B_T^c, T)/\delta(B_T^c, T))K)$. Let $B^c(K, r, \delta, t)$ denote the solution. The optimal exercise policy for the call is to exercise at the stopping time

$$\tau^c(S, K, r, \delta) = \inf\left\{t \in [0, T] : S R_{0,t}^{-1} \exp\left(-\int_0^t \delta(S_v, v) dv\right) M_{0,t} = B^c(K, r, \delta, t)\right\}. \quad (27)$$

² If the exercise region is up-connected the exercise boundary is unique. Failure of this property may imply the existence of multiple boundaries.

In this context put–call symmetry leads to

Proposition 8 *Consider an American put option with characteristics K and T written on an asset with price S given by (22) in the market with interest rate $r(S, t)$. Let $P(S, K, r, \delta, t)$ denote the American put price process and $\tau^p(S, K, r, \delta)$ the optimal exercise time. Then, prior to exercise, the put price is*

$$P(S_t, K, r, \delta, t) = C(S_t^*, S, \delta, r, t) \quad (28)$$

where $C(S_t^*, S, \delta, r; t)$ is value of an American call with strike price $S = S_t$ and maturity date T in a financial market with stochastic interest rate δ and in which the underlying asset price S^* satisfies the stochastic differential equation

$$dS_v^* = S_v^* \left[\left(\delta \left(\frac{SK}{S_v^*}, v \right) - r \left(\frac{SK}{S_v^*}, v \right) \right) dv + \sigma \left(\frac{SK}{S_v^*}, v \right) dz_v^* \right], \text{ for } v \in [t, T] \quad (29)$$

with initial value $S_t^* = K$ and with z^* defined by (16). The optimal exercise time for the put option is $\tau^p(S, K, r, \delta) = \tau^c(K, S, \delta, r)$ and the exercise boundaries are related by

$$B^p(K, r, \delta, t) = \frac{SK}{B^c(S, \delta, r, t)}. \quad (30)$$

In the financial market setting of (22) all the information relevant for future payoffs is embedded in the current stock price. Any strictly monotone transformation of the price is also a sufficient statistic. Thus the passage from the original economy to the auxiliary economy with stock price (29) preserves the information required to price derivatives with future payoffs. No information beyond the current price S_t^* is required to assess the correct evolution of the coefficients of the underlying asset price process. This stands in contrast with the general model with Itô price processes in which the path of the Brownian motion needs to be recorded in the auxiliary economy for proper evaluation of future distributions.

Note also that the change of measure converts the original underlying asset into a symmetric asset with inverse price up to a multiplicative factor depending only on the initial conditions. Since the change of measure can be performed independently of the structure of the coefficients the results are valid even in the absence of symmetry-like restrictions on the volatility coefficient.

Proof of Proposition 8 The first part of the proposition follows from Theorem 6. To prove the relationship between the exercise boundaries note that the call boundary at maturity equals

$$B^c = \max(K, b^c)$$

where b^c solves the nonlinear equation

$$r\left(\frac{SK}{b^c}, T\right)b^c - \delta\left(\frac{SK}{b^c}, T\right)S = 0.$$

In this expression we used the relation $S_T = SK/S_T^*$. Now with the change of variables $b^p = SK/b^c$ it is clear that b^p solves

$$r(b^p, T)K - \delta(b^p, T)b^p = 0$$

and that the put boundary at the maturity date satisfies (30). To establish the relation prior to the maturity date it suffices to use the recursive integral equation for the call boundary, pass to the Q^* -measure and perform the change of variables indicated. The resulting expression is the recursive integral equation for the put boundary. $\square$

The results in this section can be easily extended to multivariate diffusion models (S, Y) where Y is a vector of state variables impacting the coefficients of the underlying asset price process. Passage to the measure Q^* , in this case, introduces a risk premium correction in the state variables processes. Multivariate models in that class are discussed extensively in Schroder (1999).

4 Options with random expiration dates

We now consider a class of American derivatives which mature automatically if certain prespecified conditions are satisfied. Let τ_l denote a stopping time of the filtration and let $H = \{H_t : t \in [0, T]\}$ denote a progressively measurable process. A call option with maturity date T , strike K , automatic liquidation time τ_l and liquidation payoff H pays $(S - K)^+$ if exercised by the holder at date $t < \tau_l$. If τ_l materializes prior to T the option automatically matures and pays off H_{τ_l} . A random maturity put option with characteristics (K, T, τ_l, H) has similar provisions but pays $(K - S)^+$ if exercised prior to the automatic liquidation time τ_l . Options with such characteristics are referred to as *random maturity options*.

Popular examples of such contracts are barrier options such as down and out put options and up and out call options. Both of these contracts become worthless when the underlying asset price reaches a prespecified level L (i.e. the liquidation payoff is a constant $H = 0$).

Another example is an American capped call option with automatic exercise at the cap L . This option is automatically liquidated at the random time

$$\tau_l = \tau_L \equiv \inf\{t \in [0, T] : S_t = L\}$$

or $\tau_L = \infty$ if no such time materializes in $[0, T]$ and pays off the constant $H =$

$L - K$ in that event. If $\tau_L > T$ the option payoff is $(S - K)^+$.³ Capped options with growing caps and automatic exercise at the cap are examples in which the automatic liquidation payoff is time dependent

Consider again the general financial market model with underlying asset price given by (13). Recall the definitions of the discount factor

$$R_{t,s} = \exp\left(-\int_t^s r_v dv\right)$$

for $t, s \in [0, T]$ and the Q -martingale

$$M_{t,s} \equiv \exp\left(-\frac{1}{2} \int_t^s \sigma_v^2 dv + \int_t^s \sigma_v d\tilde{z}_v\right)$$

for $t, s \in [0, T]$, $s \geq t$.

Let $P_t = P(S, K, T, \tau_l, H, r, \delta; \mathcal{F}_t)$ denote the value of an American random maturity put with characteristics (K, T, τ_l, H) . In this financial market the put value is given by

$$\begin{aligned} P_t = \sup_{\tau \in S_{t,T}} & \tilde{E}\left[R_{t,\tau} \left(K - S_t R_{t,\tau}^{-1} \exp\left(-\int_t^\tau \delta_v dv\right) M_{t,\tau}\right)^+ 1_{\{\tau < \tau_l\}} \right. \\ & \left. + R_{t,\tau_l} H_{\tau_l} 1_{\{\tau \geq \tau_l\}} \mid \mathcal{F}_t\right]. \end{aligned}$$

Performing the same change of measure as in the previous section enables us to rewrite the put value P_t as

$$\begin{aligned} & \sup_{\tau \in S_{t,T}} \tilde{E}\left[\exp\left(-\int_t^\tau \delta_v dv\right) M_{t,\tau} \left[\left(K R_{t,\tau} \exp\left(\int_t^\tau \delta_v dv\right) M_{t,\tau}^{-1} - S_t\right)^+ 1_{\{\tau < \tau_l\}} \right. \right. \\ & \quad \left. \left. + \frac{S_t H_{\tau_l}}{S_{\tau_l}} 1_{\{\tau \geq \tau_l\}} \right] \mid \mathcal{F}_t\right] \\ & = \sup_{\tau \in S_{t,T}} E^*\left[\exp\left(-\int_t^\tau \delta_v dv\right) \left[\left(K R_{t,\tau} \exp\left(\int_t^\tau \delta_v dv\right) M_{t,\tau}^{-1} - S_t\right)^+ 1_{\{\tau < \tau_l\}} \right. \right. \\ & \quad \left. \left. + H_{\tau_l}^* 1_{\{\tau \geq \tau_l\}} \right] \mid \mathcal{F}_t\right] \end{aligned}$$

where we define the stochastic process H^* as

³ Note that, in the case of constant cap, an American capped call option without an automatic exercise clause when the cap is reached is indistinguishable from an American capped call option with an automatic exercise provision at the cap but otherwise identical features. It is indeed easy to show that the optimal exercise time for such an option is the minimum of the hitting time of the cap and the optimal exercise time for an uncapped call option with identical features (see Broadie and Detemple (1995) for a derivation of this result in a market with constant coefficients).

$$H_v^* = \frac{S_t H_v}{S_v}$$

for $v \in [t, T]$.

With these transformations it is apparent that the following result holds.

Theorem 9 (Random maturity options PCS) *Let τ_l denote a stopping time of the filtration and let $H = \{H_t : t \in [0, T]\}$ be a progressively measurable process. Consider an American random maturity put option with maturity date T , strike K , automatic liquidation time τ_l and liquidation payoff H , written on an asset with price S given by (13) in the market with stochastic interest rate r . Denote the put price by $P(S, K, T, \tau_l, H, r, \delta; \mathcal{F}_t)$ and the optimal exercise time by $\tau^p(S, K, T, \tau_l, H, r, \delta)$. Then, prior to exercise, the put price equals*

$$P(S_t, K, T, \tau_l, H, r, \delta; \mathcal{F}_t) = C(S_t^*, S, T, \tau_l^*, H^*, \delta, r; \mathcal{F}_t) \quad (31)$$

where $C(S_t^*, S, T, \tau_l^*, H^*, \delta, r; \mathcal{F}_t)$ is the value of an American random maturity call with strike price $S = S_t$, maturity date T , automatic liquidation time τ_l^* and liquidation payoff H^* in a financial market with interest rate δ and in which the underlying asset price follows the Itô process (14) with initial value $S_t^* = K$ and with z^* defined by (16). The liquidation payoff is given by

$$H_t^* = \frac{S H_t}{S_t} = \frac{S_t^* H_t}{K}$$

and the liquidation time is $\tau_l^* = \tau_l$. The optimal exercise time for the put option is

$$\tau^p(S, K, \tau_l, H, r, \delta) = \tau^c(K, S, \tau_l^*, H^*, \delta, r) \quad (32)$$

where $\tau^c(K, S, \tau_l^*, H^*, \delta, r)$ denotes the optimal exercise time for the random maturity call option.

Remark 10 Suppose that the automatic liquidation provision of the random maturity put is defined as

$$\tau_l = \inf\{t \in [0, T] : S_t \in A\}$$

where A is a closed set in $\mathbb{R}^+$, i.e. τ_l is the hitting time of the set A . Then the liquidation time of the corresponding random maturity call can be expressed in terms of the underlying asset price in the auxiliary market as

$$\tau_l^* = \inf\{t \in [0, T] : S_t^* \in A^*\}$$

where $A^* = \{x \in \mathbb{R}^+ : x = KS/y \text{ and } y \in A\}$. Given the definition of the process for S^* and the fact that the information filtration is the same in the auxiliary market it is immediate to verify that $\tau_l^* = \tau_l$.

As an immediate corollary of Theorem 9 we get the symmetry property for down and out put options and up and out call options. This generalizes results of Gao, Huang and Subrahmanyam (2000) who consider barrier options when the underlying asset price follows a geometric Brownian motion process.

Corollary 11 (Barrier options PCS) *Let $\tau_L = \inf\{t \in [0, T] : S_t = L\}$. Consider an American down and out put option with maturity date T , strike price K and automatic liquidation time τ_L (and liquidation payoff $H = 0$), written on an asset with price S given by (13) in the market with stochastic interest rate r . Prior to exercise or liquidation, the put price equals*

$$P(S_t, K, T, \tau_L, 0, r, \delta; \mathcal{F}_t) = C(S_t^*, S, T, \tau_{L^*}, 0, \delta, r; \mathcal{F}_t) \quad (33)$$

where $C(S_t^*, S, T, \tau_{L^*}, 0, \delta, r; \mathcal{F}_t)$ is the value of an American up and out call with strike price $S = S_t$, maturity date T and automatic liquidation time τ_{L^*} (and liquidation payoff $H^* = 0$) in a financial market with interest rate δ and in which the underlying asset price follows the Itô process (14) with initial value $S_t^* = K$ and with z^* defined by (16). The liquidation time is

$$\tau_{L^*} = \inf\left\{t \in [0, T] : S_t^* = L^* \equiv \frac{KS}{L}\right\}.$$

The optimal exercise time for the put option is

$$\tau^p(S, K, \tau_L, 0, r, \delta) = \tau^c(K, S, \tau_{L^*}, 0, \delta, r) \quad (34)$$

where $\tau^c(K, S, \tau_{L^*}, 0, \delta, r)$ denotes the optimal exercise time for the up and out call option.

Another corollary covers the case of American capped put and call options.

Corollary 12 (Capped options PCS) *Let $\tau_L = \inf\{t \in [0, T] : S_t = L\}$. Consider an American capped put option with maturity date T , strike price K , cap $L < K$ and automatic liquidation time τ_L (and liquidation payoff $H = K - L$), written on an asset with price S given by (13) in the market with stochastic interest rate r . Prior to exercise, the put price equals*

$$P(S_t, K, T, \tau_L, 0, r, \delta; \mathcal{F}_t) = C(S_t^*, S, T, \tau_{L^*}, 0, \delta, r; \mathcal{F}_t) \quad (35)$$

where $C(S_t^*, S, T, \tau_{L^*}, 0, \delta, r; \mathcal{F}_t)$ is the value of an American capped call with strike price $S = S_t$, maturity date T , cap $L^* = KS/L$ and automatic liquidation time τ_{L^*} (and liquidation payoff $H^* = L^* - S$) in a financial market with interest rate δ and in which the underlying asset price follows the Itô process (14) with

initial value $S_t^* = K$ and with z^* defined by (16). The liquidation time is

$$\tau_{L^*} = \inf \left\{ t \in [0, T] : S_t^* = L^* \equiv \frac{KS}{L} \right\}.$$

The optimal exercise time for the capped put option is

$$\tau^p(S, K, \tau_L, 0, r, \delta) = \tau^c(K, S, \tau_{L^*}, 0, \delta, r) \quad (36)$$

where $\tau^c(K, S, \tau_{L^*}, 0, \delta, r)$ denotes the optimal exercise time for the capped call option.

5 Multiasset derivatives

In this section we consider American-style derivatives whose payoffs depend on the values of n underlying asset prices.

The setting is as follows. The underlying filtration is generated by an n -dimensional Brownian motion process z . The price S^j of asset j follows the Itô process

$$dS_t^j = S_t^j[(r_t - \delta_t^j)dt + \sigma_t^j d\tilde{z}_t] \quad (37)$$

where r , δ^j and σ^j are progressively measurable and bounded processes, $j = 1, \dots, n$. The financial market is complete, i.e. the volatility matrix σ of the vector of prices is invertible. Let $S = (S^1, \dots, S^n)$ denote the vector of prices.

The derivatives under consideration have payoff function $f(S, K)$ with parameter K . In some applications the parameter K can be interpreted as a strike price; in others it represents a cap. We assume that the function f is continuous and homogeneous of degree one in the $n + 1$ -dimensional vector (S, K) . Examples of such contracts are call and put options on the maximum or the minimum of n assets, spread options, exchange options, capped exchange options and options on a weighted average of assets. Capped multiasset options such as capped options on the maximum or minimum of multiple assets are also obtained if K is a vector.

For a constant λ define $\lambda \circ_j S$ as

$$\lambda \circ_j S = (S^1, \dots, S^{j-1}, \lambda S^j, S^{j+1}, \dots, S^n)$$

i.e. $\lambda \circ_j S$ represents the vector of prices whose j th component has been rescaled by the factor λ . Also for a given f -claim with parameter K and for any j we define the associated f^j -claim obtained by permutation of the j th argument and the parameter

$$f^j(S, K) = f(\lambda^j \circ_j S, S^j)$$

with $\lambda^j = K/S^j$, $j = 1, \dots, n$.

For the contracts under consideration the approach of the previous sections applies and leads to the following symmetry results.

Theorem 13 *Consider an American f -claim with maturity date T and a continuous and homogeneous of degree one payoff function $f(S, K)$. Let $V(S, K, r, \delta; \mathcal{F}_t)$ denote the value of the claim in the financial market with filtration $\mathcal{F}_{(\cdot)}$, asset prices S_t satisfying (37) and progressively measurable interest rate r . Pick some arbitrary index j and define*

$$\lambda^j \equiv \frac{K}{S^j} \quad \text{and} \quad \lambda^j(\delta) \equiv \frac{r}{\delta^j}.$$

Prior to exercise the value of the claim is

$$V(S_t, K, r, \delta; \mathcal{F}_t) = V^j(S_t^*, S^j, \delta^j, \lambda^j(\delta) \circ_j \delta; \mathcal{F}_t)$$

where $V^j(S_t^*, S^j, \delta^j, \lambda^j(\delta) \circ_j \delta; \mathcal{F}_t)$ is the value of the f^j -claim with parameter S^j and maturity date T in an auxiliary financial market with interest rate δ^j and in which the underlying asset prices follow the Itô processes

$$\begin{cases} dS_v^{i*} = S_v^{i*}[(\delta_v^j - \delta_v^i)dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}]; & \text{for } i \neq j \text{ and } v \in [t, T] \\ dS_v^{j*} = S_v^{j*}[(\delta_v^j - r_v)dv + \sigma_v^j dz_v^{j*}]; & \text{for } i = j \text{ and } v \in [t, T] \end{cases}$$

with respective initial conditions $S_t^{i*} = S^i$ for $i \neq j$ and $S_t^{j*} = K$ for $i = j$. The process z^{j*} is defined by

$$dz_v^{j*} = -\tilde{d}z_v + \sigma_v^{j'} dv, \text{ for } v \in [0, T]; z_0^{j*} = 0.$$

The optimal exercise time for the f -claim is the same as the optimal exercise time for the f^j -claim in the auxiliary financial market.

Theorem 13 is a natural generalization of the one asset case. It establishes a symmetry property between a claim with homogeneous of degree one payoff in the original financial market and related claims whose payoffs are obtained by permutation of the original one in auxiliary financial markets $j = 1, \dots, n$. In the j th auxiliary market the interest rate is the dividend rate of asset j in the original economy, the dividend rate of asset i is δ^i for $i \neq j$ and r for asset j , and the volatility coefficients of asset prices are $\sigma^j - \sigma^i$ for $i \neq j$ and σ^j for asset j . The initial (date t) value of asset j is the payoff parameter K of the f -claim under consideration. Clearly the results of the previous sections are recovered when we specialize the payoff function to the earlier cases considered.

Proof of Theorem 13 Define $S^j = S_t^j$. Proceeding as in Section 2 we can write the

value of the contract

$$\begin{aligned}
V(S_t, K, r, \delta; \mathcal{F}_t) &= \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) f(S_\tau, K) \mid \mathcal{F}_t \right] \\
&= \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) \frac{S_\tau^j}{S_t^j} f \left(S_\tau \frac{S^j}{S_\tau^j}, K \frac{S^j}{S_\tau^j} \right) \mid \mathcal{F}_t \right] \\
&= \sup_{\tau \in \mathcal{S}_{t,T}} E^{j*} \left[\exp \left(- \int_t^\tau \delta_v^j dv \right) f \left(S_\tau \frac{S^j}{S_\tau^j}, S_\tau^{j*} \right) \mid \mathcal{F}_t \right] \\
&= \sup_{\tau \in \mathcal{S}_{t,T}} E^{j*} \left[\exp \left(- \int_t^\tau \delta_v^j dv \right) f^j(S_\tau^*, S^j) \mid \mathcal{F}_t \right] \\
&= V^j(S_t^*, S^j, \delta^j, \lambda^j(\delta) \circ_j \delta; \mathcal{F}_t).
\end{aligned}$$

The second equality above uses the homogeneity property of the payoff function, the third is based on the definition $S_\tau^{j*} = K S^j / S_\tau^j$ and the passage to the measure Q^{j*} and the fourth relies on the definition of the permuted payoff f^j . The final equality uses the definition of the value function V^j .

To complete the proof of the theorem it suffices to use Itô's lemma to identify the dynamics of the asset prices in the auxiliary economy. This leads to the processes stated in the theorem. $\square$

The interest of the theorem becomes apparent when we specialize the payoff function to familiar ones. The following results are valid.

1. Call max-option on two assets ($f(S^1, S^2, K) = (\max(S^1, S^2) - K)^+$): One symmetric contract is an option to exchange the maximum of an asset and cash against another asset (or, equivalently, an exchange option with put floor) whose payoff is

$$f^2(S^{1*}, S^{2*}, K') = (\max(S^{1*}, K') - S^{2*})^+ = (S^{1*} - S^{2*})^+ \vee (K' - S^{2*})^+$$

where $K' = S^2$ in the auxiliary financial market obtained by taking $j = 2$ as reference. A similar contract emerges if $j = 1$ is taken as reference. The theorem implies that the valuation of any one of these contracts is obtained by a simple reparametrization of the values of the symmetric contracts.

2. Exchange option on two assets ($f(S^1, S^2) = (S^1 - S^2)^+$): A symmetric contract is a standard call option with payoff

$$f^2(S^{1*}, K') = (S^{1*} - K')^+$$

and $K' = S^2$ in the auxiliary market $j = 2$ in which S^{1*} satisfies

$$\begin{aligned}
dS_t^{1*} &= S_t^{1*}[(\delta_t^2 - \delta_t^1)dt + (\sigma_t^2 - \sigma_t^1)dz_t^{2*}] \\
&= S_t^{1*}[(\delta_t^2 - \delta_t^1)dt + (\sigma_{1t}^2 - \sigma_{1t}^1)dz_{1t}^{2*} + (\sigma_{2t}^2 - \sigma_{2t}^1)dz_{2t}^{2*}].
\end{aligned}$$

In the second equality we used $\sigma^i = (\sigma_1^i, \sigma_2^i)$, for $i = 1, 2$. Bjerk Sund and Stensland (1993) prove this result for financial markets with constant coefficients using PDE methods (see also Rubinstein (1991) for a proof in a binomial setting and Broadie and Detemple (1997) for a proof based on the EEP representation). The case of European options is treated in Margrabe (1978). Our theorem establishes the validity of this symmetry in a much broader setting. The second symmetric contract is a standard put option with strike price $K' = S^1$ in the auxiliary market $j = 1$.

3. Capped exchange option with proportional cap ($f(S^1, S^2) = LS^2 \wedge (S^1 - S^2)^+$): In this instance one symmetric contract (in the auxiliary financial market $j = 2$) is a capped call option with constant cap whose payoff is

$$f^2(S^{1*}, K') = LK' \wedge (S^{1*} - K')^+$$

where $K' = S^2$. The theorem thus provides a simple and immediate proof of this result derived in Broadie and Detemple (1997) for models with constant coefficients. Alternatively we can also consider the symmetric contract in the auxiliary market $j = 1$. We find the payoff

$$f^1(K', S^{2*}) = LS^{2*} \wedge (K' - S^{2*})^+,$$

with $K' = S^1$. In other words the capped exchange option with proportional cap is symmetric to a put option with proportional cap in the market in which asset 1 is chosen as the numeraire.

4. Capped exchange option with constant cap ($f(S^1, S^2, K) = (S^1 \wedge K - S^2)^+$): The symmetric contract in any auxiliary market $j = 2$ is a call option on the minimum of two assets with payoff

$$f^2(S^{1*}, S^{2*}, K') = (S^{1*} \wedge S^{2*} - K')^+$$

where $K' = S^2$. An analysis of min-options in the context of the model with constant coefficients is carried out in Detemple, Feng and Tian (2000).

5. The symmetry relations of Theorem 13 also apply to multiasset derivatives whose payoffs are homogeneous of degree one relative to a subset of variables. An interesting example is provided by quantos. These are derivatives written on foreign asset prices or indices but whose payoff is denominated in domestic currency. For instance a quanto call option on the Nikkei pays off $(S - K)^+$ dollars at the exercise time where S is the value of the Nikkei quoted in yen. The payoff in foreign currency is $e(S - K)^+$ where e is the Y/\$ exchange rate. From the foreign perspective the contract is homogeneous of degree $\nu = 2$ in the triplet (e, S, K) . However, for interpretation purposes it is more advantageous to treat it as a contract homogeneous of degree $\nu = 1$ in the exchange rate e . If

r_f denotes the foreign interest rate and the dividend rate on the index is δ the American quanto call is valued at

$$C_t^Q = \sup_{\tau \in S_{t,T}} \tilde{E}^f \left[\exp \left(- \int_t^\tau r_v^f dv \right) e_\tau (S_\tau - K)^+ | \mathcal{F}_t \right]$$

in yen where the expectation is taken relative to the foreign risk neutral measure and

$$\begin{cases} dS_t = S_t[(r_t^f - \delta_t)dt + \sigma_t d\tilde{z}_t^f] \\ de_t = e_t[(r_t^f - r_t)dt + \sigma_t^e d\tilde{z}_t^f]. \end{cases}$$

Here r is the domestic interest rate and σ, σ^e are the volatility coefficients of the foreign index and the exchange rate. The process $\tilde{z}^f$ is a two-dimensional Brownian motion relative to the foreign risk neutral measure. Using the exchange rate as new numeraire yields

$$C_t^Q = \sup_{\tau \in S_{t,T}} E^{f*} \left[\exp \left(- \int_t^\tau r_v dv \right) (S_\tau - K)^+ | \mathcal{F}_t \right]$$

where

$$dS_t = S_t \left[(r_t^f - \delta_t + \sigma_t \sigma_t^{e'}) dt - \sigma_t dz_t^{f*} \right].$$

Hence, from the foreign perspective the quanto call option is symmetric to a standard call option on an asset paying dividends at the rate $\delta - \sigma \sigma^{e'}$ in an auxiliary financial market with interest rate r . Similarly a quanto forward contract is symmetric to a standard forward contract in the same auxiliary financial market. The forward price is

$$F_t = \frac{E^{j*} [\exp(-\int_t^\tau r_v dv) S_\tau | \mathcal{F}_t]}{E^{j*} [\exp(-\int_t^\tau r_v dv) | \mathcal{F}_t]}.$$

For the case of constant coefficients $F_t = S_t \exp((r^f - \delta + \sigma \sigma^{e'})(T - t))$. Alternative representations for these prices can be derived by using the homogeneity of degree 2 relative to (e, S, K) ; they are discussed in Section 7.

6. Lookback options: The exercise payoff depends on an underlying asset value and its sample path maximum or minimum. A lookback put pays off $f(S_v, M_v) = (M_v - S_v)^+$ where $M_v = \sup_{s \in [0, v]} S_s$; the lookback call payoff is $f(S_v, m_v) = (S_v - m_v)^+$ where $m_v = \inf_{s \in [0, v]} S_s$. Even though there is only one underlying asset the contract depends on two state variables, namely the underlying asset price and one of its sample path statistics. Since renormalizations do not affect the order of a sample path statistic it is easily verified that the lookback call is symmetric to a put option on the minimum of the price expressed in a new numeraire $(S - m_v^*)$ where $m_v^* = (S/S_v) \inf_{s \in [0, v]} S_s = \inf_{s \in [0, v]} (SS_s/S_v)$.

Likewise, a lookback put is related to a call option on the maximum of the price expressed in a new numeraire. European lookback option pricing is discussed in Goldman, Sosin and Gatto (1979) and Garman (1989) in the context of the model with constant coefficients. Similar symmetry relations can be established for average options (Asian options).

6 Occupation time derivatives

An occupation time derivative is a derivative whose payoff has been modified to reflect the time spent by the underlying asset price in certain regions of space. Various special cases have been considered in the recent literature such as Parisian and cumulative barrier options (Chesney, Jeanblanc-Picqué and Yor (1997)), step options (Linetsky (1999)) and quantile options (Miura (1992), Akahori (1995), Dassios (1995)). The general class of occupation time claims is introduced by Hugonnier (1998) who discusses their valuation and hedging properties. So far the literature has focused exclusively on European-style derivatives when the underlying asset follows a geometric Brownian motion process. In this section we provide symmetry results applying to both European and American-style contracts and when the underlying asset follows an Itô process. Extensions to multiasset occupation time derivatives are also discussed.

We consider an American occupation time f -claim with exercise payoff

$$f(S, K, O^{S,A})$$

at time t , where S satisfies the Itô process (1), K is a constant representing a strike price or a cap and $O^{S,A}$ is an occupation time process defined by

$$O_t^{S,A} = \int_0^t 1_{\{S_v \in A_v\}} dv, \quad t \in [0, T].$$

for some random, progressively measurable, closed set $A(\cdot, \cdot) : [0, T] \times \Omega \rightarrow \mathcal{B}(\mathbb{R}^+)$. Thus $O_t^{S,A}$ represents the amount of time spent by S in the set A during the time interval $[0, t]$. Examples treated in the literature involve occupation times of constant sets of the form $A = \{x \in \mathbb{R}^+ : x \geq L\}$ or $A = \{x \in \mathbb{R}^+ : x \leq L\}$ with L constant, which represent time spent above or below a constant barrier L . Simple generalizations of these are when the barrier L is a function of time or a progressively measurable stochastic process.

The value of this American claim is

$$V(S_t, K, O^{S,A}, r, \delta; \mathcal{F}_t) = \sup_{\tau \in S_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) f(S_\tau, K, O_\tau^{S,A}) \mid \mathcal{F}_t \right].$$

Assume that the claim is homogeneous of degree one in (S, K) . Then we can perform the usual change of measure and obtain

Theorem 14 Consider an American occupation time f -claim with maturity date T and a payoff function $f(S, K, O)$ which is homogeneous of degree one with respect to (S, K) . Let $V(S, K, O^{S,A}, r, \delta; \mathcal{F}_t)$ denote the value of the claim in the financial market with filtration $\mathcal{F}_t$, asset price S satisfying (1) and progressively measurable interest rate r . Prior to exercise the value of the claim is

$$V(S_t, K, O^{S,A}, r, \delta; \mathcal{F}_t) = V^1(S_t^*, S, O^{S^*,A^*}, \delta, r; \mathcal{F}_t)$$

where $A^* = \{A^*(v, \omega), v \in [t, T]\}$ with $A^*(v, \omega) = \{x \in \mathbb{R}^+ : x = \frac{KS}{y} \text{ and } y \in A(v, \omega)\}$ and $O_t^{S^*,A^*} \equiv O_t^{S,A}$. Also $V^1(S_t^*, S, O^{S^*,A^*}, \delta, r; \mathcal{F}_t)$ is the value of the permuted claim $f^1(S_t^*, S, O_t^{S^*,A^*}) = f(S, K \frac{S}{S_t^*}, O_t^{S^*,A^*})$ with parameter $S = S_t$, occupation time $O_t^{S^*,A^*}$, and maturity date T in an auxiliary financial market with interest rate δ and in which the underlying asset price follows the Itô process

$$dz_v^* = S_v^*[(\delta_v - r_v)dt + \sigma_v dz_v^*], \quad \text{for } v \in [t, T]$$

with initial condition $S_t^* = K$. The process z^* is defined by $dz_v^* = -d\tilde{z}_v + \sigma_v dv$, $v \in [0, T]$, $z_0^* = 0$. The optimal exercise time for the f -claim is the same as the optimal exercise time for the f^1 -claim in the auxiliary financial market.

Proof of Theorem 14 Fix $t \in [0, T]$ and set $O_t^{S,A} = O_t^{S^*,A^*}$. For any stopping time $\tau \in \mathcal{S}_{t,T}$ the occupation time can be written

$$O_\tau^{S,A} = O_t^{S,A} + \int_t^\tau 1_{\{S_v \in A_v\}} dv = O_t^{S^*,A^*} + \int_t^\tau 1_{\{S_v^* \in A_v^*\}} dv = O_\tau^{S^*,A^*}$$

where $S_v^* = KS/S_v$, $v \in [t, T]$ and $O_\tau^{S^*,A^*}$ denotes the occupation time of the random set A^* by the process S^* . Performing the change of measure leads to the results. $\square$

Special cases of interest are as follows.

1. Parisian options (Chesney, Jeanblanc-Picqué and Yor (1997)): Let $g(L, t) = \sup\{s \leq t : S_s = L\}$ denote the last time the process S has reached the barrier L (if no such time exists set $g(L, t) = t$) and consider the random time

$$O_t^{S,A^+(t,L)} = \int_{g(L,t)}^t 1_{\{S_v \geq L\}} dv = \int_0^t 1_{\{(v, S_v) \in A^+(t,L)\}} dv$$

where $A^+(t, L) = \{(v, S) : v \geq g(L, t), S \geq L\}$. Note that $O_t^{S,A^+(t,L)}$ measures the age of a current excursion above the level L . A Parisian up and out call with window D has null payoff as soon as an excursion of age D above L takes place. If no such event occurs prior to exercise the exercise payoff is $(S - K)^+$. A Parisian down and out call with window D loses all value if there is an excursion of length D below the prespecified level L . Parisian put options

are similarly defined. Fix $t \in [0, T]$ and suppose that no excursion of age D has occurred before t . The symmetry relation for Parisian options can be stated as

$$C(S_t, K, O_t^{S, A^+(t, L)}, D, r, \delta; \mathcal{F}_t) = P(S_t^*, S, O_t^{S^*, A^-(t, KS/L)}, D, \delta, r; \mathcal{F}_t). \quad (38)$$

This follows from $g(L, t) = \sup\{s \leq t : S_s = L\} = \sup\{s \leq t : S_s^* = KS/L\} = g^*(KS/L, t)$ and

$$O_t^{S, A^+(t, L)} = \int_{g(L, t)}^t 1_{\{KS/L \geq KS/S_v\}} dv = \int_{g^*(KS/L, t)}^t 1_{\{KS/L \geq S_v^*\}} dv = O_t^{S^*, A^-(t, KS/L)},$$

with $A^-(t, KS/L) = \{(v, S^*) : v \geq g^*(KS/L, t), KS/L \geq S^*\}$, which ensures that the stopping times

$$H_t(L, D) = \inf\{v \in [t, T] : O_v^{S, A^+(v, L)} \geq D\}, \quad \text{and} \\ H_t^*(KS/L, D) = \inf\{v \in [t, T] : O_v^{S^*, A^-(v, KS/L)} \geq D\}$$

at which the call and put options lose all value coincide. In summary a Parisian up and out call with window D has the same value as a Parisian down and out put with window D , strike $S = S_t$, occupation time $O_t^{S^*, A^-(t, KS/L)}$, and maturity date T in an auxiliary financial market with interest rate δ and in which the underlying asset price follows the Itô process described in Theorem 14. Chesney, Jeanblanc-Picqué and Yor derive this symmetry property for European Parisian options in a financial market with constant coefficients. In this context they also provide valuation formulas for such contracts involving Laplace transforms.

2. Cumulative (Parisian) barrier options (Chesney, Jeanblanc-Picqué and Yor (1997)): The contract payoff is affected by the (cumulative) amount of time spent above or below a constant barrier L . For instance let $A^\pm(L) = \{x \in \mathbb{R}^+ : (x - L)^\pm \geq 0\}$ and consider a call option that pays off if the amount of time spent above L exceeds some prespecified level D (up and in call). The following symmetry result applies:

$$C(S_t, K, O_t^{S, A^+(L)}, D, r, \delta; \mathcal{F}_t) = P(S_t^*, S, O_t^{S^*, A^-(KS/L)}, D, \delta, r; \mathcal{F}_t). \quad (39)$$

Here the left hand side is the value of the cumulative barrier call with payoff $(S - K)^+ 1_{\{O^{S, A^+(L)} \geq D\}}$ in the original economy; the right hand side is the value of a cumulative barrier put option with payoff $(S - S^*)^+ 1_{\{O^{S^*, A^-(KS/L)} \geq D\}}$ in an auxiliary economy with interest rate δ , dividend r and asset price process S^* . Chesney, Jeanblanc-Picqué and Yor (1997) and Hugonnier (1998) examine the valuation of European cumulative barrier options when the underlying asset price follows a geometric Brownian motion process. European cumulative barrier digital calls and puts satisfy similar symmetry relations and are discussed

by Hugonnier. An analysis of these contracts is relegated to the next section since their payoffs are homogeneous of degree zero.

3. Step options (Linetsky (1999)): A step option is discounted at a rate which depends on the occupation time of a set. For instance the step call option payoff is $(S - K)^+ \exp(-\rho O_t^{S, A^\pm(L)})$ for some $\rho > 0$ where $A^\pm(L)$ is defined above. Again the PCS relation (39) holds in this case. Put and call step options are special cases of the occupation time derivatives in which the payoff function involves exponential discounting. Closed form solutions are provided by Linetsky for geometric Brownian motion price process.

Occupation time derivatives can be easily generalized to the multiasset case. For a progressively measurable stochastic closed set $A \in \mathbb{R}_+^n$ and a vector of asset prices $S \in \mathcal{B}(\mathbb{R}_+^n)$ a multiasset f -claim has payoff $f(S, K, O^{S, A})$ where

$$O_t^{S, A} = \int_0^t 1_{\{S_v \in A_v\}} dv, \quad t \in [0, T].$$

A natural generalization of Theorem 13 is

Theorem 15 *Consider an American occupation time f -claim with maturity date T and a payoff function $f(S, K, O^{S, A})$ which is homogeneous of degree one in (S, K) . Let $V(S, K, O^{S, A}, r, \delta; \mathcal{F}_t)$ denote the value of the claim in the financial market with filtration $\mathcal{F}_{(\cdot)}$, asset prices S satisfying (37) and progressively measurable interest rate r . Pick some arbitrary index j and define*

$$\lambda^j \equiv \frac{K}{S^j} \quad \text{and} \quad \lambda^j(\delta) \equiv \frac{r}{\delta^j}.$$

Prior to exercise the value of the multiasset occupation time f -claim is

$$V(S_t, K, O^{S, A}, r, \delta; \mathcal{F}_t) = V^j(S_t^*, S^j, O^{S^*, A^*}, \delta^j, \lambda^j(\delta) \circ_j \delta; \mathcal{F}_t)$$

where $A^* = \{A^*(v, \omega), v \in [t, T]\}$ with $A^*(v, \omega) = \{x \in \mathbb{R}_+^n : x_i = y_i S / y_j, \text{ for } i \neq j, x_j = K S / y_j \text{ and } y = (y_1, \dots, y_n) \in A(v, \omega)\}$ and $O_t^{S^*, A^*} \equiv O_t^{S, A}$. Also $V^j(S_t^*, S^j, O_t^{S^*, A^*}, \delta^j, \lambda^j(\delta) \circ_j \delta; \mathcal{F}_t)$ is the value of the f^j -claim with parameter $S^j = S_t^j$, maturity date T and occupation time $O_t^{S^*, A^*}$ in an auxiliary financial market with interest rate δ^j and in which the underlying asset prices follow the Itô processes

$$\begin{cases} dS_v^{i*} = S_v^{i*}[(\delta_v^j - \delta_v^i)dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}]; & \text{for } i \neq j \text{ and } v \geq t \\ dS_v^{j*} = S_v^{j*}[(\delta_v^j - r_v)dv + \sigma_v^j dz_v^{j*}]; & \text{for } i = j \text{ and } v \geq t \end{cases}$$

with respective initial conditions S^i for $j \neq i$ and K for $j = i$. The process z^{j*} is defined by

$$dz_v^{j*} = -d\tilde{z}_v + \sigma_v^{j'} dv$$

for all $v \in [0, T]$, $z_0^{j*} = 0$. The optimal exercise time for the f -claim is the same as the optimal exercise time for the f^j -claim in the auxiliary financial market.

Some particular cases are the natural counterpart of standard multiasset options.

1. Cumulative barrier max- and min-options: When there are two underlying assets call options in this category have payoff functions of the form $(S_t^1 \vee S_t^2 - K)^+ 1_{\{O_t^{S,A} \geq b\}}$ (max-option) or $(S_t^1 \wedge S_t^2 - K)^+ 1_{\{O_t^{S,A} \geq b\}}$ (min-option), where $b \in [0, T]$. Similarly for put options. It is easily verified that a cumulative barrier call max-option is symmetric to a cumulative barrier option to exchange the maximum of an asset and cash against another asset for which the occupation time has been adjusted.
2. Cumulative barrier exchange options: The payoff function takes the form $(S^1 - S^2) 1_{\{O_t^{S,A} \geq b\}}$. This exchange option is symmetric to cumulative barrier call and put options with suitably adjusted occupation times.
3. Quantile options (Miura (1992), Akahori (1995), Dassios (1995)): An α -quantile call option pays off $(M(\alpha, t) - K)$ upon exercise where $M(\alpha, t) = \inf\{x : \int_0^t 1_{\{S_v \leq x\}} dv > \alpha t\} = \inf\{x : O_t^{S,A^-(x)} > \alpha t\}$. Consider an α -quantile strike put with payoff $(M(\alpha, t) - S_t)$. Note that

$$\begin{aligned} M(\alpha, t) &= \inf\left\{x : \int_0^t 1_{\{S_v \leq x\}} dv > \alpha t\right\} = \inf\left\{x : \int_0^t 1_{\{SS_v/S_t \leq Sx/S_t\}} dv > \alpha t\right\} \\ &= (S_t/S) \inf\left\{y : \int_0^t 1_{\{SS_v/S_t \leq y\}} dv > \alpha t\right\} \equiv (S_t/S) M^*(\alpha, t) \end{aligned}$$

where $M^*(\alpha, t)$ is the α -quantile of the normalized price $S_{v,t}^* \equiv SS_v/S_t$ for $v \leq t$. Thus $M(\alpha, t) = (S_t/S) M^*(\alpha, t)$ and an α -quantile strike put is seen to be symmetric to an α -quantile call option with (fixed) strike price S and quantile based on the normalized asset price $S_{v,t}^*$, $v \leq t$.

Multiasset step options can be also be defined in a natural manner and satisfy symmetry properties akin to those of standard multiasset options.

7 Symmetry property without homogeneity of degree one

Several derivative securities have payoffs that are not homogeneous of degree one. Examples include digital options and quantile options (homogeneous of degree $v = 0$) or product options (homogeneous of degree $v \neq 0, 1$). Product options (options on a product of assets) include options on foreign indices with payoff in domestic currency such as quanto options. As we show below, even in these cases, symmetry-like properties link various types of contracts.

Consider an f -claim on n underlying assets whose payoff is homogeneous of degree ν , i.e.,

$$f(\lambda S, \lambda K) = \lambda^\nu f(S, K)$$

for some $\nu \geq 0$ and for all $\lambda > 0$. The following result is then valid.

Theorem 16 *Consider an American f -claim with maturity date T and a continuous and homogeneous of degree ν payoff function $f(S, K)$. Let $V(S, K, r, \delta; \mathcal{F}_t)$ denote the value of the claim in the financial market with filtration $\mathcal{F}_{(\cdot)}$, asset prices S_t satisfying (37) and progressively measurable interest rate r . For $j = 1, \dots, n$, define*

$$\begin{aligned} r^{j*} &= (1 - \nu)r + \nu\delta^j + \frac{1}{2}\nu(1 - \nu)\sigma^j\sigma^{j'} \\ \delta^{i*} &= (1 - \nu)r + \delta^i + (\nu - 1)\delta^j + (1 - \nu)\left(-1 + \frac{1}{2}\nu\right)\sigma^j\sigma^{j'} + (1 - \nu)\sigma^i\sigma^{j'}, \\ &\quad \text{for } i \neq j \\ \delta^{j*} &= (2 - \nu)r + (\nu - 1)\delta^j + (1 - \nu)\left(-1 + \frac{1}{2}\nu\right)\sigma^j\sigma^{j'}. \end{aligned}$$

Prior to exercise the value of the claim is, for any $j = 1, \dots, n$,

$$V(S_t, K, r, \delta; \mathcal{F}_t) = V^j(S_t^*, S^j, r^{j*}, \delta^{j*}; \mathcal{F}_t)$$

where $V^j(S_t^*, S^j, r^{j*}, \delta^{j*}; \mathcal{F}_t)$ is the value of the f^j -claim with parameter S^j and maturity date T in an auxiliary financial market with interest rate r^{j*} and in which the underlying asset prices follow the Itô processes

$$\begin{cases} dS_v^{i*} = S_v^{i*}[(r_v^{j*} - \delta_v^{i*})dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}]; & \text{for } i \neq j \text{ and } v \in [t, T] \\ dS_v^{j*} = S_v^{j*}[(r_v^{j*} - \delta_v^{j*})dv + \sigma_v^j dz_v^{j*}]; & \text{for } i = j \text{ and } v \in [t, T] \end{cases}$$

with respective initial conditions $S_t^{i*} = S^i$ for $i \neq j$ and $S_t^{j*} = K$ for $i = j$. The process z^{j*} is defined by

$$dz_v^{j*} = -d\tilde{z}_v + \nu\sigma_v^{j'}dv, \text{ for } v \in [0, T]; z_0^{j*} = 0.$$

The optimal exercise time for the f -claim is the same as the optimal exercise time for the f^j -claim in the auxiliary financial market.

Proof of Theorem 16 Define $S^j = S_t^j$. Let

$$r_v^{j*} = (1 - \nu)r_v + \nu\delta_v^j + \frac{1}{2}\nu(1 - \nu)\sigma_v^j\sigma_v^{j'}$$

and note that

$$\begin{aligned} \exp\left(-\int_t^\tau r_v dv\right) \left(\frac{S_\tau^j}{S_t^j}\right)^\nu \\ = \exp\left(-\int_t^T r_v^{j*} dv\right) \exp\left(-\frac{1}{2}\nu^2 \int_t^T \sigma_v^j \sigma_v^{j'} dv + \nu \int_t^T \sigma_v^j d\tilde{z}_v\right). \end{aligned}$$

Defining the equivalent measure

$$dQ^{j*} = \exp\left(-\frac{1}{2}\nu^2 \int_0^T \sigma_v^j \sigma_v^{j'} dv + \nu \int_0^T \sigma_v^j d\tilde{z}_v\right) dQ$$

enables us to write

$$\begin{aligned} V(S_t, K, r, \delta; \mathcal{F}_t) &= \sup_{\tau \in S_{t,T}} \tilde{E} \left[\exp\left(-\int_t^\tau r_v dv\right) f(S_\tau, K) | \mathcal{F}_t \right] \\ &= \sup_{\tau \in S_{t,T}} \tilde{E} \left[\exp\left(-\int_t^\tau r_v dv\right) \left(\frac{S_\tau^j}{S_t^j}\right)^\nu f\left(S_\tau \frac{S^j}{S_\tau^j}, K \frac{S^j}{S_\tau^j}\right) | \mathcal{F}_t \right] \\ &= \sup_{\tau \in S_{t,T}} E^{j*} \left[\exp\left(-\int_t^\tau r_v^{j*} dv\right) f\left(S_\tau \frac{S^j}{S_\tau^j}, S_\tau^{j*}\right) | \mathcal{F}_t \right] \\ &= \sup_{\tau \in S_{t,T}} E^{j*} \left[\exp\left(-\int_t^\tau r_v^{j*} dv\right) f^j(S_\tau^*, S^j) | \mathcal{F}_t \right] \\ &= V^j(S_t^*, S^j, r^{*j}, \delta^*; \mathcal{F}_t). \end{aligned}$$

Under Q^{j*} the process

$$dz_v^{j*} = -d\tilde{z}_v + \nu \sigma_v^{j'} dv$$

is a Brownian motion and S^{i*} satisfies, for $i \neq j$ and $v \in [t, T]$

$$\begin{aligned} dS_v^{i*} &= S_v^{i*}[(\delta_v^j - \delta_v^i + (\sigma_v^j - \sigma_v^i)\sigma_v^{j'})dv - (\sigma_v^j - \sigma_v^i)d\tilde{z}_v] \\ &= S_v^{i*}[(\delta_v^j - \delta_v^i + (\sigma_v^j - \sigma_v^i)\sigma_v^{j'})dv + (\sigma_v^j - \sigma_v^i)[dz_v^{j*} - \nu \sigma_v^{j'} dv]] \\ &= S_v^{i*}[(\delta_v^j - \delta_v^i + (1-\nu)(\sigma_v^j - \sigma_v^i)\sigma_v^{j'})dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}] \\ &= S_v^{i*}[(r_v^{j*} - \delta_v^{i*})dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}] \end{aligned}$$

where

$$\delta_v^{i*} = (1-\nu)r_v + \delta_v^i + (\nu-1)\delta_v^j + (1-\nu)\left(-1 + \frac{1}{2}\nu\right)\sigma_v^j \sigma_v^{j'} + (1-\nu)\sigma_v^i \sigma_v^{j'}$$

and for $i = j$ and $v \in [t, T]$

$$\begin{aligned} dS_v^{j*} &= S_v^{j*}[(\delta_v^j - r_v + \sigma_v^j \sigma_v^{j'})dv - \sigma_v^j d\tilde{z}_v] \\ &= S_v^{j*}[(\delta_v^j - r_v + (1-\nu)\sigma_v^j \sigma_v^{j'})dv + \sigma_v^j dz_v^{j*}] \\ &= S_v^{j*}[(r_v^{j*} - \delta_v^{j*})dv + \sigma_v^j dz_v^{j*}] \end{aligned}$$

where

$$\delta_v^{j*} = (2 - \nu)r_v + (\nu - 1)\delta_v^j + (1 - \nu)\left(-1 + \frac{1}{2}\nu\right)\sigma_v^j\sigma_v^{j'}.$$

This completes the proof of the theorem. $\square$

Remark 17 When the claim is homogeneous of degree 1 the interest rate and the dividend rates in the economy with numeraire j become $r_v^{j*} = \delta_v^j$, $\delta_v^{i*} = \delta_v^i$, for $i \neq j$, and $\delta_v^{j*} = r_v$. Thus we recover the prior results of Theorem 13.

Another special case of interest is when the payoff function is homogeneous of degree 0. The economy with numeraire j then has characteristics

$$\begin{aligned} r^{j*} &= r \\ \delta^{i*} &= r + \delta^i - \delta^j - (\sigma^j - \sigma^i)\sigma^{j'}, \text{ for } i \neq j \\ \delta^{j*} &= 2r - \delta^j - \sigma^j\sigma^{j'} \end{aligned}$$

and the underlying asset prices follow the Itô processes

$$\begin{cases} dS_v^{i*} = S_v^{i*}[(r_v^{j*} - \delta_v^{i*})dv + (\sigma_v^j - \sigma_v^i)dz_v^{j*}]; & \text{for } i \neq j \text{ and } v \in [t, T] \\ dS_v^{j*} = S_v^{j*}[(r_v^{j*} - \delta_v^{j*})dv + \sigma_v^j dz_v^{j*}]; & \text{for } i = j \text{ and } v \in [t, T] \end{cases}$$

with respective initial conditions $S_t^{i*} = S^i$ for $i \neq j$ and $S_t^{j*} = K$ for $i = j$. The process z_v^{j*} is defined by $dz_v^{j*} = -d\tilde{z}_v$, for $v \in [0, T]$. It is a Brownian motion under $Q^* = Q$.

Examples of contracts in this category are

1. Digital options: A digital call option ($f(S, K) = 1_{\{S \geq K\}}$) is symmetric to a digital put option with strike $S = S_t$, written on an asset with dividend rate $\delta^* = 2r - \delta - \sigma^2$, in an economy with interest rate $r^* = r$.
2. Digital multiasset options: A digital call max-option ($f(S^1, S^2, K) = 1_{\{S^1 \vee S^2 \geq K\}}$) is symmetric to a digital option to exchange the maximum of an asset and cash against another asset ($f^2(S^1, S^2, K') = 1_{\{S^{*1} \vee K' \geq S^{*2}\}}$, where $K' = S^2$) in the economy with asset $j = 2$ as numeraire (with characteristics $r^{2*} = r$, $\delta^{1*} = r + \delta^1 - \delta^2 - (\sigma^2 - \sigma^1)\sigma^{2'}$, and $\delta^{2*} = 2r - \delta^2 - \sigma^2\sigma^{2'}$). A digital call min-option ($f(S^1, S^2, K) = 1_{\{S^1 \wedge S^2 \geq K\}}$) is symmetric to a digital option to exchange the minimum of an asset and cash against another asset ($f^2(S^1, S^2, K') = 1_{\{S^{*1} \wedge K' \geq S^{*2}\}}$, where $K' = S^2$) in the same auxiliary economy. Similar relations hold for digital multiasset put options.
3. Cumulative barrier digital options: Symmetry properties for occupation time derivatives with homogeneous of degree zero payoffs can be easily identified by drawing on the previous section. A cumulative barrier digital call option with barrier L (i.e. payoff $f(S, K, O^{S, A^+(L)}) = 1_{\{S \geq K\}} 1_{\{O_t^{S, A^+(L)} \geq b\}}$ where

$A^+(L) = \{x \in \mathbb{R}^+ : (x - L)^+ \geq 0\}$ is symmetric to a cumulative barrier digital put option with barrier $L^* = KS/L$ (i.e. payoff $f^1(S^*, K', O^{S^*, A^-(L^*)}) = 1_{\{K' \geq S^*\}} 1_{\{O_t^{S^*, A^-(L^*)} \geq b\}}$ where $K' = S$ and $A^-(L^*) = \{x \in \mathbb{R}^+ : (x - L^*)^- \geq 0\}$). A similar symmetry relation can be established for Parisian digital call and put options.

4. Quanto options: Consider again the quanto call option with payoff $e(S - K)^+$ in foreign currency where e is the Y/\$ exchange rate. From the foreign perspective the contract is homogeneous of degree $v = 2$ in the triplet (e, S, K) . The results of Theorem 16 imply that the quanto call is symmetric to an exchange option in an economy with interest rate

$$r^{f*} = -r^f + 2r - \sigma^e \sigma^{e'}$$

and which underlying assets have dividend rates

$$\begin{aligned} \delta^{1*} &= -r^f + \delta + r - \sigma \sigma^{e'} \\ \delta^{2*} &= r. \end{aligned}$$

The call value can be written

$$C_t^Q = e_t \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E}^{f*} \left[\exp \left(- \int_t^\tau r_v^{f*} dv \right) (S_\tau^{1*} - S_\tau^{2*})^+ | \mathcal{F}_t \right]$$

where

$$\begin{cases} dS_v^{1*} = S_v^{1*} [(r_v^{f*} - \delta_v^{1*})dv + (\sigma_v^e - \sigma_v)dz_v^{f*}]; & \text{for } v \in [t, T] \\ dS_v^{2*} = S_v^{2*} [(r_v^{f*} - \delta_v^{2*})dv + \sigma_v^e dz_v^{f*}]; & \text{for } v \in [t, T], \end{cases}$$

with the initial conditions $S_t^{1*} = S_t$ and $S_t^{2*} = K$. An alternative representation for the quanto call was provided in Section 7.

Remark 18 Representation formulas involving the change of measure introduced in earlier sections can also be obtained with payoffs that are homogeneous of degree v . In this case the coefficients of the underlying asset price processes reflect the homogeneity degree of the payoff function. Indeed letting $S^j = S_t^j$ we can always write

$$\begin{aligned} V(S_t, K, r, \delta; \mathcal{F}_t) &= \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) f(S_\tau, K) | \mathcal{F}_t \right] \\ &= \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{E} \left[\exp \left(- \int_t^\tau r_v dv \right) \left(\frac{S_\tau^j}{S^j} \right) \right. \\ &\quad \left. \times f \left(S_\tau \left(\frac{S^j}{S_\tau^j} \right)^{1/v}, K \left(\frac{S^j}{S_\tau^j} \right)^{1/v} \right) | \mathcal{F}_t \right] \end{aligned}$$

$$= \sup_{\tau \in \mathcal{S}_{t,T}} E^{j*} \left[\exp \left(- \int_t^\tau \delta_v^j dv \right) f(\widehat{S}_\tau, \widehat{S}_\tau^{n+1}) | \mathcal{F}_t \right]$$

where $\widehat{S}_v^i = S_v^i (\frac{S_v^j}{S_v^i})^{1/\nu}$ for $i = 1, \dots, n$ and $\widehat{S}_v^{n+1} = K (\frac{S_v^j}{S_v^i})^{1/\nu}$ for $v \in [t, T]$. The auxiliary economy has interest rate δ^j and the equivalent measure Q^{j*} is

$$dQ^{j*} = \exp \left(-\frac{1}{2} \int_0^T \sigma_v^j \sigma_v^{j'} dv + \int_0^T \sigma_v^j d\widetilde{z}_v \right) dQ.$$

The process $dz_v^{j*} = -d\widetilde{z}_v + \sigma_v^{j'} dv$, for $v \in [0, T]$ is a Q^{j*} -Brownian motion process.

8 Changes of numeraire and representation of prices

In the financial markets of the previous sections the price of a contingent claim is the expectation of its discounted payoff where discounting is at the riskfree rate and the expectation is taken under the risk neutral measure. This standard representation formula is implied by the ability to replicate the claim's payoff using a suitably constructed portfolio of the basic securities in the model. Since symmetry properties are obtained by passing to a new numeraire a natural question is whether contingent claims that are attainable in the basic financial markets are also attainable in the economy with new numeraire. This question is in fact essential for interpretation purposes since the symmetry properties above implicitly assume that the renormalized claims can be priced in the new numeraire economy and that their price corresponds to the one in the original economy.

For the case of nondividend paying assets Geman, El Karoui and Rochet (1995) prove that contingent claims that are attainable in one numeraire are also attainable in any other numeraire and that the replicating portfolios are the same. Our next theorem provides an extension of this result to dividend-paying assets. The framework of section 2 with Brownian filtration is adopted for convenience only; the results are valid for more general filtrations.

Theorem 19 *Consider an economy with Brownian filtration and complete financial market with n risky assets and one riskless asset. Suppose that risky assets pay dividends and that their prices follow Itô processes (37), and that the riskless asset pays interest at the rate r . Assume that all the coefficients are progressively measurable and bounded processes. If a contingent claim's payoff is attainable in a given numeraire then it is also attainable in any other numeraire. The replicating portfolio is the same in all numeraires.*

Proof of Theorem 19 Let $i = 0$ denote the riskless asset. The gains from trade in

the primary assets are

$$\begin{aligned} dG_t^i &\equiv dS_t^i + S_t^i \delta_t^i dt = S_t^i [r_t dt + \sigma_t^i d\tilde{z}_t], \text{ for } i = 1, \dots, n \\ dG_t^0 &\equiv dB_t = B_t r_t dt, \text{ for } i = 0. \end{aligned}$$

For $i = 0, \dots, n$, gains from trade expressed in numeraire j are

$$G_t^{i,j} = \frac{S_t^i}{S_t^j} + \int_0^t \frac{1}{S_v^j} \delta_v^i S_v^i dv \quad (40)$$

so that

$$\begin{aligned} dG_t^{i,j} &= \frac{1}{S_t^j} dS_t^i + S_t^i d\left(\frac{1}{S_t^j}\right) + \frac{1}{S_t^j} S_t^i \delta_t^i dt + d\left[S^i, \frac{1}{S^j}\right]_t \\ &= \frac{1}{S_t^j} dG_t^i + S_t^i d\left(\frac{1}{S_t^j}\right) + d\left[S^i, \frac{1}{S^j}\right]_t. \end{aligned}$$

Now let π^i represent the amount invested in asset i and consider a portfolio $(\pi^0, \pi) \in \mathbb{R}^{n+1}$ such that $\int_0^T \pi_v \sigma_v \sigma_v' \pi_v' dv < \infty$, (P -a.s.). The wealth process X generated by N , where $N^j = \pi^j / S^j$, $j = 0, \dots, n$ represents the number of shares of each asset in the portfolio, satisfies

$$dX_t = \sum_{i=0}^n N_t^i dG_t^i$$

and $X_t = \sum_{i=0}^n N_t^i S_t^i$ (this portfolio is self financing since all dividends are reinvested). Using Itô's lemma gives

$$\begin{aligned} d\left(\frac{X_t}{S_t^j}\right) &= \sum_{i=0}^n N_t^i \left(\frac{dG_t^i}{S_t^j}\right) + X_t d\left(\frac{1}{S_t^j}\right) + \sum_{i=0}^n N_t^i d\left[G^i, \frac{1}{S^j}\right]_t \\ &= \sum_{i=0}^n N_t^i \left(\frac{dG_t^i}{S_t^j} + S_t^i d\left(\frac{1}{S_t^j}\right) + d\left[S^i, \frac{1}{S^j}\right]_t\right) \\ &= \sum_{i=0}^n N_t^i dG_t^{i,j} \end{aligned}$$

i.e. the normalized wealth process can be synthesized in the new numeraire economy in which all asset prices have been deflated by the numeraire asset j . Furthermore the investment policy which achieves normalized wealth is the same as in the original economy. Consequently, any deflated payoff is attainable in the new numeraire economy when the (undeflated) payoff is attainable in the original economy. $\square$

Remark 20 (i) *The proper definition of gains from trade in the new numeraire is instrumental in the proof above. Since dividends are paid over time they must be*

deflated at a discount rate which reflects the timing of the cash flows. This explains the discount factor inside the integral of dividends in (40).

(ii) Note that Theorem 19 applies even if the numeraire chosen is a portfolio of assets or any other progressively measurable process instead of one of the primitive assets. It also applies when the portfolio is not self financing, for example when there are infusions or withdrawal of funds over time.

(iii) The results above apply for payoffs that are received at fixed time as well as stopping times of the filtration: if there exists a trading strategy that attains the random payoff X_τ where $\tau \in \mathcal{S}_{0,T}$ in the original financial market then the normalized payoff X_τ/S_t^j is attainable in the economy with numeraire asset j .

Our next result now follows easily from the above.

Theorem 21 Suppose that asset j serves as numeraire and that S^j satisfies (37). Define the probability measure Q^{j*} by

$$\begin{aligned} dQ^{j*} &= \frac{\exp(-\int_0^T (r_v - \delta_v)dv) S_T^j}{S_0^j} dQ \\ &= \exp\left(-\frac{1}{2} \int_0^T \sigma_v^j \sigma_v^{j'} dv + \int_0^T \sigma_v^j d\tilde{z}_v\right) dQ \end{aligned} \quad (41)$$

and consider the discount rate δ^j . Then the discounted prices of primary securities expressed in numeraire j are Q^{j*} -supermartingales (discounted gains from trade in numeraire j are Q^{j*} -martingales) and the price of any attainable security in the original economy can be represented as the expected discounted value of its cash flows expressed in numeraire j where the discount rate is δ^j and the expectation is under the Q^{j*} -measure.

Proof of Theorem 21 Using definition (40) of gains from trade expressed in numeraire j and Itô's lemma gives

$$\begin{aligned} dG_t^{i,j} &= \frac{1}{S_t^j} dS_t^i + S_t^i d\left(\frac{1}{S_t^j}\right) + \frac{1}{S_t^j} S_t^i \delta_t^i dt + d\left[S^i, \frac{1}{S^j}\right]_t \\ &= \frac{1}{S_t^j} S_t^i [r_t dt + \sigma_t^i d\tilde{z}_t] + S_t^i \frac{1}{S_t^j} [(\delta_t^j - r_t + \sigma_t^j \sigma_t^{j'}) dt - \sigma_t^j d\tilde{z}_t] \\ &\quad - S_t^i \frac{1}{S_t^j} \sigma_t^i \sigma_t^{j'} dt \\ &= \frac{1}{S_t^j} S_t^i [(\delta_t^j + (\sigma_t^j - \sigma_t^i) \sigma_t^{j'}) dt + (\sigma_t^i - \sigma_t^j) d\tilde{z}_t] \\ &= \frac{1}{S_t^j} S_t^i [\delta_t^j dt + (\sigma_t^j - \sigma_t^i) dz_t^{j*}], \end{aligned}$$

where $dz_t^{j*} = -d\tilde{z}_t + \sigma_t^{j'} dt$ is a Q^{j*} -Brownian motion process. Defining $S_t^{i*} = S_t^i / S_t^j$ we can then write

$$dS_t^{i*} = S_t^{i*}[(\delta_t^j - \delta_t^i)dt + (\sigma_t^j - \sigma_t^i)dz_t^{j*}]$$

i.e. the discounted price of asset i in numeraire j , $\exp(-\int_0^t \delta_v^j dv) S_t^{i*}$, is a Q^{j*} -supermartingale where discounting is at the rate δ^j . Alternatively the discounted gains from trade process

$$\exp\left(-\int_0^t \delta_v^j dv\right) S_t^{i*} + \int_0^t \exp\left(-\int_0^v \delta_u^j du\right) S_v^{i*} \delta_v^i dv$$

is a Q^{j*} -martingale. Thus, we can write the representation formula

$$S_t^{i*} = E_t^{j*} \left[\exp\left(-\int_t^T \delta_v^j dv\right) S_T^{i*} + \int_t^T \exp\left(-\int_t^v \delta_u^j du\right) S_v^{i*} \delta_v^i dv \mid \mathcal{F}_t \right].$$

The relations satisfied by primary asset prices also apply to portfolios of primary assets and therefore to any contingent claim that is attainable. This completes the proof of the theorem. $\square$

Remark 22 When a dividend-paying primary asset price is chosen as deflator the auxiliary economy has an interest rate equal to the dividend rate of the deflator. In this new numeraire cash is converted into an asset that pays a dividend rate equal to the interest rate in the original economy. If we choose the discounted price $\widehat{S}_t^j = \exp(-\int_0^t (r_v - \delta_v^j) dv) S_t^j$, which is a martingale, as numeraire the process $S_t^{i*} = S_t^i / \widehat{S}_t^j$ satisfies

$$dS_t^{i*} = S_t^{i*}[(r_t - \delta_t^i)dt + (\sigma_t^j - \sigma_t^i)dz_t^{j*}]$$

and its discounted value at the riskfree rate is a Q^{j*} -supermartingale where Q^{j*} is defined in (41). With this choice of numeraire the interest rate remains unchanged in the auxiliary economy. Cash is converted into an asset that pays a dividend rate equal to the interest rate and thus has null drift (martingale).

Remark 23 (i) Note that a payoff expressed in a new numeraire is not necessarily the same as the payoff evaluated at normalized underlying asset prices (i.e. prices expressed in the new numeraire). There is clearly equivalence when the payoff is homogeneous of degree one. With homogeneity of degree v the payoff in the new numeraire is equivalent to the payoff function evaluated at underlying asset prices that are normalized by a power of the numeraire price. Normalized asset prices (in the payoff function) then differ from asset prices expressed in the new numeraire.

(ii) A byproduct of Theorem 21 is a generalized “symmetry” property which applies to any payoff function. In this interpretation of the property the symmetric contract is simply the payoff expressed in the new numeraire.

Some extensions are worth mentioning.

Remark 24 *Note that the results on the replication of attainable contingent claims, their financing portfolios and their representation under new measures are valid even when markets are incomplete. Indeed if the claims under consideration can be replicated in a given incomplete market equilibrium (i.e. if the claims' payoffs live in the asset span) so can they under a change of numeraire. The results are also valid when the market is effectively complete (single agent economies). In this case even when claims payoffs cannot be duplicated they have a unique price which can be expressed in different forms corresponding to various choices of numeraire.*

9 Conclusion

In this paper we have reviewed and extended recent results on PCS. Features of the models considered include (i) financial markets with progressively measurable coefficients, (ii) random maturity options, (iii) options on multiple underlying asset, (iv) occupation time derivatives and (v) payoff functions that are homogeneous of degree $\nu \neq 1$. One important element in the proofs is the ability to renormalize a vector of prices and parameters which determine the payoff of the contract. Homogeneity of degree ν is sufficient in that regard but it is not a necessary condition. Another important element in the proofs is the separation between the role of informational variables and the change of measure (numeraire). Indeed while the change of measure converts the underlying assets into normalized or symmetric assets in the auxiliary financial market the information sets in the two markets are kept the same. This separation enables us to derive symmetry properties even for financial markets in which prices do not follow Markov processes. In the context of diffusion models the change of measure is instrumental for obtaining symmetry properties of option prices without restricting volatility coefficients.

Some of the results in the paper can be readily extended. Symmetry-like properties hold for multiasset contracts even when the payoff functions are not homogeneous of some degree ν (for instance when homogeneity of different degrees holds relative to different subsets of the underlying asset prices). In this instance normalized prices in the auxiliary economy involve further adjustments to dividends and volatilities. Likewise the methodology reviewed in this paper also applies, in principle, to complete financial markets with general semimartingales or even to incomplete markets provided that the securities under consideration lie in the asset span.

References

- Akahori, J. (1995), Some formulae for a new type of path-dependent option *Annals of Applied Probability* **5**, 383–8.
- Bensoussan, A. (1984), On the theory of option pricing *Acta Applicandae Mathematicae* **2**, 139–58.
- Bjerkstrand, P. and Stensland, G. (1993), American exchange options and a put–call transformation: a note *Journal of Business, Finance and Accounting* **20**, 761–4.
- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities *Journal of Political Economy* **81**, 637–54.
- Broadie, M. and Detemple, J.B. (1995), American capped call options on dividend-paying assets *Review of Financial Studies* **8**, 161–91.
- Broadie, M. and Detemple, J.B. (1997), The valuation of American options on multiple assets *Mathematical Finance* **7**, 241–85.
- Carr, P. and Chesney, M. (1996), American put call symmetry. Working paper.
- Carr, P., Jarrow, R. and Myneni, R. (1992), Alternative characterizations of American put options *Mathematical Finance* **2**, 87–106.
- Chesney, M. and Gibson, R. (1993), State space symmetry and two factor option pricing models, in J. Janssen and C. H. Skiadas, eds, *Applied Stochastic Models and Data Analysis*. World Scientific Publishing Co, Singapore.
- Chesney, M., Jeanblanc-Picqué, M. and Yor, M. (1997), Brownian excursions and Parisian barrier options *Advances in Applied Probability* **29**, 165–84.
- Dassios, A. (1995), The distribution of the quantile of a Brownian motion with drift and the pricing of related path-dependent options *Annals of Applied Probability* **5**, 389–98.
- Detemple, J. B., Feng, S. and Tian W., (2000), The valuation of American options on the minimum of dividend-paying assets. Working paper, Boston University.
- Gao, B., Huang, J.Z. and Subrahmanyam, M. (2000), The valuation of American barrier options using the decomposition technique *Journal of Economic Dynamics and Control*, to appear.
- Garman, M., (1989), Recollection in Tranquility *Risk* **24**, 1783–827.
- Geman, E., El Karoui, N. and Rochet, J.C. (1995), Changes of numeraire, changes of probability measure and option pricing *Journal of Applied Probability* **32**, 443–58.
- Girsanov, I.V., (1960), On transforming a certain class of stochastic processes by absolutely continuous substitution of measures *Theory of Probability and Its Applications* **5**, 285–301.
- Goldman, B., Sosin, H. and Gatto, M. (1979), Path-dependent options: buy at the low, sell at the high *Journal of Finance* **34**, 1111–27.
- Grabbe, O., (1983), The pricing of call and put options on foreign exchange *Journal of International Money and Finance* **2**, 239–53.
- Hugonnier, J. (1998), The Feynman–Kac formula and pricing occupation time derivatives. Working paper, ESSEC.
- Jacka, S. D. (1991), Optimal stopping and the American put *Mathematical Finance* **1**, 1–14.
- Karatzas, I. (1988), On the pricing of American options *Appl. Math. Optim.* **17**, 37–60.
- Karatzas, I. and Shreve, S. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, 1988.
- Kholodnyi, V.A. and Price, J.F. *Foreign Exchange Option Symmetry*. World Scientific Publishing Co., New Jersey, 1998.
- Kim, I.J. (1990), The analytic valuation of American options *Review of Financial Studies* **3**, 547–72.

- Linetsky, V. (1999), Step options *Mathematical Finance* **9**, 55–96.
- Margrabe, W. (1978), The value of an option to exchange one asset for another *Journal of Finance* **33**, 177–86.
- McDonald, R. and Schroder, M. (1990), A parity result for American options *Journal of Computational Finance*. Working paper, Northwestern University.
- McKean, H.P. (1965), A free boundary problem for the heat equation arising from a problem in mathematical economics *Industrial Management Review* **6**, 32–9.
- Merton, R.C. (1973), Theory of rational option pricing *Bell Journal of Economics and Management Science* **4**, 141–83.
- Miura, R. (1992), A note on look-back option based on order statistics *Hitosubashi Journal of Commerce and Management* **27**, 15–28.
- Rubinstein, M. (1991), One for another *Risk*.
- Schroder, M. (1999), Changes of numeraire for pricing futures, forwards and options *Review of Financial Studies* **12**, 1143–63.

Purely Discontinuous Asset Price Processes

Dilip B. Madan

1 Introduction

Prices of assets determined in highly liquid financial markets are generally viewed as continuous functions of time. This is true of the Black–Scholes (1973), and Merton (1973) model of geometric Brownian motion for the dynamics of the price of a stock, and of its many successors that include the stochastic volatility models of Hull and White (1987), Heston (1993) and the more recent advances into modeling the evolution of the local volatility surface by Derman and Kani (1994), and Dupire (1994). Jumps or discontinuities, when considered, have been added on as an additional orthogonal compound Poisson process also impacting the stock, as for example in Press (1967), Merton (1976), Cox and Ross (1976), Naik and Lee (1990), Bates (1996), and Bakshi and Chen (1997). This class of models is broadly referred to as jump-diffusion models and as the name suggests they are mixture models studying the high activity and low activity events by using two orthogonal modeling strategies.

The purpose of this chapter is to present the case for an alternative approach that stands in sharp contrast to the above mentioned models and synthesizes the study of high and low activity price movements using a class of purely discontinuous price processes. The contrast with the above class of models is that the processes advocated here have no continuous component, as all jump-diffusions must have, and furthermore, the discontinuities are infinite in number with moves of larger sizes coming at a slower rate than moves of smaller sizes. Additionally the jump-diffusion models have what is called infinite variation, in that the sum of absolute price moves is infinity in any interval and one must square these moves before their sum is finite (the property of finite quadratic variation) while the processes we advocate are of finite variation. Unlike jump-diffusions, our processes model price up ticks and down ticks separately and the price process can be decomposed as the difference of two increasing processes representing the increases and decreases of

prices. We shall also demonstrate that the finite variation property of the proposed models also enhances their robustness and thereby their relevance for economic modeling.

This chapter summarizes the findings of research that I have conducted over the past 15 years in collaboration with a number of coauthors. The research is still on going with a number of new and interesting developments already in place, but we shall focus attention on what has been learned to date. The papers that are summarized here are Madan and Seneta (1990) , Madan and Milne (1991), Madan, Carr and Chang (1998), Carr and Madan (1998), (1998), Geman, Madan and Yor (2000), Bakshi and Madan (1998a,b).¹

The case for purely discontinuous price processes is, as it should be, an argument with many facets. First we summarize the empirical findings on the study of both the statistical and risk neutral processes and observe the empirical need to consider discontinuous processes as relevant candidates. Statistical reality by itself, however, is not a convincing argument. Unsupported by a theoretical understanding of market fundamentals, statistical modeling is at best a spurious coincidence. One must consider the implications of a fundamental economic analysis. We show that economic analysis with the help of some deep structural mathematical results points in the same direction: the use of purely discontinuous price processes. Statistical reality and theoretical conviction are ultimately no match for success. If the wrong model is brilliantly successful in delivering results, while the right one is relatively barren then we have little choice but to work with the incorrect model, bearing in mind its limitations. To address this concern we present some of the successes of modeling with a purely discontinuous price process. We match the success of Brownian motion in option pricing and portfolio management with the success of the purely discontinuous VG process obtained on time changing Brownian motion by a gamma process. The improvement in option pricing is clear, eliminating the implied volatility smile in the strike direction, and we are able to go further in portfolio management and study the optimal management of portfolios of derivative securities, a question that is relatively untouched in the diffusion context. In fact we successfully calibrate observed derivative portfolios as optimal and employ revealed preference methods to infer what we call the position measure but is better known as the personalized state price density. The perspective of purely discontinuous price processes, we conclude, is not only correct from a statistical and theoretical viewpoint, but is also rich in results and interesting applications.

The statistical findings we summarize confirm from a variety of perspectives that the local motion of the stock price is not Gaussian. This is true of both

¹ The last of these papers is a working paper and can be obtained from my web site: www.dilip-madan.com.

the time series of moves and the pricing distribution of moves as reflected in option prices. Apart from these standard tests of normality we also consider the behavior of extremal events. Relying on asymptotic laws of maxima and minima of independent sampled observations (see Embrechts, Kluppelberg and Mikosch (1997)), we employ long time series of returns and reject the hypothesis that asset return distributions are locally Gaussian. They lie in the domain of attraction of the Fréchet distribution that includes the log gamma formulation of the VG process. Additionally we investigate empirically the relationship between arrival rates of jumps of different sizes with the jump size. The focus of our attention is on whether arrival rates display a monotonicity with respect to size, decreasing as the size rises, and whether the assumption of an infinite arrival rate is supported by a casual analysis of arrival rates. We conclude in favor of infinite and decreasing arrival rates.

From a theoretical perspective, we concentrate on the implications of no arbitrage, a property that is fundamental to all models for the asset price process. This property is shown to imply that asset prices in continuous time must be modeled by a time changed Brownian motion. The question at issue is then the nature of the time change. We investigate whether the time change could be continuous, with the resultant implication of the continuity of the price process, and show that this is possible only in economies where returns are locally Gaussian and time is locally deterministic and non-random. Given the overwhelming evidence on the lack of a locally Gaussian return distribution we are led to entertain the lack of continuity of the price process. This modeling choice is also consistent with observations on studying the relationship between time changes and economic activity, whereby we learn that time changes are related to some measure of the rate of arrival of orders or trades. As the latter have a random element, and are not locally deterministic, this suggests that such properties are inherited by the time change and hence once again we are led to the class of discontinuous price processes.

Within the class of discontinuous processes we begin our search by focusing attention in the first instance on processes with identical and independently distributed increments: a property shared with Brownian motion, the base model for the underlying uncertainty in the continuous case. This leads naturally via the Lévy–Khintchine theorem for such processes to considering Lévy processes characterized by their Lévy densities whose empirical counterparts are precisely the relationship between arrival rates of jumps of different sizes and the jump size noted earlier in our empirical analysis. When the Lévy density integrates the absolute value of the jump size in the neighborhood of zero, a case we restrict attention to, the process has finite variation and can be decomposed into the difference of two increasing processes that constitute our models for the price up and down ticks. We suggest this model as a partial equilibrium model that clears market buy orders with

an up tick price response as the order is cleared through the limit sell book. The converse being the case for market sell orders cleared through the limit buy book at a price down tick.

An alternative and interesting economic model for price responses goes back to traditional dynamic models of price adjustment that represent the rate of adjustment as a function of the level of excess demand in the economy. We term this function relating the rate of change of prices to excess demand, the *force function* of the economy. Modeling excess demand by Brownian motion we may write the price process as the difference between price increases occurring during positive excursions of Brownian motion less the cumulated decreases that occur on negative excursions of Brownian motion. Such a price process is of course open to arbitrage by trades that reverse themselves during a single excursion of Brownian motion. For example, on a single positive excursion, one buys at a price and then sells at a higher price in the same excursion. To avoid such arbitrage, we restrict equilibrium trading to equilibrium times by requiring these to occur at the zero set of Brownian motion. This is organized by evaluating the disequilibrium price process at the inverse local time of Brownian motion. The resulting price process inherits the property of being purely discontinuous from inverse local time, and the process is the difference of two increasing processes that cumulate price responses during positive and negative excursions.

The two models of discontinuous price processes, (i) Lévy processes and (ii) integrals of force functionals of Brownian motion to inverse local time, are surprisingly related under the hypothesis of complete monotonicity of the Lévy density.² Every force function has associated with it a completely monotone Lévy density and for every completely monotone Lévy density there exists an equivalent representation of the price process using a force function. The equivalence is however a consequence of some deep results from number theory and hence the surprise.

We also consider the issue of robustness of the economic model with respect to tolerance of a heterogeneity of views on parameters and observe that the property of bounded variation in the price process is critical for delivering such robustness. Our concern in robustness with respect to views on parameters is that different beliefs should naturally allow for different probabilities, but the probabilities should remain equivalent and not become singular. With infinite variation there are many cases where a change in certain parameters induces singularity of measures.

With the theoretical and statistical foundations in sufficient harmony, and two broad classes of models outlined in sufficient detail, we turn our attention to the

² The Lévy density is completely monotone if each of its two halves on the positive and negative side have the property of sign alternating derivatives or equivalently can be expressed as Laplace transforms of positive functions on the positive half line. Hence, they are essentially mixtures of exponential densities.

study of particularly rich examples in this class of models. The basic generalization of geometric Brownian motion we introduce is the VG process that introduces two additional parameters providing control over skewness and kurtosis. The model arises on evaluating Brownian motion with drift at a random time given by a gamma process. The volatility of the gamma process provides control over kurtosis while the drift in the Brownian motion before the time change controls skewness. We show that this model is successful in option pricing, eliminating the smile in the strike direction with relative ease.

Fundamental to the world of purely discontinuous price processes is the property of options being market completing assets with a genuine role to play in the economy and a natural demand for these assets by investors. Recognizing these properties, we reconsider the problem of optimal derivative investment in continuous time, keeping in place Mertonian (1971) objective functions for the investor but expanding the asset space to include all European options on the underlying stock for all strikes and maturities. We find that for HARA utilities and VG statistical and risk neutral measures the derivative investment problem may be solved in closed form and leads in such economies to a healthy demand for at-the-money short maturity options: precisely the options with the greatest liquidity in financial markets. One may view the Black–Scholes economy as teaching us about stock delta positions in option hedging, while the first lessons of investment in purely discontinuous high activity price processes are about positioning in short maturity at-the-money options.

With some courage we consider replicating actual trader derivative positions as optimal ones, allowing in the process adjustments in the level of risk aversion in power utility and a view on subjective kurtosis that may differ from the statistically observed kurtosis level. Kurtosis is particularly hard to estimate as its variance is of the order of the eighth moment. With this two dimensional flexibility, we are amazingly successful in many instances in calibrating actual spot slides as optimal wealth responses from the perspective of our continuous time optimal derivative investment model.³ Having inferred risk aversion and the characteristics of subjective probability consistent with replicating observed positions as optimal, we may construct the personalized state price density that values options at a dollar amount yielding a marginal utility that matches the future expected marginal utility from holding the option. We call this state price density the position measure and provide explicit constructions of position measures, contrasting them with the risk neutral and statistical measures. We find generally that position measures are closer to the statistical measure and lie between the statistical and risk neutral measure. This is consistent with the view that traders are aware of relative frequency of

³ The spot slide of a derivatives book graphs the value of the book as a function of the level of the underlying, typically varying the underlying in the range plus or minus 30% of spot for equity assets.

occurrence of market moves and their prices and accordingly make markets in option contracts.

The outline for the rest of the chapter is as follows. Section 2 presents a summary of the statistical results. The economic consequences of no arbitrage are described in section 3, while the two equivalent but apparently different economic models of the price process are summarized in section 4. The task of constructing specific examples consistent with the statistical and economic observations of these sections is taken up in section 5. The basic operating model of the VG process is introduced in section 6. Its successes in option pricing are summarized in section 7. Optimal solutions to the asset allocation problem with derivatives are presented in section 8 and employed to infer position measures in section 9. Section 10 concludes.

2 Properties of the price process

This section summarizes some of the broad properties of the statistical and risk neutral price process. We address issues related to the normality of the motion, the behavior of extreme moves and the shape of the density of arrival rates of price moves. The emphasis in all cases is on the movement over short horizons as we view the macro moves as cumulated short moves.

2.1 Long-tailedness of historical returns

We begin by considering some well known results about the long-tailedness of the statistical return distribution and standard chi-square goodness of fit tests of normality of the return distribution. Early results on these issues go back to Fama (1965) where both the independence of daily returns and their long-tailedness is documented. We now have data at much higher frequencies of observation and report in Table 1 results on S&P 500 futures returns at these frequencies. We focus attention on the level of the observed kurtosis and on χ^2 goodness of fit tests for normality.

We observe from Table 1 that the kurtosis is substantially higher than three, the kurtosis level of a normal distribution. The goodness of fit tests also overwhelmingly reject the hypothesis of normality for returns over short durations. We will note later, in the next section, that this has very significant implications for modeling the dynamics of the price process.

2.2 Long-tailedness in risk neutral distribution

Apart from the statistical return distribution we are also interested in the risk neutral or pricing distribution as implied by option prices. This distribution assesses the

Table 1. *High frequency tests of normality S&P 500
Futures Returns Nov. 1992–Feb. 1993.*

	1 Min.	15 Min.	Hourly	Daily
Kurtosis	58.59	13.85	5.97	10.31
χ^2 test statistic	437.12	931.85	98.323	123.84
χ^2 critical value 5%	9.26	5.7	3.57	0.989

Source: Dissertation of Thierry Ané, University of Paris IX Dauphine and ESSEC 1997.

futures price of a binary derivative that pays a dollar at a future date if the stock price is in a certain interval, as opposed to the likelihood of the occurrence of this event. The distribution may be recovered from observed option prices with the density being given by the second derivative of the European call option price, of maturity matching the future date, with respect to the option strike as derived in Ross (1976a) and Breeden and Litzenberger (1978). If the distribution describing the current prices of derivatives written on future stock price events is Gaussian then an implication is that the implied volatility obtained from equating the option price to the value given by the Black–Scholes formula, should be constant as one varies the strike for a fixed maturity. On the other hand, if this density is symmetric about a point, then the implied volatilities, though no longer necessarily flat with respect to strike, should be symmetric about a point as well. Both these implications are contradicted by what has come to be known as the implied volatility smile.

We present in Table 2 below, the implied volatility smile on S&P 500 index options, based on out of the money options using only puts for strikes below, and calls for strikes above, the spot price. These are the more liquid option markets. The time period covered is June 1988 to May 1991 and we focus attention just on the short maturity options. The choice of this focus is motivated by our intention of studying the dynamics of the stock price process, which is but the cumulation of short maturity moves.

We observe from Table 2, reading up the columns, that as the strike level rises, the implied volatility falls sharply followed by a smaller rise as one crosses the level of the spot price. We therefore clearly have a smile shape in the short maturity implied volatility, but the left and right sides are not symmetric. We may conclude from these observations that the left tail of the pricing distribution is fatter than the right tail, and this reflects a negative skewness in the distribution. The existence of the smile itself is evidence of excess kurtosis (relative to the normal distribution) in this density.

Table 2. *The smile in implied volatilities at shorter maturities below 60 days.*

Moneyiness <i>spot/strike</i>	June 1988– May 1989	June 1989– May 1990	June 1990– May 1991
<0.94	17.27	16.16	19.70
0.94–0.97	16.21	15.10	18.23
0.97–1.00	16.33	15.83	18.65
1.00–1.03	17.42	17.81	20.87
1.03–1.06	19.04	20.65	22.27
>1.06	21.84	25.70	25.57

Source: Bakshi, Cao and Chen, *Journal of Finance* (1997), page 2015.

2.3 *The behavior of extreme moves*

Tables 1 and 2 are classical results on the statistical properties of densities associated with price movements in financial markets. They summarize essentially the narrow behavior of the return distribution as may be evidenced by noting that most of the returns considered in the time series analysis are the ones with the smaller magnitudes, and the range of moneyiness reported in the implied volatility curves is just within six percentage points over an average period of a month. Hence the evidence presented is that of lack of normality in the neighborhood of the zero return and one might wonder whether at least the tail of the distributions is Gaussian. For the risk neutral distribution this has the implication that the implied volatility curve flattens out as one gets into deep out-of-the-money options on both sides, though the level at which the curves flatten out may be different on each side.

To focus attention on the behavior of the tails of the distribution with a view to addressing whether this may be Gaussian, we consider the behavior of extremes. It is shown in Embrechts, Kluppelberg and Mikosch (1997) that the asymptotic distribution of the maximum and minimum of independent drawings from a Gaussian distribution is given up to shift and scale by the Gumbel distribution. The other possible asymptotic distributions for these extremal events are, again up to shift and scaling, the Weibull and Fréchet distributions. For distributions that have as support the positive half line, the candidate limiting distributions are just the Gumbel and Fréchet distributions.

The analysis of extreme events requires long time series of data and for this purpose we obtained data on daily returns on the Dow–Jones industrial average (DJIA) for 100 years from 1897–1997. Partitioning this data into non-overlapping intervals of 100 days, we constructed a series on the maximum percentage daily rise and the maximum percentage daily drop in the DJIA over the 100 days. We

Table 3. *Log-likelihoods of the distribution of extremal price movements maximum daily percentage rise and fall in the DJIA over 100 day nonoverlapping intervals for 100 years.*

	Maximum daily drop 100 days		
	Gumbel	Fréchet	<i>P</i> -value
1897–1997	768.37	808.58	0.00
1897–1945	380.22	389.98	0.01
1946–1997	409.93	434.74	0.00
	Maximum daily rise 100 days		
	Gumbel	Fréchet	<i>P</i> -value
1897–1997	811.66	833.77	0.01
1897–1945	395.79	408.92	0.01
1946–1997	358.33	432.95	0.01

Source: Bakshi and Madan (1998), *What is the probability of a stock market crash*, working paper, University of Maryland.

then artificially nested the Gumbel and Fréchet log likelihoods and tested the null hypothesis that the distribution of the extreme event is Gumbel, the limit of the Gaussian tail. Table 3 presents these results.

Table 3 demonstrates that the normality hypothesis may also be rejected as a model for the tails of the statistical distribution of daily returns. Given the evidence on excess kurtosis, we would conjecture that these tails are heavier than Gaussian and if the property is shared with the risk neutral distribution, as we suspect it is, then implied volatilities must continue to rise as we get deeper out-of-the-money, i.e., the implied volatility curves do not flatten out at either end of the strike range. At this point we do not have documentary evidence on very deep out-of-the-money implied volatilities but observations from current market quotes on S&P 500 index options would suggest that this may well be the case.

2.4 *The structure of the arrival rates of price moves*

The arguments of this chapter lead us to considering as models for the dynamics of stock prices, purely discontinuous processes. Such processes, when they have independent and identically distributed increments, are characterized by their Lévy densities that essentially count the rate of arrival of jumps of different sizes. These are a wide class of processes, and structural properties if supported by data are beneficial in limiting the class of models that need to be considered. One such structural property is complete monotonicity of the Lévy density, whereby large

jumps occur at a smaller rate than small jumps. This is a reasonable property to expect as market participants facing price increases on buy orders and decreases on sell orders have an incentive to minimize these impacts. Another structural property is the aggregate arrival rate of jumps or moves, that could be finite or infinite. We note in this regard that Brownian motion is an infinite activity process as the actual sum of absolute price moves is itself infinite for Brownian motion as it is a process of infinite variation. We note further that jump-diffusions employ a compound-Poisson process for the arrival of jumps that have a finite arrival rate with the magnitude of jumps having, once again, a normal distribution.

The models we propose in this chapter have infinite arrival rates of jumps and in this regard they are closer to Brownian motion, but unlike Brownian motion they are processes of finite variation. This requires that the integral of the Lévy density be infinite, but the density times the jump size should have a finite integral near zero. A typical Lévy density meeting these conditions is of the form $\alpha \exp(-\beta |x|)/|x|^{1+\rho}$ for jump size x with $\rho > 0$. The log arrival rate is in this case linear in the jump size and the log of the jump size, with the coefficient on the log of the jump size being above unity. For $\rho > 1$ we have infinite variation and $\rho = 0$ is the case of the gamma process, or in this case the difference of two gamma processes which we will note later is the VG model. On the other hand if the jump sizes are exponentially distributed with a finite arrival rate, as postulated for example in Das and Foresi (1996) then the log arrival rates are linear in just the size with the coefficient on log size being 0 or $\rho = -1$. In contrast the log arrival rate of the compound-Poisson process with Gaussian jump sizes (see Cox and Ross (1976)) is linear in the size and the square of the size. Since the exponential of a negative quadratic shifts from being concave near zero to convex near infinity, such a Lévy density is not completely monotone.

A cursory evaluation of these structural properties may be simply made by regressing log arrival rates on the size of jumps, their log and their square. For our 100 year data on daily returns on the DJIA we counted the number of arrivals of jumps in the different size categories and then regressed the log of the empirically observed arrival rate on the size of the jump, its log and its square. For the Cox and Ross (1976) model the log arrival rates have a single representation that is not distinguished by the sign of the jump, while for the Das and Foresi and VG type models, the parameters vary with sign, so the latter two model estimates allow for this by separating out the positive and negative moves. Table 4 presents the results of these regressions.

From Table 4 we observe that the coefficient of log size in the first two regressions is significantly different from zero and may even be close to two, which definitely argues against a process with a finite arrival rate, as in Das and Foresi (1996). As in a number of cases the coefficient is estimated above two, the process

Table 4. *Regression of log arrival rates on the sizes of jumps. Standard errors are in parentheses.*

Log arrival rates of drops				
	Constant	Jump size	Log size	R^2
1897–1997	–9.88 (1.44)	–31.6 (8.36)	–1.92 (0.32)	0.97
1897–1945	–8.51 (1.45)	–33.0 (8.53)	–1.65 (0.32)	0.97
1946–1997	–12.35 (2.22)	–32.0 (17.78)	–2.41 (0.45)	0.95
Log arrival rates of rises				
	Constant	Jump size	Log size	R^2
1897–1997	–11.55 (1.71)	–24.5 (9.10)	–2.25 (0.38)	0.96
1897–1945	–10.29 (1.65)	–25.4 (8.97)	–1.99 (0.37)	0.97
1946–1997	–13.66 (3.23)	–25.8 (24.45)	–2.67 (0.65)	0.93
Arrival rates for jump diffusion				
	Constant	Jump size	Size ²	R^2
1897–1997	–3.66 (0.53)	–1.73 (3.86)	–447 (66)	0.70
1897–1945	–3.36 (0.48)	–1.77 (3.66)	–421 (62)	0.71
1946–1997	–3.17 (0.65)	1.54 (8.98)	–928 (191)	0.64

Source: Bakshi and Madan (1998), *What is the probability of a stock market crash*, working paper, University of Maryland.

may be one of infinite variation. However, we cannot reject the hypothesis that this coefficient is below two and hence we may have a process of finite variation. As will be argued later, there are other reasons for entertaining a finite variation process and in the absence of strong evidence to the contrary we conclude in favor of finite variation processes with infinite arrival rates.

Regarding the comparison with the Cox and Ross (1976) process with quadratic log arrival rates, we note that the linear term is in all cases insignificant, suggesting a pure quadratic model, but note further that one explains only up to 70% of the variation in arrival rates compared with up to 97% of the variation using the completely monotone density.

2.5 Summary of empirical observations

We note from Tables 1 and 2 that both the statistical and risk neutral distributions are for short intervals, not normal distributions. They have significant levels of excess kurtosis and the risk neutral distribution in particular is also skewed to the left with a heavier left tail than a right tail. This absence of normality continues into the tail of the densities as reflected by an analysis of extremes in Table 3. From Table 4 we infer that a reasonable model could be a pure jump model with an infinite arrival rate – Lévy density integrating to infinity – and a process of finite variation. We also infer from Table 4 some support for a completely monotone Lévy density. Heavy risk neutral tails, if confirmed, imply that implied volatilities are strictly *U*-shaped and do not flatten out as one moves deep out of the money in both directions.

3 The implications of economic theory

One of the most far reaching implications of economic theory are now recognized to be the consequences of the no arbitrage hypothesis. From early beginnings with the Ross' (1976) theory of arbitrage, and its application to option pricing by Black and Scholes (1973) and Merton (1973) to the development of the martingale theory of pricing by Harrison and Kreps (1979) and Harrison and Pliska (1981) this hypothesis has yielded many deep and interesting results. We demonstrate in this section a continuation of these lessons and draw out more exactly the implications of this hypothesis for modeling the dynamics of the asset price.

Before proceeding we note an important proviso with regard to this hypothesis. Financial markets may display arbitrage opportunities and there are many documented “so-called” anomalies that are suggestive of such a possibility, yet it remains true that models of the price process to be employed in developing derivative pricing models must be free of arbitrage. This is so for the simple reason of preventing traders from arbitraging a firm quoting arbitrageable prices. That models must be arbitrage free goes without question.

3.1 The stochastic process implications of no arbitrage

Four results, one from mathematical finance and the other three from the theory of stochastic processes, form the foundations for the stochastic process implications of the hypothesis of no arbitrage. The first of these results, from mathematical finance, demonstrates that the absence of arbitrage is equivalent to the existence of an equivalent martingale measure. The other results, from the theory of stochastic processes, characterize martingales.

3.1.1 No arbitrage and martingales

This result has many proofs or no proof depending on the context and meaning to be attached to the idea of no arbitrage. In discrete time and with finitely many states there is no ambiguity and the result is true with a proof going back to Harrison and Kreps (1979). At the other extreme we have continuous time and states given, at a minimum, by the relatively large set consisting of the paths of the stock price process. Here the existence of martingale measures easily implies the absence of arbitrage, but the implication in the reverse direction is not available, and this is the direction that concerns us here. Essentially the hypothesis of no arbitrage, merely asserting that one cannot combine a portfolio of existing assets to earn a non-negative, non-zero, cash flow at a negative current price is too weak to deduce the existence of a martingale measure. For interesting counterexamples of economies satisfying no arbitrage and yet not satisfying the existence of a martingale measure the reader is referred to Jarrow and Madan (1998).

In these richer contexts allowing an infinity of dynamic trading strategies, the hypothesis of no arbitrage must be strengthened to permit deduction of a martingale measure. The strengthening required is topological in nature and requires that one not be able to construct an approximation to an arbitrage opportunity in some limiting sense, and then it does follow that there exists an equivalent martingale measure. The first results in this direction are due to Kreps (1981). The difficulty with the result of Kreps (1981) is the weak sense in which the limit is taken, as the definition of approximation lacks a sense of uniformity, and what is regarded as an approximation may not be so from the perspective of other economic agents.

The strongest results in this direction are due to Delbaen and Schachermayer (1994). They employ a strong and uniform sense of no arbitrage and show that if there is no random sequence of zero cost trading strategies converging in this strong sense to a non-negative, non-zero cash flow, with the random sequence being uniformly bounded below by a negative constant, then there exists a martingale measure and the converse holds as well. They term this hypothesis No Free Lunch with Vanishing Risk (NFLVR) and prove that it is equivalent to the existence of an equivalent martingale measure.

3.1.2 Martingales and semimartingales

The second important result in ascertaining the stochastic process implications of the hypothesis of no arbitrage is Girsanov's theorem. This is pointed out by Delbaen and Schachermayer (1994) and amounts to noting that if there exists a change of measure from the true statistical measure P to a martingale measure or risk neutral measure Q such that under Q discounted asset prices are martingales, then it must be that under P the price process was a semimartingale to begin with.

This is a very useful realization as it informs us that models for price processes may safely be restricted to the class of semimartingale processes. Since the class of semimartingales is very wide indeed, one might argue that this is not a very important insight. On the other hand, a lot is known about the structure of semimartingales and for a modeler it is useful to know that the search may be constrained by this structure. Some recent examples of proposals for stock price processes that are not semimartingales include the use of fractional Brownian motion with the arbitrage demonstrated in Rogers (1997).

Semimartingales are a difficult concept to communicate in precision, as they go beyond the idea of a simple concept and are in fact a fairly complete and very general theory of random processes, yet given their established importance to the field of mathematical finance today, it is imperative that we communicate some of the flavor of this theory, and do so with brevity. There are at least two approaches, one analytical and the other structural and it is best to consider the structural approach. From this perspective a semimartingale is described by its decomposition into a martingale plus a very general model for the drift of the process. This certainly includes linear drift but also more general models of the drift. One merely requires that this process be of finite and integrable variation, as well as being predictable (i.e. the limit of left continuous functions). Examples include Brownian motion with drift, solutions to stochastic differential equations like the mean reverting Cox, Ingersoll and Ross (1985) interest rate process and the VG model (Madan, Carr and Chang (1998)) with drift to be discussed later in the chapter. To appreciate what is not a semimartingale, we consider the discrete time continuous state context studied by Jacod and Shiryaev (1998) where they show that the no arbitrage property is lost if zero is not in the relative interior of the support of the multivariate return distribution over the discrete time step and hence the arbitrage. We also learn from this paper that not all semimartingales are stock price models, as calendar time is a semimartingale with a zero martingale component and has arbitrage if it was a price process. The important property is to get zero into the relative interior of the support, at least in discrete time. Price processes must be semimartingales with a non-zero martingale component.

3.1.3 Semimartingales and time changed Brownian motion

The next result we employ in developing our understanding of the stochastic process implications of no arbitrage is a fundamental characterization of all semimartingales, due to Monroe (1978). This remarkable result shows that every semimartingale can be written as a Brownian motion (possibly defined on some adequately extended probability space) evaluated at a random time. This result is somewhat surprising at first, since Brownian motion, even if evaluated at a random time, is suggestive of a martingale and as noted earlier semimartingales include

simple linear drifts like time itself. However, this is only a problem at first glance as the time change need not be independent of the Brownian motion and calendar time t , for example, is Brownian motion $W(t)$ evaluated at the first time $T(t)$ at which this same Brownian motion reaches t .

By this result the study of price processes is reduced to the study of time changes for Brownian motion and one may consider both independent and dependent time changes. One might ask what the time change represents? Ignoring price changes that are the possible result of noise or liquidity trades, changes in the price of an asset occur through trades motivated primarily for reasons of information. The cumulated arrival of relevant information is a reasonable, economically meaningful measure of the time change, that gets translated into buy or sell orders. Geman, Madan and Yor (2000) consider many models for the process of buy and sell orders and relate the time change in all these cases to some measure of economic activity. In some cases the measure is just the number of trades while in other cases time is measured by the weighted sum of order arrivals, where the weights vary with the size of the order.

When time is viewed in this economically fundamental manner the question of dependence or independence of the time change becomes an interesting and meaningful question. Certainly, some part of the order process and hence the time change, one would expect, is motivated by observations of the price process. This is the phenomenon of herding or runs on the asset. On the other hand if the market is dominated by independent analysts who view the market price as always providing us with the most efficient and accurate valuation of the asset, i.e. it is a discounted martingale under the right measure, then there is no information to be extracted from prices that the market has not already extracted and so no analysts are motivated in their trades by observations of price movements. They are bound to seek independent, and as far as possible, private information, as the motivating basis of their trading decisions. This interpretation of the process suggests an independent time change. We also note that from a mathematical modeling viewpoint, it would be easier to work with independent time changes though it is possible and we shall see cases where both representations are possible for the same process. Generally, the independent time change is the more tractable alternative and so far most of our successes come from processes of this type. The broad consistency of this hypothesis with the efficient markets hypothesis is therefore an attractive feature.

3.1.4 Continuous time changes and semimartingales

We come now to the crux of the issue, the continuity of the price process or otherwise. This brings us to the third and final result from the theory of stochastic processes shedding light on the nature of the price process as a consequence of no arbitrage. We note first that as the price process is a time changed Brownian

motion, it will be a continuous process essentially only if the time change is continuous. The implications of supposing such continuity in the time change rely on results characterizing continuous semimartingales (Revuz and Yor (1994), page 190).

Let $X(t)$ be a continuous semimartingale, be it the price process or the time change. Let $V(t)$ be the quadratic characteristic of the semimartingale $X(t)$ which exists by virtue of X being a semimartingale. In the terminology of Wall Street the process $V(t)$ is akin to the realized total variance on the process $X(t)$. If the process $X(t)$ has a well defined sense of a variance rate per unit time, or equivalently $V(t)$ is differentiable in t then the quadratic characteristic is absolutely continuous with respect to Lebesgue measure and in this case we may write the process $X(t)$ as a stochastic integral with respect to Brownian motion. Under these conditions there exist processes $a(t)$, $b(t)$ and a standard Brownian motion $W(t)$ such that

$$X(t) = X(0) + \int_0^t a(s)ds + \int_0^t b(s)dW(s). \quad (1)$$

Consider now the implications of $X(t)$ being a time change and the price process in turn. If $X(t)$ is a time change, then it is an increasing process and so $b(t)$ must be identically zero. This implies that the time change is locally deterministic with no uncertainty in local rate of time change which is then $a(t)$. If we view the time change, as suggested earlier, as a measure of economic activity, proxied by the rate of arrival of information, orders, or size weighted orders then one would expect some local uncertainty in the time change and this argues against the use of a locally deterministic time change and hence, by implication, a continuous semimartingale as a model for the price process.

On the other hand if one views $X(t)$ directly as a price process, the representation (1) argues that the local motion of the stock return must be Gaussian. Given the considerable evidence cited against the likelihood of this possibility, we conclude once again that a continuous semimartingale is not an appropriate model for the price process. Now it is possible that there is a continuous martingale component in the price process in addition to a jump component as is the case of jump diffusions, but the necessity of introducing such a diffusion term onto a functioning purely discontinuous model must be separately argued for. As we will observe, the latter class of models contain many alternatives capable of approximating very closely the structural characteristics of diffusions.

3.1.5 Summary of the consequences of no arbitrage

We showed in this section that no arbitrage implies, via the existence of an equivalent martingale measure, that the price process is a semimartingale. We then observed that all semimartingales are time changed Brownian motions, time changed

by a random increasing time change. The resulting process could be continuous only if the time change is locally deterministic. Relating time changes to measures of economic activity with some local uncertainty we argued that the price process was not a continuous process. We also observed that such continuity implies that the process is locally Gaussian, for which we have ample evidence to the contrary, and so once again we concluded that the process cannot be continuous. The remaining sections will take up the issue of modeling using purely discontinuous processes and demonstrate their effectiveness. The need to add on an additional continuous process onto a functioning purely discontinuous process must in our view be argued for on theoretical and empirical grounds. Carr, Geman, Madan and Yor (2000) present evidence to the contrary.

4 Economic models of finite variation for asset price processes

Statistical and economic analysis suggests that we entertain purely discontinuous price processes with possibly infinite arrival rates, and finite variation. An attractive feature of finite variation processes is that they may be decomposed as the difference of two increasing processes, a property lost in Brownian motion and other processes of infinite variation. This permits, for the first time, a separation of the price process into the process of up ticks and down ticks. Our analysis of optimal contracting in such economies indicates that the major demand for short maturity at-the-money options in such economies arises from a desire on the part of investors to be positioned differently with respect to upward and downward movements in the market, a position not attainable by direct stock investment alone. Hence options, and short maturity at-the-money options in particular, play a fundamental role in such economies: a role that may be consistent with casual observations of high activity in these markets. The next step forward from correctly adjusting one's delta or stock position is the optimal positioning of the up and down deltas via option trades. To effectively answer these questions it is imperative that we focus attention, separately, on the up and down forces of the market. We propose here two classes of models, accomplishing this objective. The models differ in their primitives and are structurally distinct, yet we show in the next section that under some fairly reasonable conditions, they are in fact equivalent. However, tractability is enhanced by working with both specifications as it can be difficult to find the equivalent formulation from the alternate perspective.

The first class of models takes as primitives two increasing processes that represent cumulated orders to buy and sell at market and models the price responses as these orders are cleared through the limit sell and buy books respectively. Economic activity and the related concepts of economic time reflect cumulated orders

of both types in this representation of the price process. We term this class of models the Order Processing Models (OPM).

The second class of models is related to traditional models of dynamic price adjustment with price changes expressed as a function of the level of excess demand in the economy. This response function is termed the force function of the economy as it measures price pressure in its relationship with excess demand. The excess demand itself is modeled by a Brownian motion with the equilibrium points given by the zero set of Brownian motion. Economic time in these models is given by cumulated squared price responses or the realized variance. This class of models we refer to as Dynamic Price Adjustment Models (DPA).

4.1 Prices in the order processing model (OPM)

The primitives in this view of the price process are two increasing processes that represent cumulated market buy orders, $U(t)$, and cumulated market sell orders $V(t)$. We have noted in our discussion of time changes that increasing random processes with local uncertainty are necessarily purely discontinuous. By taking as primitives such increasing random processes, the fundamental uncertainties of the economy are discontinuous and prices modeled as market responses to such inherit this property. Defining the jumps in the processes $U(t)$ at time t by $\Delta U(t) = U(t) - U(t_-)$ where we note that the processes are by construction right continuous with left limits and $U(t) = \lim_{s \downarrow t} U(s)$ while $U(t_-) = \lim_{s \uparrow t} U(s)$ and likewise for $V(t)$, $V(t_-)$ and $\Delta V(t)$. The property of being increasing and purely discontinuous implies that

$$U(t) = \sum_{s \leq t} \Delta U(s)$$

$$V(t) = \sum_{s \leq t} \Delta V(s)$$

so that the current value of each process is just the sum of all the jumps that have occurred to date.

Price changes are modeled in Geman, Madan and Yor (2000) by market responses to these market buy orders. Here we describe the process of price increases. The magnitude $\Delta U(t)$ is viewed as a buy order at the prevailing price of $p(t_-)$ which by construction cannot be accessed. There is a downward sloping demand curve $q^{du}(p(t)/p(t_-), \Delta U(t), t)$ that is $\Delta U(t)$ at $p(t) = p(t_-)$ and an upward sloping supply curve $q^{su}(p(t)/p(t_-), \Delta U(t), t)$ that is zero at $p(t) = p(t_-)$ that must be equated to determine both the quantity transacted $q^u = q^{du} = q^{su}$ and

the price response $p(t)$. The solution gives the price response in log form by

$$\ln \left(\frac{p(t)}{p(t_-)} \right) = \Phi^u(\Delta U(t), t).$$

A similar analysis yields the price response to a market sell order

$$\ln \left(\frac{p(t)}{p(t_-)} \right) = \Phi^v(\Delta V(t), t).$$

The price process is obtained as an aggregation of the price responses to market buy and sell orders

$$\ln(p(t)) = \ln(p(0)) + \sum_{s \leq t} \Phi^u(\Delta U(s), s) - \sum_{s \leq t} \Phi^v(\Delta V(s), s)$$

and is by construction the difference of two increasing processes, and therefore a finite variation process. It is also purely discontinuous in that it is precisely the sum of all its jumps. Geman, Madan and Yor (2000) rewrite such processes in many cases as time changed Brownian motion and study the relationship between the time change and the market primitives, showing that the time change is generally a size weighted sum of the market buy and sell order processes. Hence their interpretation as measures of the level of economic activity.

4.2 The dynamic adjustment model (DPA)

This formulation of the price process begins with a traditional price adjustment model of the form

$$\frac{d \ln(p)}{dt} = f(z(t))$$

where $z(t)$ is a measure of excess demand and f represents the force by which prices respond to excess demand in the economy. This function we term the force function of the economy. By construction $f(x) \geq 0$ for $x > 0$ and $f(x) \leq 0$ for $x < 0$.

Excess demand is exogeneously modeled as dominated by new information and is given by a Brownian motion $W(t)$. It follows that

$$\ln(p(t)) = \ln(p(0)) + \int_0^t f(W(s))ds.$$

Equilibrium times are of course given by the zero set of Brownian motion and there are arbitrage opportunities to be made during upward or downward rallies by buying or selling and then reversing the trade before the end of the rally. Such intra rally trades are not available to general market participants whose price access is only at equilibrium times. The restriction to equilibrium times, the zero set of

Brownian motion, is accomplished by evaluating the above process at the inverse local time of Brownian motion at zero, $\sigma(t)$. We therefore define

$$\ln(p(t)) = \ln(p(0)) + \int_0^{\sigma(t)} f(W(s))ds. \quad (2)$$

This process is once again a purely discontinuous process, inheriting this property from that of inverse local time. It may be decomposed as the difference of two increasing processes

$$\ln(p(t)/p(0)) = \int_0^{\sigma(t)} f^+(W(s))ds - \int_0^{\sigma(t)} f^-(W(s))ds$$

where $f^+(x) = f(x)\mathbf{1}_{(x \geq 0)}$; $f^-(x) = f(x)\mathbf{1}_{(x \leq 0)}$, and is a process of finite variation under the condition

$$\int_{-K}^K |f(x)| dx < \infty \quad \text{for all } K.$$

It is interesting to enquire into the nature of the force function in the economy. For example, if $f(x) > 0$ for all $x > 0$ and $f(x) < 0$ for $x < 0$ then the price process is one with an infinite arrival rate of jumps. On the other hand there are finitely many jumps in any interval if $f(x) = 0$ in a neighborhood of zero. Another interesting question is whether the force is immediately infinite and decreasing for larger excess demands or whether it rises with the level of excess demand. Geman, Madan and Yor (2000) present many explicit solutions that may be employed to answer such questions. They also show that such a process may be written as Brownian motion evaluated at a time change that aggregates the squared price responses and is thereby a measure of realized variance.

5 Prices as Lévy processes

Finite variation asset price processes are by construction the difference of two increasing processes and section 4 has described two classes of economic models that give rise to such processes. We now wish to construct specific examples of such processes that may be evaluated empirically in their adequacy as models for the statistical dynamics of the price process, and as models for the pricing densities reflected in option prices. This statistical evaluation is enhanced if one has effective descriptions of the transition densities for use in maximum likelihood estimation and closed form or otherwise fast and accurate computation methods for the prices of European options when the underlying process is in the described class.

Both these objectives are simultaneously met by an analytic closed form for the characteristic function of the log of the stock price at a future date. The density is then easily evaluated by Fourier inversion and maximum likelihood estimation

is feasible, alternatively one may also follow the methods outlined in Madan and Seneta (1989) and estimate parameters by maximum likelihood on transformed variates. Option prices are easily obtained from the characteristic function and this is described in Bakshi and Madan (1998) and a faster algorithm is provided in Carr and Madan (1998). Carr and Madan show how to analytically write the Fourier transform in log strike of an exponentially damped call price, in terms of the characteristic function of the log stock price. The damped call price and call price are then obtained by a single Fourier inversion that may even invoke the fast Fourier transform. The characteristic function of the log stock price is therefore seen as the key to efficient model validation from both a statistical and risk neutral perspective.

5.1 The characteristic function of log price relatives

In constructing alternatives to Brownian motion as models of the fundamental uncertainty driving the stock price, that may meet our requirements of being a purely discontinuous process of finite variation with a possibly infinite arrival rate of shocks, we focus in the first instance on keeping all the properties of Brownian motion except those that must be given up. We are well aware that just as more complex models allowing for stochastic volatility and correlations of various sorts can be constructed out of Brownian motions by combining them in various ways, the same can be done with any candidate process that replaces Brownian motion.

The first property of Brownian motion that we seek to keep is the analytically rich property of being a process of independent increments, identically distributed over non-overlapping intervals of equal lengths of time. This introduces a homogeneity of the base uncertainty across time, that may be altered through parametric shifts in later developments. In any case, for modeling the local motion, homogeneity should be a reasonable hypothesis from at least the perspective of a local approximation that employs some average density of moves, even if the actual ones are state contingent and time varying.

The second property, which we may or may not keep, is that of finite moments of all orders. We are modeling continuously compounded returns and this should in principle be a bounded random variable, even if it is difficult to organize this within a modeling context, and hence the finiteness of moments is really a non-issue. Considerations of analytical tractability may on occasion require us to consider processes with infinite moments, but my priority is to avoid them as far as possible.

The theory of stochastic processes has a lot to teach us about processes meeting these conditions. Such processes are called infinitely divisible and the Lévy–Khintchine theorem (see Feller (1971) and Bertoin (1996)) provides us with a complete characterization of the characteristic function. Specifically, let $X(t) =$

$\log(S(t))$ be the continuous time process for the log of the stock price with mean μt , and further suppose that $X(t)$ is a finite variation process of independent identically distributed increments. Then there exists a unique measure Π defined on $\mathbb{R} - \{0\}$ such that

$$\phi_{X(t)}(u) \stackrel{\text{def}}{=} E [\exp(iuX(t))] = \exp \left(iu\mu t + t \int_{-\infty}^{\infty} (e^{iux} - 1) \Pi(dx) \right).$$

The measure Π is called the Lévy measure of the process and $X(t)$ is a Lévy process. When the measure has a density $k(x)$, we may write

$$\phi_{X(t)}(u) = \exp \left(iu\mu t + t \int_{-\infty}^{\infty} (e^{iux} - 1) k(x) dx \right) \quad (3)$$

and we refer to the function $k(x)$ as the Lévy density.

Heuristically the density $k(x)$ specifies the arrival rate of jumps of size x and the Lévy process $X(t)$ is a compound Poisson process with a finite arrival rate if the integral of the Lévy density is finite. We shall primarily be concerned with Lévy processes with an infinite arrival rate. The Lévy process may always be approximated by a compound Poisson process obtained by truncating the Lévy density in a neighborhood of zero, and using as an arrival rate

$$\lambda = \int_{|x| > \varepsilon} k(x) dx$$

and as a density for the jump magnitude conditional on the arrival, the density

$$g(x) = \frac{k(x) \mathbf{1}_{|x| > \varepsilon}}{\lambda}.$$

The convergence occurs as we let $\varepsilon \rightarrow 0$. Geman, Madan and Yor (2000) present many examples of candidate Lévy processes that are associated with the two economic models OPM and DPA of section 4.

5.2 Robustness of finite variation Lévy processes

Continuous time processes with continuous sample paths have a certain lack of robustness best illustrated by considering geometric Brownian motion under two different but close volatilities. Two individuals could perhaps hold such different views on volatility but as a consequence their probability measures are no longer equivalent but are in fact singular. The set of paths receiving probability 1 under one measure has probability 0 under the other measure. The measures are not robust, in the sense of equivalence, to different volatility beliefs. This lack of robustness is really a consequence, not of continuity, but of infinite variation.

Hence, remaining in the class of finite variation processes enhances robustness of the models to heterogeneity of views on various parameters.

To appreciate this point we note (Jacod and Shiryaev (1980), page 159) that when two Lévy processes with Lévy densities $k(x)$ and $k'(x)$ are equivalent then there exists a positive measurable function $Y(x)$ such that

$$k'(x) = Y(x)k(x) \quad (4)$$

and

$$\int_{-\infty}^{\infty} (|x| \wedge 1) (Y(x) - 1) k(x) dx < \infty. \quad (5)$$

One may rewrite (5) on employing (4) as

$$\int_{k' < k} (|x| \wedge 1) (k(x) - k'(x)) dx + \int_{k' > k} (|x| \wedge 1) (k'(x) - k(x)) dx < \infty \quad (6)$$

and observe that on the set $|x| > 1$ the required integrability holds by virtue of the integrability of the Lévy densities on this set. On the set $|x| < 1$ we have the integrability condition

$$\int_{k' < k} |x| (k(x) - k'(x)) dx + \int_{k' > k} |x| (k'(x) - k(x)) dx < \infty$$

and this condition essentially requires that the difference between the two Lévy measures be a finite variation process and holds automatically if both Lévy processes are of finite variation. Hence for finite variation processes, equivalence just requires absolute continuity of the measures with respect to each other or the condition (4) with no integrability conditions. Restrictions on the ability to change parameters like volatility in geometric Brownian motion follow from the integrability conditions for equivalence and apply to processes with infinite variation.

In this regard one may consider the Lévy measure studied in Geman, Madan and Yor (2000) of the form

$$k(x) = \frac{e^{-x}}{x^{2+\alpha}} \text{ for } x > 0.$$

For $\alpha > 0$ this process has infinite variation and the parameter generating the infinite variation is α . This parameter cannot be changed if equivalence is to be preserved. Specifically, if

$$k'(x) = \frac{e^{-x}}{x^{2+\beta}}$$

for $\alpha \neq \beta$ and $\alpha, \beta > 0$ the two measures are no longer equivalent and it is the integrability condition (5) that fails.

5.3 Complete monotonicity (CM)

There are of course many Lévy densities that one may employ in modeling the price process. It is therefore useful if the collection of possible choices can be reduced by invoking some structural properties. One such property is that of complete monotonicity. The idea is to require the arrival rates of large jumps to be less than the arrival rates of small jumps. This suggests that $k(x)$ be decreasing in $|x|$ or that $k'(x) \leq 0$ for $x > 0$ and $k'(x) \geq 0$ for $x < 0$. The first derivative of the Lévy density is therefore of one sign on each side of zero. The property of complete monotonicity requires that all the derivatives, and not just the first, have this property of having the same sign on each side of zero. By a result of Bernstein this property is equivalent to requiring $k(x)$ for $x > 0$ to be the Laplace transform of a positive measure on the positive half line and similarly for $k(x)$ for $x < 0$. Specifically we require that there exist measures G_p and G_n ,

$$k(x) = \int_0^\infty e^{-ax} G_p(da) \text{ for } x > 0$$

$$k(x) = \int_0^\infty e^{ax} G_n(da) \text{ for } x < 0.$$

The Lévy density is then a mixture of exponential densities. An important result that follows for such Lévy densities is that the two classes of economic models OPM and DPA are equivalent under the CM property.

5.3.1 Equivalence of OPM and DPA under CM

In particular, for every force function defining the price response under DPA, the resulting price process of equation 2 is a Lévy process with a completely monotone Lévy density. Geman, Madan and Yor (2000) give numerous examples of force functions and their associated Lévy densities. For example, if the force function is x^m for some integer $m > 0$ then the process is one of independent stable increments with index $\alpha = (1/2 + m)^{-1}$.

Conversely, every Lévy process with such a completely monotone Lévy density can be written as the integral of a functional of Brownian motion up to the inverse local time of the Brownian motion. This equivalence result is an application of analytical results from number theory called Krein's theory and the specification construction of the force function from the Lévy density and vice versa remains a difficult, if not impossible task. Specifically, for the variance gamma model that we introduce next, we know the Lévy density quite explicitly but are not aware of what the force function is in this case.

6 The variance gamma model

Purely discontinuous processes of finite variation with infinite arrival rates contain a particularly tractable and parametrically parsimonious subclass of processes that is constructed from two very well known processes, Brownian motion and the gamma process. This is the “so-called” variance gamma process first studied by Madan and Seneta (1990). The process studied in Madan and Seneta (1990), was the symmetric variance gamma process that is obtained on evaluating Brownian motion at gamma time. An asymmetric risk neutral process was developed by Madan and Milne (1991) by assuming that a Lucas representative agent with power utility had to hold the risk exposure in a symmetric variance gamma process. It was shown in Madan, Carr and Chang (1998) that the resulting risk neutral process was equivalent to evaluating Brownian motion with drift at gamma time. Given the importance of asymmetry or skewness in option pricing, we focus directly on this asymmetric variance gamma process but will refer to it as the variance gamma process. The process is parametrically parsimonious in that only two additional parameters are involved beyond the volatility introduced by Black and Scholes, and these two parameters give us control over skewness and kurtosis, that are precisely the primary concern in modeling and assessing derivative risks.

6.1 The variance gamma process

Let $Y(t; \sigma, \theta)$ be a Brownian motion with drift θ and variance rate σ^2 . If $W(t)$ is a standard Brownian motion, we may write the process $Y(t; \sigma, \theta)$ in terms of $W(t)$ as

$$Y(t; \sigma, \theta) = \theta t + \sigma W(t).$$

The variance gamma process is obtained on evaluating the process Y at an independent random time given by a gamma process. For this we define the process $G(t; \nu)$ with independent increments, identically distributed over non-overlapping intervals of length h , with the increments, $G(t + h; \nu) - G(t; \nu) = g$, having the gamma density

$$p(g, h) = \frac{g^{h/\nu-1} \exp(-g/\nu)}{\nu^{h/\nu} \Gamma(h/\nu)}.$$

The mean of the gamma density is h and the variance is νh . Hence the average random time change in h units of calendar time is h and its variance is proportional to the length of the interval. The gamma density is infinitely divisible with characteristic function

$$E[\exp(iug)] = \left(\frac{1}{1 - iuv} \right)^{h/\nu}$$

and the gamma process is an increasing Lévy process with a one sided Lévy density

$$k(x) = \frac{\exp(-x/\nu)}{\nu x}, \quad \text{for } x > 0.$$

Both the gamma process and Brownian motion are highly tractable processes about which a lot is known and each process has seen many domains of application. The variance gamma process is the process $X(t; \sigma, \nu, \theta)$ defined by

$$\begin{aligned} X(t; \sigma, \nu, \theta) &= Y(G(t; \nu); \sigma, \theta) \\ &= \theta G(t; \nu) + \sigma W(G(t; \nu)) \end{aligned} \quad (7)$$

or Brownian motion with drift θ and variance rate σ^2 evaluated at the gamma time $G(t; \nu)$. Apart from the variance rate of the Brownian motion σ^2 , the two other parameters are θ and ν . We shall observe that it is θ that generates skewness while kurtosis is primarily controlled by ν .

6.1.1 Characteristic function of the variance gamma process

The characteristic function of the variance gamma process is easily evaluated by conditioning on the gamma process first and then employing the characteristic function of the gamma process itself. It has a simple analytic form of a quadratic raised to a negative power. Specifically,

$$\phi_{X(t)}(u) \stackrel{\text{def}}{=} E [\exp(iuX(t))] = \left(\frac{1}{1 - iu\theta\nu + \frac{\sigma^2\nu}{2}u^2} \right)^{\frac{t}{\nu}}. \quad (8)$$

The Black–Scholes and Merton model employing Brownian motion is a limiting case of this model since the process converges to Brownian motion with drift as one lets the volatility of the time change ν tend to zero. This may also be observed from the characteristic function on letting t/ν tend to infinity as ν tends to zero and noting that the limit is precisely $\exp(iu\theta t - \sigma^2 u^2 t/2)$ the characteristic function of Brownian motion with drift.

We also note that if θ is zero, the characteristic function is real valued and the process is therefore symmetric and there is no skewness, hence validating the claim that skewness is generated by $\theta \neq 0$. This observation is even clearer once we have constructed the Lévy measure for the VG process.

6.1.2 Moments of the variance gamma process

The moments of the VG process are easily obtained by exploiting the structure of the process or by differentiating the characteristic function. It is shown in Madan, Carr and Chang (1998) that

$$E[X(t)] = \theta t$$

$$\begin{aligned}
E[(X(t) - E[X(t)])^2] &= (\theta^2 v + \sigma^2) t \\
E[(X(t) - E[X(t)])^3] &= (2\theta^3 v^2 + 3\sigma^2 \theta v) t \\
E[(X(t) - E[X(t)])^4] &= (3\sigma^4 v + 12\sigma^2 \theta^2 v^2 + 6\theta^4 v^3) t \\
&\quad + (3\sigma^4 + 6\sigma^2 \theta^2 v + 3\theta^4 v^2) t^2.
\end{aligned}$$

We observe again that skewness is zero if $\theta = 0$. Furthermore, in the case of $\theta = 0$ we have that the fourth central moment divided by the square of the second central moment or the kurtosis is $3(1 + v)$. This leads to the interpretation that the parameter v controls kurtosis and is in fact (for $\theta = 0$) the percentage excess kurtosis over the kurtosis of the normal distribution, which is three.

6.1.3 The variance gamma process as a process of finite variation

The variance gamma process is a finite variation process and the two increasing processes whose difference is the variance gamma process are both gamma processes. This is observed by considering two independent gamma processes $\gamma_p(t)$ and $\gamma_n(t)$ with mean rates of μ_p, μ_n and variance rates v_p, v_n respectively for the positive and negative components. The characteristic functions of the two gamma processes are

$$E[\exp(iu\gamma_k(t))] = \left(\frac{1}{1 - iuv_k/\mu_k} \right)^{\mu_k^2 t / v_k} \quad \text{for } k = p, n.$$

Supposing that the two gamma processes have the same coefficients of variation and $v_k/\mu_k^2 = v$ for $k = p, n$, we may write the characteristic function of the difference of the two gamma processes as

$$E[\exp(iu(\gamma_p(t) - \gamma_n(t)))] = \left(\frac{1}{1 - iu\left(\frac{v_p}{\mu_p} - \frac{v_n}{\mu_n}\right) + u^2 \frac{v_p}{\mu_p} \frac{v_n}{\mu_n}} \right)^{t/v}.$$

The result follows on comparing this characteristic function with that of the variance gamma process and defining the mean and variance rates of the two gamma processes to be differenced accordingly. Specifically

$$\begin{aligned}
\mu_p &= \frac{1}{2} \sqrt{\theta^2 + \frac{2\sigma^2}{v}} + \frac{\theta}{2}, \\
\mu_n &= \frac{1}{2} \sqrt{\theta^2 + \frac{2\sigma^2}{v}} - \frac{\theta}{2}, \\
v_p &= \mu_p^2 v, \\
v_n &= \mu_n^2 v.
\end{aligned}$$

6.1.4 The Lévy density for the variance gamma process

The Lévy density for the variance gamma process is easily constructed from its representation as the difference of two gamma processes using the well known form for the Lévy density of the gamma process. It follows that the Lévy density of the variance gamma process is

$$k_X(x) = \begin{cases} \frac{1}{v} \frac{\exp(-\frac{\mu_n}{v_n} |x|)}{|x|} & \text{for } x < 0 \\ \frac{1}{v} \frac{\exp(-\frac{\mu_p}{v_p} x)}{x} & \text{for } x > 0. \end{cases}$$

The basic form of the Lévy density is that of a negative exponential scaled by the reciprocal of the jump size. Just as in the gamma process, the integral of the Lévy density is infinite and the process is therefore a finite variation process with infinite arrival rates of jumps. It is helpful to write the Lévy density in terms of the original parameters of the process and this leads to the expression

$$k_X(x) = \frac{\exp(\theta x / \sigma^2)}{v |x|} \exp\left(-\frac{\sqrt{2/v + \theta^2 / \sigma^2}}{\sigma} |x|\right). \quad (9)$$

The special case of $\theta = 0$ is a symmetric Lévy measure and hence the absence of skew. Negative values of θ give a fatter left tail and induce negative skewness. We also observe that as v is increased the rate of exponential decay in the Lévy measure is reduced thus raising the arrival rate of jumps of the larger size. This induces the higher kurtosis related to this parameter. The two additional parameters therefore give direct control of the two moments that data analysis indicates we need to be able to control.

6.1.5 The return density for the variance gamma process

The density of $X(t; \sigma, v, \theta)$ is available in closed form and is derived in Madan, Carr and Chang (1998). This is a closed form, in that it is expressible in terms of the special functions of mathematics, in particular the modified Bessel function of the second kind. Specifically we have that the density of $X(t) = x$ given $X(0) = 0$, $h(x, t; \sigma, v, \theta) = h(x)$ is

$$h(x) = \frac{2 \exp(\theta x / \sigma^2)}{v^{t/v} \sqrt{2\pi} \sigma \Gamma(t/v)} \left(\frac{x^2}{2\sigma^2/v + \theta^2} \right)^{\frac{t}{2v} - \frac{1}{4}} K_{\frac{t}{v} - \frac{1}{2}} \left(\frac{1}{\sigma^2} \sqrt{x^2 \left(\frac{2\sigma^2}{v} + \theta^2 \right)} \right). \quad (10)$$

There are three terms in the density, an exponential, a real power and the modified Bessel function. This is useful for maximum likelihood estimation of parameters from time series and it is also useful in providing density plots of results. Later

we report on closed forms for option prices and this incorporates a closed form for the cumulative distribution function as well, that may be used to determine critical values for extreme points in value at risk calculations.

6.2 The stock price process driven by a VG process

We replace Brownian motion in the classical formulation of the geometric Brownian motion model by the VG process and define the risk neutral process for the stock price $S(t)$ by

$$S(t) = S(0) \exp \left(rt + X(t; \sigma, \nu, \theta) + \frac{t}{\nu} \ln \left(1 - \theta\nu - \frac{\sigma^2\nu}{2} \right) \right) \quad (11)$$

where r is the constant continuously compounded interest rate. Observe from the characteristic function of the VG process that

$$\begin{aligned} E [\exp(X(t))] &= \phi_X(-i) \\ &= \left(\frac{1}{1 - \theta\nu - \sigma^2\nu/2} \right)^{\frac{t}{\nu}} \\ &= \exp \left(-\frac{t}{\nu} \ln \left(1 - \theta\nu - \frac{\sigma^2\nu}{2} \right) \right) \end{aligned}$$

and hence the mean rate of return on the stock, under the risk neutral process, is the interest rate by construction.

We note further that the limit as ν tends to zero of $\frac{1}{\nu} \ln(1 - \theta\nu - \sigma^2\nu/2)$ is by L'Hopital's rule $-\theta - \sigma^2/2$ and so for small ν this term is $-\theta t - \sigma^2 t/2$. Noting that $X(t) = \theta G(t) + \sigma W(G(t))$ but for small ν , $G(t)$ is essentially t , we get that

$$\ln S(t) = \ln S(0) + \left(r - \frac{\sigma^2}{2}\right)t + W(t)$$

or the familiar geometric Brownian motion model for the log of the stock price. Hence we have a generalization of the Black–Scholes and Merton models for the stock price. The generalization has introduced two new parameters ν, θ that we have observed give us control over skewness and kurtosis in the process.

6.2.1 Characteristic function of the log of the stock price

The characteristic function of the $\ln(S(t))$ is easily derived from that of $X(t)$, and is useful in deriving option prices by Fourier methods. Specifically we have that

$$\begin{aligned} \phi_{\ln(S(t))}(u) &\stackrel{def}{=} E [\exp(iu \ln(S(t)))] \\ &= \exp \left(iu \left(\ln(S(0)) + rt + \frac{t}{\nu} \ln \left(1 - \theta\nu - \frac{\sigma^2\nu}{2} \right) \right) \right) \phi_{X(t)}(u) \quad (12) \end{aligned}$$

where $\phi_{X(t)}(u)$ is the characteristic function of the VG process given in (8).

6.3 Variance gamma option pricing

When the risk neutral process for the stock is described by the variance gamma process for the log of stock price as in equation (11), European call options on stock of strike K and maturity t have a price, $c(S(0); K, t)$ that is given by evaluating the expected discounted cash flow

$$c(S(0); K, t) = E \left[e^{-rt} \max(S(t) - K, 0) \right]. \quad (13)$$

This valuation result is an application of the defining property of a risk neutral probability, that traded asset prices, when discounted by the value of the money market account, are martingales under this probability. The valuation result follows on noting that option prices at maturity equal the promised payoff.

The computation of the call price in equation (13) is accomplished in closed form in Madan, Carr and Chang (1998). Other approaches at efficient computation employ Fourier inversion as described in Bakshi and Madan (1998) or improvements thereof as explained in Carr and Madan (1998). We present here a brief summary of these results. The reader is referred to the original papers for further details.⁴

6.3.1 The Madan, Carr and Chang closed form

The method employed by Madan, Carr and Chang (1998) to develop a closed form for the VG option price relies on integrating the Black–Scholes formula applied to a random gamma time, with respect to the gamma density for this time. This approach requires the explicit computation of expressions of the form

$$\Psi(a, b, \gamma) = \int_0^\infty N \left(\frac{a}{\sqrt{u}} + b\sqrt{u} \right) \frac{u^{\gamma-1} \exp(-u)}{\Gamma(\gamma)} du, \quad (14)$$

where $N(x)$ is the cumulative distribution function of the standard normal variate. The call option price can be explicitly computed in terms of this Ψ function. Specifically we have that

$$\begin{aligned} c(S(0); K, t) = S(0) \Psi \left(d \sqrt{\frac{1-c_1}{v}}, (\alpha + s) \sqrt{\frac{v}{1-c_1}}, \gamma \right) \\ - K \exp(-rt) \Psi \left(d \sqrt{\frac{1-c_2}{v}}, \alpha \sqrt{\frac{v}{1-c_2}}, \gamma \right) \end{aligned}$$

⁴ Matlab programs are available for performing these computations in all the three ways described here.

where

$$\begin{aligned}
 s &= \frac{\sigma}{\sqrt{1 + \left(\frac{\theta}{\sigma}\right)^2 \frac{v}{2}}} \\
 \alpha &= -\frac{\theta}{\sigma \sqrt{1 + \left(\frac{\theta}{\sigma}\right)^2 \frac{v}{2}}} \\
 \gamma &= \frac{t}{v} \\
 c_1 &= \frac{v(\alpha + s)^2}{2} \\
 c_2 &= \frac{v\alpha^2}{2} \\
 d &= \frac{\ln\left(\frac{S(0)}{K}\right) + rt}{s} + \frac{\gamma}{s} \ln\left(\frac{1 - c_1}{1 - c_2}\right).
 \end{aligned}$$

A reduction of the Ψ function (14) to the special functions of mathematics is accomplished in terms of the modified Bessel function of the second kind and the degenerate hypergeometric function of two variables with integral representation (Humbert (1920))

$$\Phi(\alpha, \beta, \gamma; x, y) = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma - \alpha)} \int_0^1 u^{\alpha-1} (1-u)^{\gamma-\alpha-1} (1-ux)^{-\beta} e^{uy} du.$$

Explicitly we have that

$$\begin{aligned}
 \Psi(a, b, \gamma) &= \frac{c^{\gamma+\frac{1}{2}} \exp(\text{sign}(a)c) (1+u)^\gamma}{\sqrt{2\pi}\Gamma(\gamma)\gamma} \\
 &\quad \times K_{\gamma+\frac{1}{2}}(c) \Phi\left(\gamma, 1-\gamma, 1+\gamma; \frac{1+u}{2}, -\text{sign}(a)c(1+u)\right) \\
 &\quad - \text{sign}(a) \frac{c^{\gamma+\frac{1}{2}} \exp(\text{sign}(a)c) (1+u)^{1+\gamma}}{\sqrt{2\pi}\Gamma(\gamma)(1+\gamma)} \\
 &\quad \times K_{\gamma-\frac{1}{2}}(c) \Phi\left(1+\gamma, 1-\gamma, 2+\gamma; \frac{1+u}{2}, -\text{sign}(a)c(1+u)\right) \\
 &\quad + \text{sign}(a) \frac{c^{\gamma+\frac{1}{2}} \exp(\text{sign}(a)c) (1+u)^\gamma}{\sqrt{2\pi}\Gamma(\gamma)\gamma} \\
 &\quad \times K_{\gamma-\frac{1}{2}}(c) \Phi\left(\gamma, 1-\gamma, 1+\gamma; \frac{1+u}{2}, -\text{sign}(a)c(1+u)\right)
 \end{aligned}$$

where

$$c = |a| \sqrt{2 + b^2}$$

$$u = \frac{b}{\sqrt{2 + b^2}}.$$

Madan, Carr and Chang (1998) go on to employ this closed form in a detailed study of the empirical properties of VG option pricing, noting in particular the importance of skewness from the risk neutral viewpoint, and the ability of the VG model to flatten the implied volatility smile in option pricing.

6.3.2 Inversion of distribution function transforms (Bakshi and Madan)

Bakshi and Madan (1998) show that very generally one may write a call option price in the form

$$c(S(0); K, t) = S(0)\Pi_1 - K \exp(-rt)\Pi_2$$

where Π_1 and Π_2 are complementary distribution functions obtained on computing the integrals

$$\begin{aligned}\Pi_1 &= \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \operatorname{Re} \left[\frac{e^{-iuk} \phi_{\ln(S(t))}(u - i)}{iu \phi_{\ln(S(t))}(-i)} \right] du \\ \Pi_2 &= \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \operatorname{Re} \left[\frac{e^{-iuk} \phi_{\ln(S(t))}(u)}{iu} \right] du\end{aligned}$$

where $k = \ln(K)$ and $\phi_{\ln(S(t))}(u)$ is the characteristic function of the log of the stock price given in this case by (12).

Bakshi and Madan (2000) study the general spanning properties of the characteristic functions and their relationship to the spanning properties of options. They also express the general relationships between the two probability elements in option pricing providing a discussion of cases where they are analytically linked in their transforms.

6.3.3 Inversion of the modified call price (Carr and Madan)

Carr and Madan (1998) define the Fourier transform of the modified call price by

$$\psi(v) = \int_{-\infty}^{\infty} e^{ivk + \alpha k} c(S(0); e^k, t) dk$$

where $k = \ln(K)$, and the multiplication by $\exp(\alpha k)$ for $\alpha > 0$ dampens the call price for negative values of log strike. They show generally that

$$\psi(v) = \frac{e^{-rt} \phi_{\ln(S(t))}(v - (\alpha + 1)i)}{\alpha^2 + \alpha - v^2 + i(2\alpha + 1)v}.$$

The call option price may then be obtained on a single Fourier inversion of ψ that may also employ the fast Fourier transform to evaluate

$$c(S(0); K, t) = \frac{\exp(-\alpha k)}{\pi} \int_0^\infty e^{-ivk} \psi(v) dv.$$

Carr and Madan (1998) also consider other strategies for speeding up the pricing of options using the characteristic function of the log of the stock price, and the methods should be useful for a variety of Lévy processes.

6.4 Results on option pricing performance

The variance gamma option pricing model was tested in Madan, Carr and Chang (1998) on data for S&P 500 options for the period January 1992 to September 1994. It was noted there that the skew is significant and the three parameter process effectively eliminates the smile in option prices in the direction of moneyness. The pricing errors are generally between 1 and 3 percent for options on the relatively liquid stocks and indices. The maturities we work with get fairly small and are as low as a couple of days at times, while the range of strikes are quite wide and may be up to 20 to 30% out-of-the-money. Yet on this wide range of strikes and low maturities the model provides adequate fits.

Here we provide some illustrations of the results for options on the SPX and Nikkei indices. Figures 1 and 2 provide graphs of the prices of out-of-the-money options on these two indices along with the theoretical price curve as fit by the VG model. For strikes above at-the-money the options are calls while puts are used for the strikes below the spot. The typical V shaped price structure observed in markets is basically consistent with that of the negative exponential in the absolute value of the size of the move, that is the local structure of the VG model. The difficulty for Gaussian based models is precisely the fact that for these models option prices of out-of-the-money options fall off too rapidly, being a negative exponential in the square of the move, compared to market. We observe here that the essential structure of price decay is consistent with the building block of completely monotone Lévy densities, the double negative exponential.

7 Asset allocation in Lévy systems

Apart from the successes of Lévy processes in option pricing, and the VG model in particular, these processes are associated with financial markets that are incomplete with respect to dynamic trading in the stock and the money market account. In such economies, with stock prices driven by an infinite arrival finite variation Lévy process, European options are market completing assets and one may study the

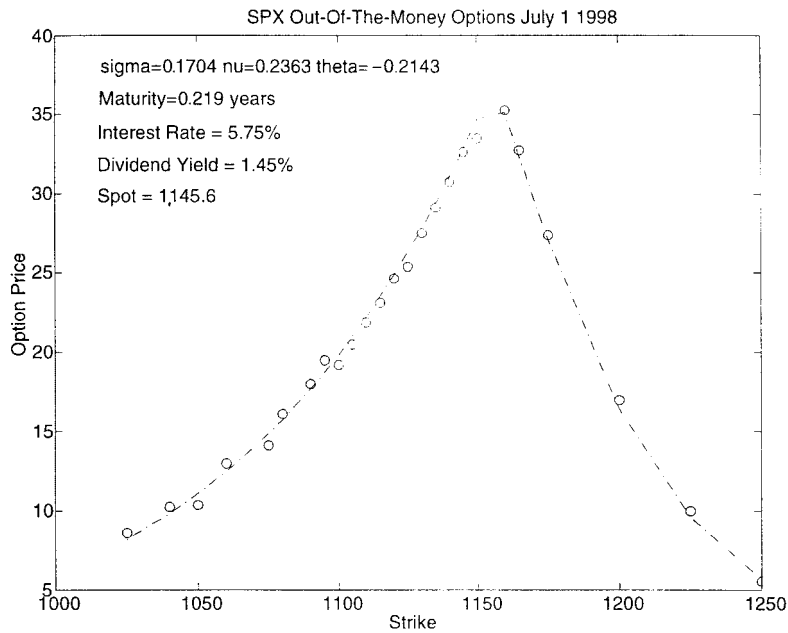


Fig. 1. Out-of-the-money option prices on the SPX index and the price curve as fit by the VG model.

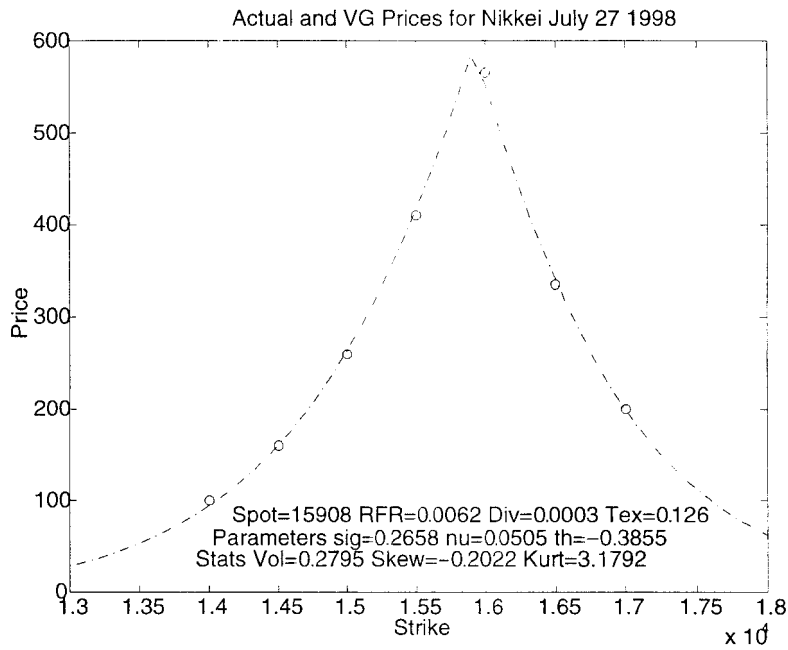


Fig. 2. Out-of-the-money option prices on the Nikkei Index and the price curve fit by the VG model.

question of the optimal demand for these assets by investors. In contrast, for the traditional economy, where options are redundant assets there is no demand for these assets.

With these observations in mind, Carr, Jin and Madan (2000) proceed to reformulate the Merton problem for optimal consumption and investment, except now the asset space is genuinely expanded to include all the European options on the stock of all strikes and maturities as well. They study the problem of optimal derivative investment and solve it in closed form for HARA utility when the statistical and risk neutral price processes are in the VG class of processes. They also show that the shape of the optimal financial derivative product is independent of preferences, time horizons and the mean rate of return on the stock, factors that influence the level of investor demand but not the shape. The latter depends primarily on the comparison between the prices of market moves and the relative frequency of their occurrence. Their analysis also suggests that demand would be highest for at-the-money low maturity options in such economies, a fact that is in accord with casual market observations.

7.1 Optimal derivative investment

Consider an economy trading a stock with price process $S(t)$ that is a homogeneous Lévy process in the interval $[0, \Upsilon]$ with a Lévy density $k_P(x)$ defined over the real line where x represents the jumps in the log of the stock price. An example is provided by the VG process of equation (11). Also trading in the economy are options on this stock with strikes $K > 0$ and maturities $T < \Upsilon$. The prices of these options are given by the processes $c(S(t); K, T)$ for $t < T$ where these prices are consistent with the absence of arbitrage and are derived in line with martingale pricing methods using the risk neutral measure that is also a homogeneous Lévy process with Lévy density $k_Q(x)$. The subscripts P and Q make the important distinction between the statistical price process and the risk neutral process, with the former assessing the relative frequency of events while the latter assesses their prices.

In such an economy we wish to study the question of optimal derivative investment. At first glance, and in analogy with the solution methods adopted in Merton (1971) this is a particularly difficult problem that is not going to be tractable from an analytical perspective. This is because we ask for the optimal positions in a doubly indexed continuum of assets, *viz.* the options of all strikes $K > 0$ and maturities $T > t$ in a context in which many of these options (i.e. those with maturities below t) are expiring on us. Furthermore, the analytical pricing of these options is generally a complex exercise reflecting all the difficulties associated with the kinked option payoff.

For reasons of tractability, we reformulate the problem with the focus on the real uncertainty which is the jump in log price of the stock, x . We view investment, not as a decision on what assets to hold, but in the first instance as a design problem where the investor wishes to design the optimal response of his or her wealth to market moves represented by x . Hence we seek to determine the optimal wealth response function $w(x, u)$ which is the jump in the investor's log wealth if the market were to jump at time u by the amount x in the log price of the stock. The actual investment in options that delivers this optimal wealth response is a secondary problem that may be solved numerically using the spanning properties of options. The structure and solution of this secondary problem is described in further detail in Carr, Jin and Madan (2000).

From the perspective of the optimal design of wealth responses, the optimal derivative investment problem may be formulated as a Markov control problem. Carr, Jin and Madan (2000) consider both the infinite time horizon problem with intermediate consumption and the finite horizon problem with no intermediate consumption. Here we present just the former. We denote by $c(t)$ the path of the flow rate of consumption per unit time and suppose the investor has a preference ordering over consumption paths represented by expected utility evaluated as

$$u = E^P \left[\int_0^\infty \exp(-\beta s) U(c(s)) ds \right] \quad (15)$$

where P is the statistical probability measure, β is the pure rate of time preference, and $U(c)$ is the instantaneous utility function. The investor wishes to choose the consumption path $c(\cdot)$ and the wealth response design $w(\cdot)$ with a view to maximizing u .

The investor is constrained by his budget constraint that describes the evolution of his wealth. The wealth, $W(t)$, transition equation is the integral equation

$$\begin{aligned} W(t) = & W(0) + \int_0^t r W(s_-) ds - \int_0^t c(s) ds \\ & + \int_0^t \int_{-\infty}^\infty W(s_-) (e^{w(x,u)} - 1) (m(\omega; dx, ds) - k_Q(x) dx ds), \end{aligned} \quad (16)$$

and the budget constraint requires that the wealth process be non-negative, $W(t) \geq 0$ almost surely. The first two terms of the wealth transition are standard and require no explanation, accounting for interest earnings and the financing of the consumption stream. The final term involves integration with respect to two measures, the first is the integer valued random measure $m(\omega; dx, ds)$ that is a Dirac delta measure counting the jumps that occur at various times of various sizes. The second is the pricing Lévy measure $k_Q(x) dx ds$. The integration with respect to m accounts for the wealth changes actually experienced by the response design

$w(x, u)$. The integration with respect to $k_Q(x)dxds$ accounts for the cost of this wealth response access that must be paid for through time.

The wealth transition equation (16) may be rewritten in a form more directly comparable to Merton's original equation by writing

$$\begin{aligned} W(t) = & W(0) + \int_0^t r W(s_-)ds - \int_0^t c(s)ds \\ & + \int_0^t \int_{-\infty}^{\infty} W(s_-) (e^{w(x,u)} - 1) (k_P(x)dxds - k_Q(x)dxds) \\ & + \int_0^t \int_{-\infty}^{\infty} W(s_-) (e^{w(x,u)} - 1) (m(\omega; dx, ds) - k_P(x)dxds) \end{aligned} \quad (17)$$

where we have just added and subtracted the integral of the wealth change with respect to the measure $k_P(x)dxds$. In this formulation the final integral in equation (17) is a martingale under the statistical measure P and matches the term representing the martingale component of stock investment in Merton (1971). The first two terms are the same as in Merton (1971). The third term matches the term that evaluates excess returns from stock investment in Merton (1971). Here excess returns are the expected wealth change less the cost or price of this change whereas in Merton we have $\mu - r$.

The investor's optimal derivative investment problem is to choose $c(\cdot)$, $w(\cdot)$, with a view to maximizing the utility u of equation (15) subject to the budget constraint of equation (16).

7.2 Optimal design of wealth responses

Let $J(W)$ be the optimized expected utility when the initial wealth $W(0) = W$. It is shown in Carr, Jin and Madan (2000) that the optimal wealth response function for the infinite time horizon problem is homogeneous in time and satisfies the equation

$$\frac{J_W(We^{w(x)})}{J_W(W)} = \frac{k_Q(x)}{k_P(x)}. \quad (18)$$

This condition has an intuitive interpretation when it is rewritten as

$$\frac{J_W(We^{w(x)})k_P(x)}{k_Q(x)} = J_W(W)$$

which is that the expected marginal utility per initial dollar spent on cash in each state, x , is equalized across states. If this is not the case then $w(x)$ should be altered to move funds from states with a lower marginal utility to states with a higher marginal utility. Alternatively, the marginal rate of transformation in utility

between two states must equal the marginal rate of transformation in markets between the same two states.

The optimal wealth response $w(x)$, is then determined from equation (18), if we know the function $J(W)$ as

$$w(x) = J_W^{-1} \left(J_W(W) \frac{k_Q(x)}{k_P(x)} \right).$$

We learn from this representation that the optimal wealth response design is a possibly smooth function J_W^{-1} applied to the ratio of two finite variation, infinite arrival rate Lévy measures. Such Lévy measures are kinked by construction at zero where the arrival rate goes to infinity. It follows that one would expect to see this property inherited by $w(x)$. This has the implication that at a minimum, optimal wealth response design positions investors with different slopes of their desired wealths with respect to up and down market movements, from at-the-money. Equivalently, there is a demand for short maturity at-the-money options.

7.2.1 HARA VG financial products

In the special case when the statistical and risk neutral processes are in the VG class and the utility function $U(c)$ is in the HARA (hyperbolic absolute risk aversion) class of utility functions, the optimal derivative investment problem of section 7.1 is shown in Carr, Jin and Madan (2000) to have a closed form solution where $J(W)$ is also in the HARA class of utility functions. The kinks in optimal designs discussed generally in section 7.2 can now be explicitly computed for this case.

Specifically, suppose the statistical Lévy measure is symmetric and given by

$$k_P(x) = \frac{1}{\kappa |x|} \exp \left(-\sqrt{\frac{2}{\kappa}} \frac{|x|}{s} \right) \quad (19)$$

where κ is the volatility of the statistical gamma time change for a symmetric Brownian motion with volatility s . Further suppose that the risk neutral Lévy measure is as given by (9) and parameters σ , ν , and θ . Let the utility function be

$$U(c) = \frac{\gamma}{1-\gamma} \left(\frac{\alpha}{\gamma} c - A \right)^{1-\gamma}.$$

In this case, defining

$$\zeta = \frac{\theta}{\sigma^2}$$

$$\lambda = \frac{1}{s} \sqrt{\frac{2}{\kappa}} - \frac{1}{\sigma} \sqrt{\frac{2}{\nu} + \frac{\theta^2}{\sigma^2}}$$

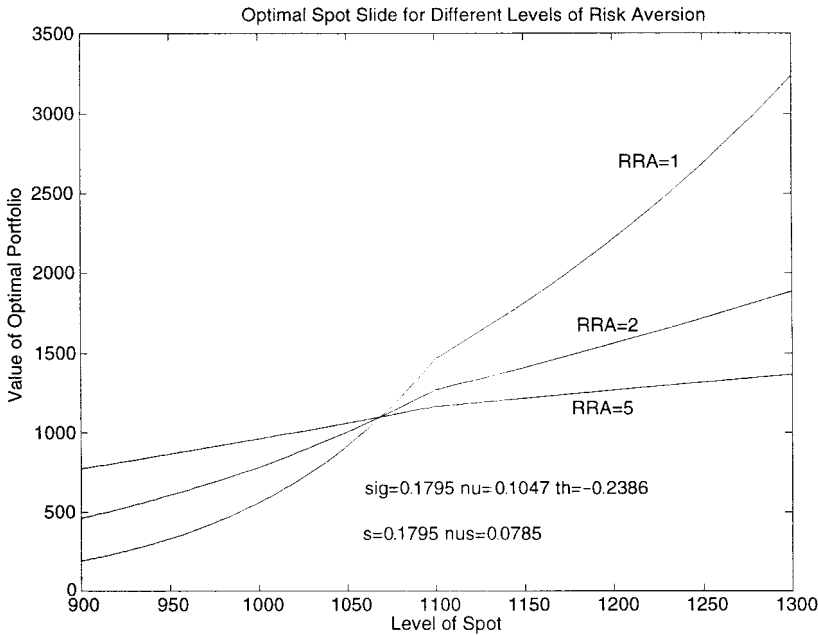


Fig. 4. Optimal spot slide for a strong skew and a mild excess kurtosis.

In Figure 3 the excess risk neutral kurtosis and skew leads to large moves being priced high relative to their likelihood and hence the optimal spot slide shorts these events and we have an inverted V shape for the spot slide.

For Figure 4 the skew is strong and the kurtosis is mild. This leads to falls being overpriced while rises are underpriced. The optimal slide is basically long the asset, but the positioning with respect to rises, the up delta, and falls, the down delta, differ.

For Figure 5 we have an excess statistical volatility making large moves relatively cheap securities. This gives rise to the V shaped optimal position.

Figure 6 is a reverse of the situation of Figure 4. The direction of the skew has been reversed and leads to a basically short position, with the kink induced by the behavior of the Lévy densities at the origin.

8 Spot slide calibration and position measures

The inputs for constructing an optimal spot slide are fairly simple and require just the specification of the statistical or time series moments of the return distribution, from which one may infer κ and s , the statistical Lévy measure parameters. The next step is to obtain data on market option prices, preferably for short

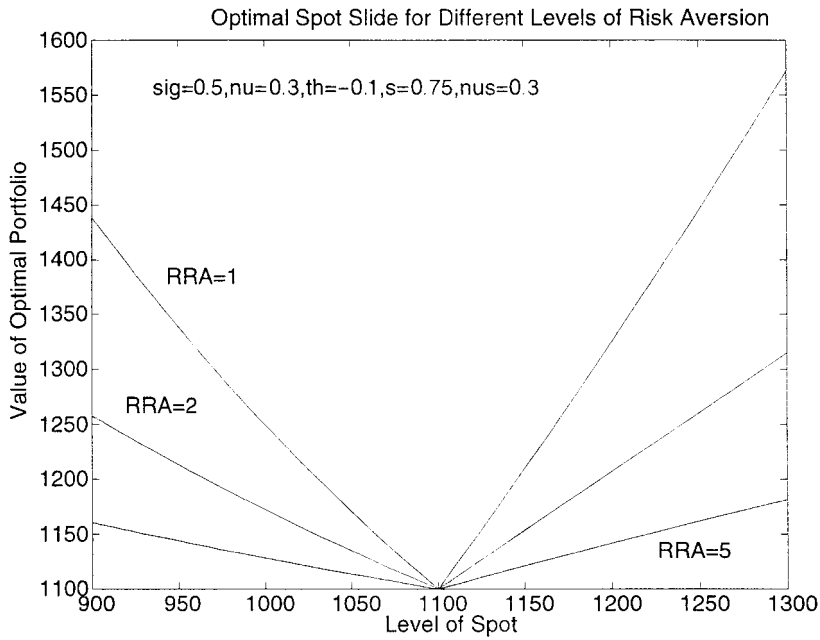


Fig. 5. Optimal spot slide when statistical volatility dominates risk neutral volatility

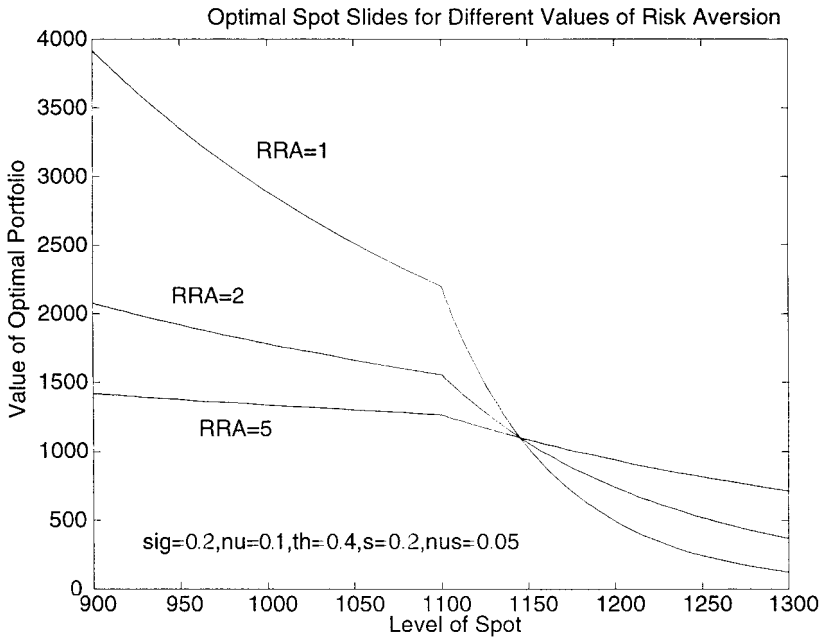


Fig. 6. Optimal spot slide for a positive skewness.

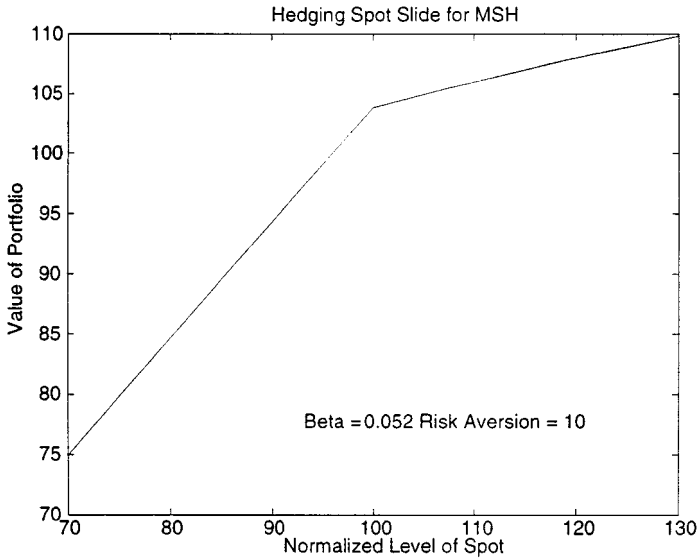


Fig. 7. Optimal spot slide as calibrated to a book of derivatives on an index.

maturity options and then to estimate the risk neutral Lévy measure and the three parameters σ , ν and θ . Finally, making some assumption on the coefficient of relative risk aversion in a power utility function gives us γ and we are ready to graph the optimal spot slide describing how one should currently be positioned in the derivatives markets.

For a contrast, one may compare with the actual spot slide that aggregates a trader's derivatives book and draws the response curve of his book value to market moves. We present here the results of calibrating optimal spot slides to data on actual spot slides. In the calibration we allowed for a reverse engineering of the coefficient of risk aversion γ as there is no other way to estimate this quantity. However, we also observed that the risk neutral excess kurtosis ν is typically an order of magnitude above its statistical counterpart κ and so we allowed this entity to be reverse engineered as well. Such an approach is defensible on noting that the variance of kurtosis estimates are of the order of the eighth moment and as the time series involved are not very long, generally two to four years, there is some leeway in an appropriate choice of this magnitude. The other parameters, σ , ν , θ , and s are taken at their estimated values.

For a variety of underlying assets and on a number of days, we reverse engineered the values of γ and κ so as to match the optimal spot slide with the actual spot slide observed for that day. Remarkably, we were able in many cases to come close to actual spot slides by just a simple choice of these two parameters (γ, κ) .

Figure 7 presents an example of an optimal spot slide as calibrated to an actual spot slide on a book of derivatives on a index. The ratio of κ to ν is referred to as β in the graph and describes the relative excess kurtosis of the subjective and risk neutral densities. Though it is often fairly small when calibrated, it is often an order of magnitude above the ratio of the statistical excess kurtosis to the risk neutral excess kurtosis.

Once all these parameters have been estimated and importantly γ and κ have been inferred from data on the actual spot slide, one may infer a personalized risk neutral density given by the subjective Lévy measure, determined by the parameters s and κ as described by equation (19), that is transformed by the marginal utility process as described in Madan and Milne (1991) to obtain the personalized risk neutral Lévy measure, $k_I(x)$ (the subscript I being indicative of an individualized measure)

$$k_I(x) = \exp(-\gamma x) \frac{1}{\kappa |x|} \exp\left(-\sqrt{\frac{2}{\kappa}} \frac{|x|}{s}\right). \quad (21)$$

The Lévy measure (21) is that of a VG process with personalized values for σ_I , ν_I , θ_I given by

$$\begin{aligned} \sigma_I &= \frac{s\sqrt{\frac{\kappa}{\nu}}}{\sqrt{1 - \frac{\gamma^2 s^2 \kappa}{2}}} \\ \theta_I &= -\gamma \frac{\kappa}{\nu} \frac{s^2}{1 - \frac{\gamma^2 s^2 \kappa}{2}} \\ \nu_I &= \kappa. \end{aligned} \quad (22)$$

We thus infer a personalized risk neutral process and this may be employed to construct a personalized return density that we term a *position measure*, as it is reverse engineered from derivative positions being viewed as optimal and therefore reflects preferences and beliefs that are obtained by a revealed preference exercise. All three densities are in the VG class of processes.

On completing this reverse engineering task we have available a statistical return density estimated from the time series of the return data, a risk neutral density as inferred from options data, and a position density as reverse engineered from the actual spot slide of the derivatives book. Figures 8, 9, 10 and 11 present a range of samples of graphs of these densities on a variety of underlying assets.

We observe a fairly diverse set of shapes of the densities, with varying degrees of skewness and kurtosis as reflected in the size of tails on the left and the right of the distribution. Furthermore, generally the position density is closer to the statistical density than the risk neutral density, reflecting the view that traders

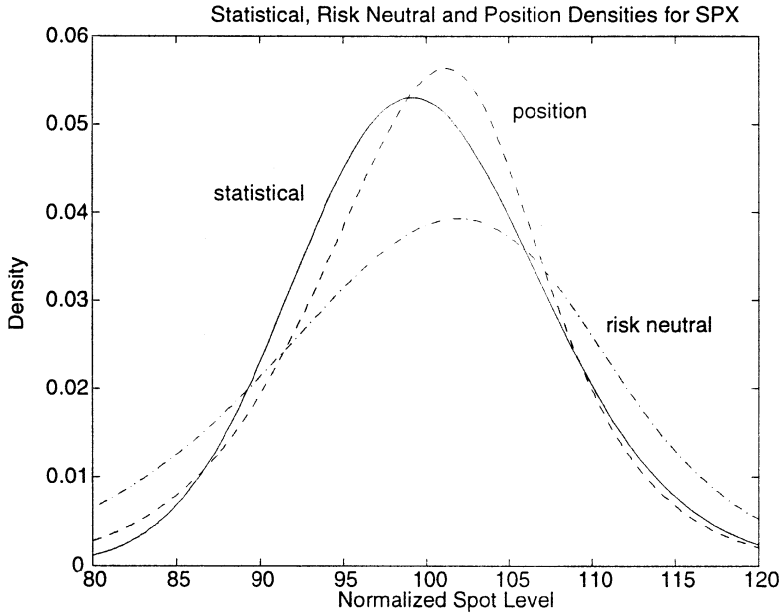


Fig. 8. Statistical, risk neutral and position densities for the SPX.

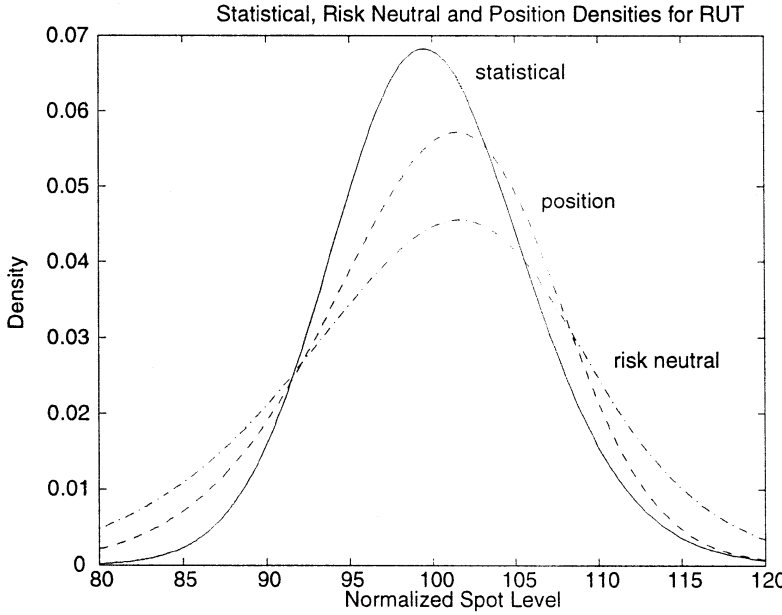


Fig. 9. Statistical, risk neutral and position densities for RUT.

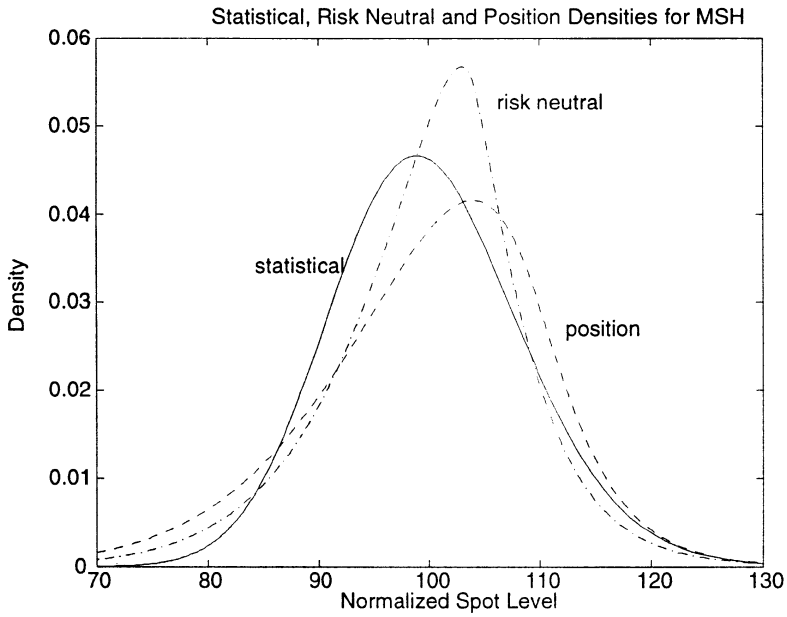


Fig. 10. Statistical, risk neutral and position densities for the MSH.

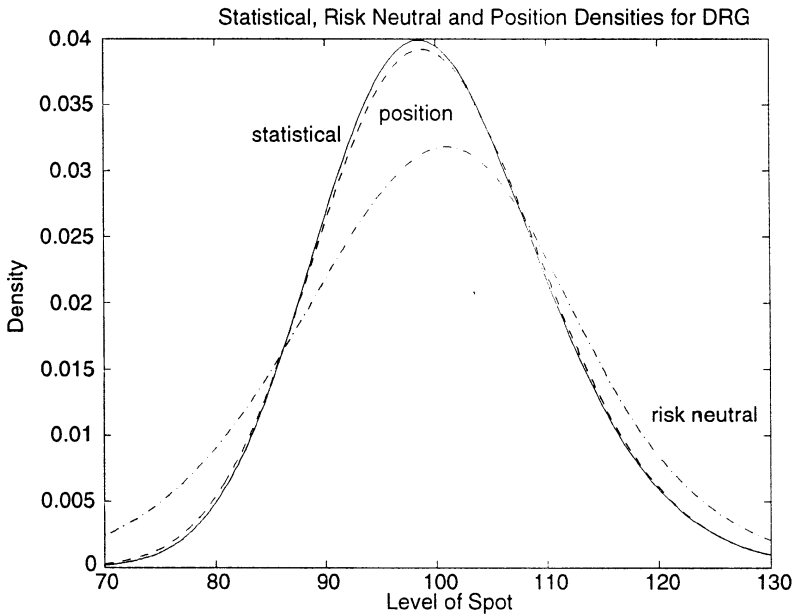


Fig. 11. Statistical, risk neutral and position densities for the DRG.

respect probability calculation as inferred from time series, and position themselves accordingly given the market prices of market moves as reflected in the risk neutral distribution. Occasionally, however, as in the case of Figure 9 the position density may be skewed further to the left than even the risk neutral density and is reflective of greater risk aversion on the part of the trader than is prevalent in the market.

9 Conclusion

We argue here that empirical evidence on the statistical and risk neutral price processes for financial assets belong to the class of purely discontinuous processes of finite variation, albeit ones of high activity, as reflected by an infinite arrival rate of jumps. Structurally, the pattern of jump arrival rates is consistent with the hypothesis of complete monotonicity whereby arrival rates at smaller size levels are higher.

Economic considerations of the absence of arbitrage point in the same direction by demonstrating that semimartingales, the candidate no arbitrage price process, is a time changed Brownian motion and the increasing random process of the time change is of necessity purely discontinuous, if it is not locally deterministic. The attribute of finite variation is attractive from two perspectives, one that allows a separation of the up and down tick modeling of the market, and we offer two representations of such price processes that are related under complete monotonicity of the Lévy density. The second attractive feature of finite variation is its robustness as reflected in its tolerance of parametric heterogeneity without the resulting measures being singular or disjoint in their sets of almost sure outcomes. This lack of robustness is an inherent property of infinite variation processes and we strongly advocate against the use of these processes as models for the price process unless there is overwhelming evidence in support of such a choice.

The class of stationary processes of independent and identically distributed increments meeting our requirements are characterized as a subclass of Lévy processes. Within this class, an important and analytically rich example is provided by Brownian motion time changed by a gamma process that combines in an interesting way two well studied processes in their own right. We summarize the properties of the resulting process termed the variance gamma process. The process has two additional parameters that enable it combat skew and kurtosis.

Option pricing under the variance gamma process is tractable using a variety of methods and we outline three such methods. The first is a closed form in terms of the modified Bessel function of the second kind and the degenerate hypergeometric function of two variables. The second involves two Fourier inversions for the complementary distribution function and the third employs direct Fourier inversion for the call price using the fast Fourier transform. The results of estimations are

illustrated for data on SPX and Nikkei Index options. It is observed that the model eliminates the smile in the strike direction, using effectively for this purpose its two additional parameters.

Infinite arrival rate, finite variation, Lévy processes with completely monotone Lévy densities are processes for the stock price for which options are market completing assets that are part of the primary assets of the economy with a genuine demand for these assets by investors. We study the Merton problem of optimal consumption and investment with the asset space expanded to include out-of-the-money European options as investment vehicles. For HARA utility and VG statistical and risk neutral processes this problem is solved in closed form with optimal portfolios that are kinked at-the-money and display a different slope with respect to upward and downward movements of the market. The positions reflect a role for at-the-money short maturity options, the most liquid end of the options market in practice.

Using our theory of optimal derivative positioning we illustrate how one may reverse engineer the preferences and beliefs of traders from observed spot slides of the derivatives book. This allows us to infer personalized risk neutral densities from observations on positions and we term this density the position density. Illustrations are provided, for comparative purposes of the statistical, risk neutral and position densities. It is observed that position densities are generally closer to the statistical density and lie between the statistical and risk neutral densities. At times however, they may be more skewed than the risk neutral density reflecting risk aversion that dominates market risk aversion.

Acknowledgment

I would like to thank all my co-authors for all the hard work on the various aspects of this project. They are in approximate chronological order, Eugene Seneta, Frank Milne, Eric Chang, Peter Carr, Helyette Geman, Marc Yor and Gurdip Bakshi. The support and encouragement offered by Claudia Albanese, Marco Avellanada, Joseph Cherian, Carl Chiarella, Jaksa Cvitanić, Nicole El Karoui, Hans Föllmer, Robert Jarrow, Yuri Kabanov, Ioannis Karatzas, Vadim Linetsky, Vincent Lacoste, Eckhardt Platen, Marc Pinsky, Stan Pliska, Phillip Protter, Raymond Rishel, Martin Schweizer, Steve Shreve, Meté Soner, and Thaleia Zariphopoulou is also greatly appreciated. Finally I would like to acknowledge the assistance and guidance I have received from my co-workers at Morgan Stanley Dean Witter, they are Doug Bonard, Steven Chung, Georges Courtadon, Peter Fraenkel, Santiago Garcia, George George, Kevin Holley, Ajay Khanna, Harry Mendell, and Lisa Polsky. Any remaining errors are solely my responsibility.

References

- Bakshi, G. and Chen, Z. (1997), An alternative valuation model for contingent claims, *Journal of Financial Economics* **44**, 123–65.
- Bakshi, G. and Madan, D.B. (2000), What is the probability of a stock market crash, Working Paper, University of Maryland.
- Bakshi, G. and Madan, D.B. (1998), Spanning and derivative security valuation, *Journal of Financial Economics* **55**, 205–38.
- Bates, D. (1996), Jumps and stochastic volatility: exchange rate processes implicit in Deutschmark options, *The Review of Financial Studies* **9**, 69–108.
- Bertoin, J. (1996), *Lévy Processes*, Cambridge University Press, Cambridge.
- Breedeen, D. and Litzenberger, R. (1978), Prices of state contingent claims implicit in option prices, *Journal of Business* **51**, 621–51.
- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–54.
- Carr, P., Geman, H., Madan, D.B. and Yor, M. (2000), The fine structure of asset returns: an empirical investigation, forthcoming in the *Journal of Business*.
- Carr, P., Jin, X. and Madan, D.B. (2000), Optimal investment in derivative securities, forthcoming in *Finance and Stochastics*.
- Carr, P. and Madan, D.B. (1999), Option valuation using the fast Fourier transform, *Journal of Computational Finance* **4**, 61–73.
- Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985), A theory of the term structure of interest rates, *Econometrica* **53**, 385–408.
- Cox, J. and Ross, S.A. (1976), The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–66.
- Das, S. and Foresi, S. (1996), Exact solutions for bond and options prices with systematic jump risk, *Review of Derivatives Research* **1**, 7–24.
- Delbaen, F. and Schachermayer, W. (1994), A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 520–63.
- Derman, E. and Kani, I. (1994), Riding on a smile, *Risk* **7**, 32–9.
- Dupire, B. (1994), Pricing with a smile, *Risk* **7**, 18–20.
- Embrechts, P. Kluppelberg, C. and Mikosch, T. (1997), *Modeling Extremal Events*, Springer-Verlag, Berlin.
- Fama, E.F. (1965), The behavior of stock market prices, *Journal of Business* **38**, 34–105.
- Feller, W.E. (1971), *An Introduction to Probability Theory and its Applications*, 2nd edition, Wiley, New York.
- Geman, H., Madan, D.B. and Yor, M. (2000), Time changes for Lévy processes, forthcoming in *Mathematical Finance*.
- Harrison, J.M. and Kreps, D. (1979), Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- Harrison, J.M. and Pliska, S.R. (1981), Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–60.
- Heston, S.L. (1993), A closed-form solution for options with stochastic volatility with applications to bond and currency options, *The Review of Financial Studies* **6**, 327–43.
- Hull, J. and White, A. (1987), The pricing of options on assets with stochastic volatility, *Journal of Finance* **42**, 281–300.
- Humbert, P. (1920), The confluent hypergeometric functions of two variables, *Proceedings of the Royal Society of Edinburgh* 73–85.
- Jacod, J. and Shiryaev, A. (1998), Local martingales and the fundamental asset pricing theorems in the discrete-time case, *Finance and Stochastics* **3**, 259–73.

- Jacod, J. and Shiryaev, A. (1980), *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.
- Jarrow, R.A. and Madan, D. (2000), Martingales and private monetary values, forthcoming in *Journal of Risk*.
- Kreps, D. (1981), Arbitrage and equilibrium in economies with infinitely many commodities, *Journal of Mathematical Economics* **8**, 15–35.
- Madan, D.B., Carr, P. and Chang, E. (1998), The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- Madan D.B. and Milne, F. (1991), Option pricing with VG martingale components, *Mathematical Finance* **1**, 39–55.
- Madan, D.B. and Seneta, E. (1989), Characteristic function estimation using maximum likelihood on transformed variables, *Journal of the Royal Statistical Society ser. B*, **51**, 281–5.
- Madan, D.B. and Seneta, E. (1990), The variance gamma (V.G.) model for share market returns, *Journal of Business* **63**, 511–24.
- Merton, R.C. (1971), Optimum consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* **3**, 373–413.
- Merton, R.C. (1973), Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–83.
- Merton, R.C. (1976), Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–44.
- Monroe, I. (1978), Processes that can be embedded in Brownian motion, *The Annals of Probability* **6**, 42–56.
- Naik, V. and Lee, M. (1990), General equilibrium pricing of options on the market portfolio with discontinuous returns, *The Review of Financial Studies* **3**, 493–522.
- Press, J.S. (1967), A compound events model for security prices, *Journal of Business* **40**, 317–35.
- Revuz, D. and Yor, M. (1994), *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin.
- Rogers, C. (1997), Arbitrage with fractional Brownian motion, *Mathematical Finance* **7**, 95–105.
- Ross, S.A. (1976a), Options and efficiency, *Quarterly Journal of Economics* **90**, 75–89.
- Ross, S.A. (1976b), Arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–60.

Latent Variable Models for Stochastic Discount Factors

René Garcia and Éric Renault

1 Introduction

Latent variable models in finance have traditionally been used in asset pricing theory and in time series analysis. In asset pricing models, a factor structure is imposed on a collection of asset returns to describe their joint distribution at a point in time, while in time series, the dynamic behavior of a series of multivariate returns depends on common factors for which a time series process is assumed. In both cases, the fundamental role of factors is to reduce the number of correlations between a large set of variables. In the first case, the dimension reduction is cross-sectional, in the second longitudinal. Factor analysis postulates that there exists a number of unobserved common factors or latent variables which explain observed correlations. To reduce dimension, a conditional independence is assumed between the observed variables given the common factors.

Arbitrage pricing theory (APT) is the standard financial model where returns of an infinite sequence of risky assets with a positive definite variance–covariance matrix are assumed to depend linearly on a set of common factors and on idiosyncratic residuals. Statistically, the returns are mutually independent given the factors. Economically, the idiosyncratic risk can be diversified away to arrive at an approximate linear beta pricing: the expected return of a risky asset in excess of a risk-free asset is equal to the scalar product of the vector of asset risks, as measured by the factor betas, with the corresponding vector of prices for the risk factors.

The latent GARCH factor model of Diebold and Nerlove (1989) best illustrates the type of time series model used to characterize the dynamic behavior of a set of financial returns. All returns are assumed to depend on a common latent factor and on noise. A longitudinal dimension reduction is achieved by assuming that the factor captures and subsumes the dynamic behavior of returns.¹ The imposed

¹ A cross-sectional dimension reduction is also achieved if the variance–covariance matrix of residuals is assumed to be diagonal.

statistical structure is a conditional absence of correlation between the factor and the noise terms, given the whole past of the factor and the noise, while the conditional variance of the factor follows a GARCH structure. This autoregressive conditional variance structure is important for financial applications such as portfolio allocations or value-at-risk calculations.

In this chapter, we aim at providing a unifying analysis of these two strands of literature through the concept of stochastic discount factor (SDF). The SDF (m_{t+1}), also called pricing kernel, discounts future payoffs p_{t+1} to determine the current price π_t of assets:

$$\pi_t = E[m_{t+1}p_{t+1}|J_t], \quad (1.1)$$

conditionally to the information set at time t , J_t . We summarize in Section 2 the mathematics of the SDF in a conditional setting according to Hansen and Richard (1987). Practical implementation of an asset pricing formula like (1.1) requires a statistical model to characterize the joint probability distribution of (m_{t+1}, p_{t+1}) given J_t . We specify in Section 3 a dynamic statistical framework to condition the discounted payoffs on a vector of state variables. Assumptions are made on the joint probability distribution of the SDF, asset payoffs and state variables to provide a state-space modeling framework which extends standard models.

Beta pricing relations amount to characterizing a vector space basis for the SDF through a limited number of factors. The coefficients of the SDF with respect to the factors are specified as deterministic functions of the state variables. Factor analysis and beta pricing with conditioning on state variables are reviewed in Section 4.

In dynamic asset pricing models, one can distinguish between reduced-form time-series models such as conditionally heteroskedastic factor models and asset pricing models based on equilibrium. We propose in Section 5 an intertemporal asset pricing model based on a conditioning on state variables which includes as a particular case stochastic volatility models. In this respect, we stress the importance of timing in conditioning to generate instantaneous correlation effects called leverage effects and show how it affects the pricing of stocks, bonds and European options. We make precise how this general model with latent variables relates to standard models such as CAPM for stocks and Black and Scholes (1973) or Hull and White (1987) for options.

2 Stochastic discount factors and conditioning information

Since Harrison and Kreps (1979) and Chamberlain and Rothschild (1983), it is well-known that, when asset markets are frictionless, portfolio prices can be characterized as a linear valuation functional that assigns prices to the portfolio payoffs.

Hansen and Richard (1987) analyze asset pricing functions in the presence of conditioning information. Their main contribution is to show that these pricing functions can be represented using random variables included in the collection of payoffs from portfolios. In this section we summarize the mathematics of a stochastic discount factor in a conditional setting following Hansen and Richard (1987). We focus on one-period securities as in their original analysis. In the next section, we will provide an extended framework with state variables to accommodate multiperiod securities.

We start with a probability space $(\Omega, \mathcal{A}, P)$. We denote the conditioning information as the information available to economic agents at date t by J_t , a sub-sigma algebra of $\mathcal{A}$. Agents form portfolios of assets based on this information, which includes in particular the prices of these assets. A one-period security purchased at time t has a payoff p at time $(t + 1)$. For such securities, an asset pricing model $\pi_t(\cdot)$ defines for the elements p of a set $P_{t+1} \subset J_{t+1}$ of payoffs a price $\pi_t(p) \in J_t$. The payoff space includes the payoffs of primitive assets, but investors can also create new payoffs by forming portfolios.

Assumption 2.1 (Portfolio formation)

$$p_1, p_2 \in P_{t+1} \implies w_1 p_1 + w_2 p_2 \in P_{t+1} \text{ for any variables } w_1, w_2 \in J_t.$$

Since we always maintain a finite-variance assumption for asset payoffs, P_{t+1} is, by virtue of Assumption 2.1, a pre-Hilbertian vectorial space included in:

$$P_{t+1}^+ = \{p \in J_{t+1}; E[p^2 | J_t] < +\infty\}$$

which is endowed with the conditional scalar product:

$$\langle p_1, p_2 \rangle_{J_t} = E[p_1 p_2 | J_t]. \quad (2.1)$$

The pricing functional $\pi_t(\cdot)$ is assumed to be linear on the vectorial space P_{t+1} of payoffs; this is basically the standard “law of one price” assumption, that is a very weak version of a condition of no-arbitrage.

Assumption 2.2 (Law of one price) For any p_1 and p_2 in P_{t+1} and any $w_1, w_2 \in J_t$:

$$\pi(w_1 p_1 + w_2 p_2) = w_1 \pi(p_1) + w_2 \pi(p_2).$$

The Hilbertian structure (2.1) will be used for orthogonal projections on the set P_{t+1} of admissible payoffs both in the proof of Theorem 2.3 below (a conditional version of the Riesz representation theorem) and in Section 4. Of course, this implies that we maintain an assumption of closedness for P_{t+1} . Indeed, Assumption 2.2 can be extended to an infinite series of payoffs to ensure not only a property of

closedness for P_{t+1} but also a continuity property for $\pi_t(\cdot)$ on P_{t+1} with appropriate notions of convergence for both prices and payoffs. With these assumptions and a technical condition ensuring the existence of a payoff with nonzero price to rule out trivial pricing functions, one can state the fundamental theorem of Hansen and Richard (1987), which is a conditional extension of the Riesz representation theorem.

Theorem 2.3 *There exists a unique payoff p^* in P_{t+1} that satisfies:*

- (i) $\pi_t(p) = E[p^*p|J_t]$ for all p in P_{t+1} ;
- (ii) $P[E[p^{*2}|J_t] > 0] = 1$.

In other words, the particular payoff which is used to characterize any asset price is almost surely nonzero. With an additional no-arbitrage condition, it can be shown to be almost surely positive.

3 Conditioning the discounted payoffs on state variables

We just stated that, given the law of one price, a pricing function $\pi_t(\cdot)$ for a conditional linear space P_{t+1} of payoffs can be represented by a particular payoff p^* such that condition (i) of Theorem 2.3 is fulfilled. In this section, we do not focus on the interpretation of the stochastic discount factor as a particular payoff. Instead, we consider a time series $(m_{t+1})_{t \geq 1}$ of *admissible* SDFs or pricing kernels, which means that, at each date t , m_{t+1} belongs to the set $\mathcal{M}_{t+1}$ defined as:

$$\mathcal{M}_{t+1} = \{m_{t+1} \in P_{t+1}^+; \pi_t(p_{t+1}) = E_t[m_{t+1}p_{t+1}|J_t], \forall p_{t+1} \in P_{t+1}\}. \quad (3.1)$$

For a given asset, we will write the asset pricing formula as:

$$\pi_t = E[m_{t+1}p_{t+1}|J_t]. \quad (3.2)$$

For the implementation of such a pricing formula, we need to model the joint probability distribution of (m_{t+1}, p_{t+1}) given J_t . To do this, we will stress the usefulness of factors and state variables. We will suppose without loss of generality² that the future payoff is the future price of the asset itself π_{t+1} . The problem is therefore to find the pricing function $\varphi_t(J_t)$ such that:

$$\psi_t(J_t) = E[m_{t+1}\psi_t(J_{t+1})|J_t]. \quad (3.3)$$

Both factors and state variables are useful to reduce the dimension of the problem to be solved in (3.3). To see this, one can decompose the information J_t into three types of variables. First, one can include asset-specific variables denoted Y_t , which

² As usual, if there are dividends or other cashflows, they may be included in the price by a convenient discounted sum. We will abandon this convenient expositional shortcut when we refer to more specific assets in subsequent sections.

should contain at least the price π_t . Dividends as well as other variables which may help characterize m_{t+1} could be included without really complicating matters. Second, the information will contain a vectorial process F_t of factors. Such factors could be suggested by economic theory or chosen purely on statistical grounds. For example, in equilibrium models, a factor could be the consumption growth process. In factor models, they could be observable macroeconomic indicators or latent factors to be extracted from a universe of asset returns. In both cases these variables are viewed as explanatory factors, possibly latent, of the collection of asset prices at time t . The purpose of these factors is to reduce the cross-sectional dimension of the collection of assets. Third, it is worthwhile to introduce a vectorial process U_t of exogenous state variables in order to achieve a longitudinal reduction of dimension.

Two assumptions are made about the conditional probability distribution of $(Y_t, F_t)_{1 \leq t \leq T}$ knowing $U_1^T = (U_t)_{1 \leq t \leq T}$ (for any T -tuple $t = 1, \dots, T$ of dates of interest) to support the claim that the processes making up U_t summarize the dynamics of the processes (Y_t, F_t) . First we assume that the state variables subsume all temporal links between the variables of interest.

Assumption 3.1 *The pairs $(Y_t, F_t)_{1 \leq t \leq T}, t = 1, \dots, T$ are mutually independent knowing $U_1^T = (U_t)_{1 \leq t \leq T}$.*

According to the standard latent factor analysis terminology, Assumption 3.1. means that the TH variables $U_t \in \mathbb{R}^H, t = 1, \dots, T$ provide a complete system of factors to account for the relationships between the variables $(Y_t, F_t)_{1 \leq t \leq T}$ (see for example Bartholomew (1987), p. 5). In the original latent variable modeling of Burt (1941) and Spearman (1927) in the early part of the century to study human intelligence, Y_t represented an individual's score to the test number t of mental ability. The basic idea was that individual scores at various tests will become independent (with repeated observations on several human subjects) given a latent factor called general intelligence. In our modeling, t denotes a date. When, with only one observation of the path of $(Y_t, F_t), t = 1, \dots, T$, we assume that these variables become independent given some latent state variables, it is clear that we also have in mind a standard temporal structure which provides an empirical content to this assumption. A minimal structure to impose is the natural assumption that only past and present values $U_\tau, \tau = 1, 2, \dots, t$ of the state variables matter for characterizing the probability distribution of (Y_t, F_t) .

Assumption 3.2 *The conditional probability distribution of (Y_t, F_t) given $U_1^T = (U_t)_{1 \leq t \leq T}$ coincides, for any $t = 1, \dots, T$, with the conditional probability distribution given $U_1^t = (U_\tau)_{1 \leq \tau \leq t}$.*

Assumption 3.2. is the following conditional independence³ property assumption:

$$(Y_t, F_t) \perp\!\!\!\perp (U_{t+1}^T) | (U_1^t) \quad (3.4)$$

for any $t = 1, \dots, T$.

Property (3.4) coincides with the definition of noncausality by Sims (1972) insofar as Assumption 3.1. is maintained and means that (Y, F) do not cause U in the sense of Sims.⁴ If we are ready to assume that the joint probability distribution of all the variables of interest is defined by a density function ℓ , Assumptions 3.1 and 3.2 are summarized by:

$$\ell[(Y_t, F_t)_{1 \leq t \leq T} | U_1^T] = \prod_{t=1}^T \ell[(Y_t, F_t) | U_1^t]. \quad (3.5)$$

The framework defined by (3.5) is very general for state-space modeling and extends such standard models as parameter driven models described in Cox (1981), stochastic volatility models as well as the state-space time series models (see Harvey (1989)). Our vector U_t of state variables can also be seen as a hidden Markov chain, a popular tool in nonlinear econometrics to model regime switches introduced by Hamilton (1989).

The merit of Assumptions 3.1 and 3.2 for asset pricing is to summarize the relevant conditioning information by the set U_1^t of current and past values of the state variables,

$$\ell[(Y_{t+1}, F_{t+1}, U_{t+1}) | (Y_\tau, F_\tau)_{1 \leq \tau \leq t} U_1^t] = \ell[(Y_{t+1}, F_{t+1}, U_{t+1}) | U_1^t]. \quad (3.6)$$

In practice, to make (3.6) useful, one would like to limit the relevant past by a homogeneous Markovianity assumption.

Assumption 3.3 *The conditional probability distribution of $(Y_{t+1}, F_{t+1}, U_{t+1})$ given U_1^t coincides, for any $t = 1, \dots, T$, with the conditional probability distribution given U_t . Moreover, this probability distribution does not depend on t .*

This assumption implies that the multivariate process U_t is homogeneous Markovian of order one.⁵

³ See Florens, Mouchart and Rollin (1990) for a systematic study of the concept of conditional independence and Florens and Mouchart (1982) for its relation with noncausality.

⁴ This noncausality concept is equivalent to the noncausality notion developed by Granger (1969). Assumption 3.2 can be equivalently replaced by an assumption stating that the state variables U can be optimally forecasted from their own past, with the knowledge of past values of other variables being useless (see Renault (1999)).

⁵ As usual, since the dimension of the multivariate process U_t is not limited *a priori*, the assumption of Markovianity of order one is not restrictive with respect to higher order Markov processes. For brevity, we will hereafter term Assumption 3.3 the assumption of Markovianity of the process U_t .

Given these assumptions, we are allowed to conclude that the pricing function, as characterized by (3.3), will involve the conditioning information only through the current value U_t of the state variables. Indeed, (3.6) can be rewritten:

$$\ell[(Y_{t+1}, F_{t+1}, U_{t+1})|(Y_\tau, F_\tau)_{1 \leq \tau \leq t} U_1^t] = \ell[(Y_{t+1}, F_{t+1}, U_{t+1})|U_t]. \quad (3.7)$$

We have seen how the dimension reduction is achieved in the longitudinal direction. To arrive at a similar reduction in the cross-sectional direction, one needs to add an assumption about the dimension of the range of m_{t+1} , given the state variables U_t . We assume that this range is spanned by K factors, $F_{kt+1}, k = 1, \dots, K$ given as components of the process F_{t+1} .

Assumption 3.4 (SDF spanning) m_{t+1} is a deterministic function of the variables U_t and F_{t+1} .

This assumption is not as restrictive as it might appear since it can be maintained when there exists an admissible SDF m_{t+1} with an unsystematic part $\varepsilon_{t+1} = m_{t+1} - E[m_{t+1}|F_{t+1}, U_t]$ that is uncorrelated, given U_t , with any feasible payoff $p_{t+1} \in P_{t+1}$. Actually, in this case, $\widehat{m}_{t+1} = E[m_{t+1}|F_{t+1}, U_t]$ is another admissible SDF since $E[m_{t+1}p_{t+1}|U_t] = E[\widehat{m}_{t+1}p_{t+1}|U_t]$ for any $p_{t+1} \in P_{t+1}$ and $\widehat{m}_{t+1}$ is by definition conformable to Assumption 3.4.

In Section 4 below, we will consider a linear SDF spanning, even if Assumption 3.4 allows for more general factor structures such as log-linear factor models of interest rates in Duffie and Kan (1996) and Dai and Singleton (1999) or nonlinear APT (see Bansal *et al.*, 1993). The linear benchmark is of interest when, for statistical or economic reasons, it appears useful to characterize the SDF as an element of a particular K -dimensional vector space, possibly time-varying through state variables. This is in contrast with nonlinear factor pricing where structural assumptions make a linear representation irrelevant for structural interpretations, even though it would remain mathematically correct.⁶ The linear case is of course relevant when the asset pricing model is based on a linear factor model for asset returns as in Ross (1976) as we will see in the next section.

4 Affine regression of payoffs on factors with conditioning on state variables

The longitudinal reduction of dimension through state variables put forward in Section 3 will be used jointly with the cross-sectional reduction of dimension through factors in the context of a conditional affine regression of payoffs or returns on factors. More precisely, the factor loadings, which are the regression coefficients on factors and which are often called beta coefficients, will be considered from

⁶ We will see in particular in Section 5 that a log-linear setting appears justified by a natural log-normal model of returns given state variables.

a conditional viewpoint, where the conditioning information set will be summarized by state variables given (3.7). We will first introduce the conditional beta coefficients and the corresponding conditional beta pricing formulas. We will then revisit the standard asset pricing theory which underpins these conditional beta pricing formulas, namely the arbitrage pricing theory of Ross (1976) stated in a conditional factor analysis setting.

4.1 Conditional beta coefficients

We first introduce conditional beta coefficients for payoffs, then for returns.

Definition 4.1 *The conditional affine regression $EL_t[P_{t+1}|F_{t+1}]$ of a payoff p_{t+1} on the vector F_{t+1} of factors given the information J_t is defined by:*

$$EL_t[p_{t+1}|F_{t+1}] = \beta_{0t} + \sum_{k=1}^K \beta_{kt} F_{kt+1} \quad (4.1)$$

with: $\varepsilon_{t+1} = p_{t+1} - EL_t[p_{t+1}|F_{t+1}]$ satisfying: $E[\varepsilon_{t+1}|J_t] = 0$, $\text{Cov}[\varepsilon_{t+1}, F_{t+1}|J_t] = 0$.

Similarly, if we denote by $r_{t+1} = p_{t+1}/\pi_t(p_{t+1})$ the return of an asset with a payoff⁷ p_{t+1} , we define the conditional affine regression of the return r_{t+1} on F_{t+1} by:

$$EL_t[r_{t+1}|F_{t+1}] = \beta_{0t}^r + \sum_{k=1}^K \beta_{kt}^r F_{kt+1}. \quad (4.2)$$

Of course, the beta coefficients of returns can be related to the beta coefficients of payoffs by:

$$\beta_{kt}^r = \frac{\beta_{kt}}{\pi_t(p_{t+1})} \quad \text{for } k = 0, 1, 2, \dots, K. \quad (4.3)$$

Moreover, the characterization of conditional probability distributions in terms of returns instead of payoffs makes more explicit the role of state variables. To see this, let us describe payoffs at time $t + 1$ from the price at the same date and a dividend process by:⁸

$$p_{t+1} = \pi_{t+1} + D_{t+1}. \quad (4.4)$$

⁷ Strictly speaking, the return is not defined for states of nature where $\pi_t(p_{t+1}) = 0$. This may complicate the statement of characterization of the SDF in terms of expected returns as in the main theorem (Theorem 4.4) of this section. However, this technical difficulty may be solved by considering portfolios which contain a particular asset with nonzero price in any state of nature. This technical condition ensuring the existence of such a payoff with nonzero price has already been mentioned in Section 2 (see also the sufficient condition 4.11 below when there exists a riskless asset). In what follows, the corresponding technicalities will be neglected.

⁸ As announced in Section 3, we depart from the expositional shortcut where the price included discounted dividends.

Following Assumption 3.1, we will assume that the rates of growth of dividends⁹ are asset-specific variables Y_t and serially uncorrelated given state variables. In other words, $Y_t = \frac{D_t}{D_{t-1}}$, $t = 1, 2, \dots, T$, are mutually independent given U_1^T . Moreover, π_{t+1} in (4.4) has to be interpreted as the price at time $(t + 1)$ of the same asset with price π_t at time t defined from the pricing functional (3.3). In other words, the pricing equation (3.3) can be rewritten:

$$\frac{\psi_t(J_t)}{D_t} = E \left[m_{t+1} \frac{D_{t+1}}{D_t} \left(\frac{\psi_t(J_{t+1})}{D_{t+1}} + 1 \right) | J_t \right]. \quad (4.5)$$

Given Assumptions 3.1, 3.2 and 3.3, we are allowed to conclude that, under general regularity conditions,¹⁰ Equation (4.5) defines a unique time-invariant deterministic function $\varphi(\cdot)$ such that:

$$\varphi(U_t) = E \left[m_{t+1} \frac{D_{t+1}}{D_t} (\varphi(U_{t+1}) + 1) | U_t \right]. \quad (4.6)$$

In other words, we get the following decomposition formulas for prices and returns:

$$\begin{aligned} \pi_t &= \varphi(U_t) D_t \\ r_{t+1} &= \frac{\pi_{t+1} + D_{t+1}}{\pi_t} = \frac{D_{t+1}}{D_t} \frac{\varphi(U_{t+1}) + 1}{\varphi(U_t)}. \end{aligned} \quad (4.7)$$

A by-product of this decomposition is that, by application of (3.7), the joint conditional probability distribution of future factors and returns $(F_\tau, r_\tau)_{\tau > t}$ given J_t depends upon J_t only through U_t in a homogeneous way. In particular, the conditional beta coefficients of returns are fixed deterministic functions of the current value of state variables:

$$\beta_{kt}^r = \beta_k^r(U_t) \quad \text{for } k = 0, 1, 2, \dots, K. \quad (4.8)$$

4.2 Conditional beta pricing

Since the seminal papers of Sharpe (1964) and Lintner (1965) on the unconditional CAPM to the most recent literature on conditional beta pricing (see e.g. Harvey (1991), Ferson and Korajczyk (1995)), beta coefficients with respect to well-chosen factors are put forward as convenient measures of compensated risk which explain the discrepancy between expected returns among a collection of financial assets. In order to document these traditional approaches in the modern setting of SDF, we have to add two fairly innocuous additional assumptions.

⁹ Stationarity (see Assumption 3.3) requires that we include the growth rates of dividends and not their levels in the variables Y_t .

¹⁰ These regularity conditions amount to the possibility of applying a contraction mapping argument to ensure the existence and unicity of a fixed point $\varphi(\cdot)$ of the functional defining the right hand side of (4.6).

Assumption 4.2 If $p_{F_{t+1}}$ denotes the orthogonal projection (for the conditional scalar product (2.1)) of the constant vector ι on the space P_{t+1} of feasible payoffs, the set $\mathcal{M}_{t+1}$ of admissible SDF does not contain a variable $\lambda_t p_{F_{t+1}}$ with $\lambda_t \in J_t$.

Assumption 4.3 Any admissible SDF has a nonzero conditional expectation given J_t .

Without Assumption 4.2, one could write for any $p_{t+1} \in P_{t+1}$:

$$\pi_t(p_{t+1}) = \lambda_t E[p_{F_{t+1}} p_{t+1} | J_t] = \lambda_t E[p_{t+1} | J_t]. \quad (4.9)$$

Therefore, all the feasible expected returns would coincide with $1/\lambda_t$. When there is a riskless asset, Assumption 4.2 simply means that an admissible SDF m_{t+1} should be genuinely stochastic at time t , that is not an element of the available information J_t at time t .

Without Assumption 4.3, one could write the price $\pi_t(p_{t+1})$ as:

$$\pi_t(p_{t+1}) = E[m_{t+1} p_{t+1} | J_t] = \text{Cov}[m_{t+1} p_{t+1} | J_t], \quad (4.10)$$

which would not depend on the expected payoff $E[p_{t+1} | J_t]$. When there is a riskless asset, Assumption 4.3 would be implied by a positivity requirement:¹¹

$$P[p > 0] = 1 \implies P[\pi_t(p) \leq 0] = 0. \quad (4.11)$$

With these two assumptions, we can state the central theorem of this section, which links linear SDF spanning with linear beta pricing and multibeta models of expected returns.

Theorem 4.4 *The three following properties are equivalent:*

P1: Linear Beta Pricing: $\exists m_{t+1} \in \mathcal{M}_{t+1}, \forall p_{t+1} \in P_{t+1}$:

$$\pi_t(p_{t+1}) = \beta_{0t} E[m_{t+1} | U_t] + \sum_{k=1}^K \beta_{kt} E[m_{t+1} F_{kt+1} | U_t], \quad (4.12)$$

P2: Linear SDF Spanning: $\exists m_{t+1} \in \mathcal{M}_{t+1}, \exists \lambda_{kt} \in J_t, k = 0, 1, 2, \dots, K$:

$$\lambda_{kt} = \lambda_k(U_t) \quad \text{and} \quad m_{t+1} = \lambda_0(U_t) + \sum_{k=1}^K \lambda_k(U_t) F_{kt+1}, \quad (4.13)$$

P3: Multibeta Model of Expected Returns: $\exists v_{kt} \in J_t, k = 0, 1, 2, \dots, K$, for any feasible return r_{t+1} :

$$E[r_{t+1} | U_t] = v_{0t} + \sum_{k=1}^K v_{kt} \beta_k^r(U_t). \quad (4.14)$$

¹¹ This positivity requirement implies the continuity of the pricing function $\pi_t(\cdot)$ needed for establishing Theorem 2.3.

Theorem 4.4 can be proved (see Renault, 1999) from three sets of assumptions: assumptions which ensure the existence of admissible SDFs (Section 2), assumptions about the state variables (Section 3), and technical Assumptions 4.2 and 4.3.

Three main lessons can be drawn from Theorem 4.4:

- (i) It makes explicit what we have called a cross-sectional reduction of dimension through factors, generally conceived to ensure SDF spanning, and more precisely linear SDF spanning, which corresponds to the specification (4.13) of the deterministic function referred to in Assumption 3.4. With a linear beta pricing formula, prices $\pi_t(p_{t+1})$ of a large cross-sectional collection of payoffs $p_{t+1} \in P_{t+1}$ can be computed from the prices of $K + 1$ particular “assets”:

$$\begin{aligned}\pi_t(\iota) &= E[m_{t+1}|J_t] = E[m_{t+1}|U_t] \\ \pi_t(F_{kt+1}) &= E[m_{t+1}F_{kt+1}|J_t] = E[m_{t+1}F_{kt+1}|U_t], \quad k = 1, 2, \dots, K.\end{aligned}\tag{4.15}$$

If there does not exist a riskless asset or if some factors are not feasible payoffs, one can always interpret suitably normalized factors as returns on particular portfolios called mimicking portfolios. Moreover, since the only property of factors which matters is linear SDF spanning, one may assume without loss of generality that $\text{Var}[F_{t+1}|U_t]$ is nonsingular to avoid redundant factors. The beta coefficients are then computed directly by:¹²

$$\begin{aligned}[\beta_{1t}, \beta_{2t}, \dots, \beta_{kt}] &= \text{Cov}[p_{t+1}, F_{t+1}|J_t] \text{Var}[F_{t+1}|U_t]^{-1} \\ \beta_{0t} &= E[p_{t+1}|J_t] - \sum_{k=1}^K \beta_{kt} E[F_{kt+1}|U_t]\end{aligned}\tag{4.16}$$

to deduce the price:

$$\pi_t(p_{t+1}) = \beta_{0t}\pi_t(\iota) + \sum_{k=1}^K \beta_{kt}\pi_t(F_{kt+1}).\tag{4.17}$$

The cross-sectional reduction of dimension consists of computing only $K + 1$ factor prices ($\pi_t(\iota), \pi_t(F_{kt+1})$) to price any payoff. The longitudinal reduction of dimension is also exploited since the pricing formula for these factors (4.15) depends on the conditioning information J_t only through U_t .

¹² When the payoffs include dividends, the only relevant conditioning information is characterized by state variables:

$$\begin{aligned}\text{Cov}[p_{t+1}, F_{t+1}|J_t] &= D_t \text{Cov}\left[\frac{p_{t+1}}{D_t}, F_{t+1}|U_t\right] \\ E[p_{t+1}|J_t] &= D_t E\left[\frac{p_{t+1}}{D_t}|U_t\right].\end{aligned}$$

- (ii) Even though the linear beta pricing formula P1 is mathematically equivalent to the linear SDF spanning property P2, it is interesting to characterize it by a property of the set of feasible returns under the maintained Assumption 2.4 of SDF spanning. More precisely, since this assumption allows us to write:

$$\pi_t(p_{t+1}) = E[m_{t+1}E[p_{t+1}|F_{t+1}, J_t]|J_t], \quad (4.18)$$

P1 is obtained as soon as a linear factor model of payoffs or returns is assumed (see e.g. Engle, Ng and Rothschild (1990)¹³). It means that the conditional expectation of payoffs given factors and J_t coincide with the conditional affine regression (given J_t) of these payoffs on these factors:

$$E[p_{t+1}|F_{t+1}, J_t] = EL_t[p_{t+1}|F_{t+1}] = \beta_{0t} + \sum_{k=1}^K \beta_{kt} F_{kt+1}. \quad (4.19)$$

Such a linear factor model can for instance be deduced from an assumption of joint conditional normality of returns and factors. This is the case when factors are themselves returns on some mimicking portfolios and returns are jointly conditionally gaussian. The standard CAPM illustrates the linear structure that is obtained from such a joint normality assumption for returns.

However, the main implication of linear beta pricing is the zero-price property of idiosyncratic risk (ε_{t+1} in the notation of Definition 4.1) since only the systematic part of the payoff p_{t+1} is compensated:¹⁴

$$\pi_t(p_{t+1}) = \pi_t(EL_t(p_{t+1}|F_{t+1})), \quad (4.20)$$

that is: $\pi_t(\varepsilon_{t+1}) = 0$. As we will see in more details in Subsection 4.3 below, this zero-price property for the idiosyncratic risk lays the basis for the APT model developed by Ross (1976). Moreover, if a factor is not compensated because $E[m_{t+1}F_{kt+1}|U_t] = 0$, it can be forgotten in the beta pricing formula. In other words, irrespective of the statistical procedure used to build the factors, only the compensated factors have to be kept:

$$\Delta_{kt} = E[m_{t+1}F_{kt+1}|U_t] \neq 0, \quad \text{for } k = 1, \dots, K. \quad (4.21)$$

- (iii) The minimal list of factors that have to be kept may also be characterized by the spanning interpretation P2. In this respect, the number of factors is purely a matter of convention: how many factors do we want to introduce to span the one-dimensional space where the SDF evolves? The existence of the SDF proves that a one-factor model with the SDF itself as

¹³ However, these authors maintain simultaneously the two assumptions of linear SDF spanning and linear factor model of returns. These two assumptions are clearly redundant as explained above.

¹⁴ The prices of the systematic and idiosyncratic parts are defined, by abuse of notation, by their conditional scalar product with the SDF m_{t+1} .

the sole factor is always correct. The definition of K factors becomes an issue for reasons such as economic interpretation, statistical procedures or financial strategies. Moreover, this definition can be changed as long as it keeps invariant the corresponding spanned vectorial space. For instance, one may assume that, conditionally to J_t , the factors are mutually uncorrelated, that is $V[F_{t+1}|J_t]$ is a nonsingular diagonal matrix. One may also rescale the factors to obtain unit variance factors (statistical motivation) or unit cost factors (financial motivation). Let us focus on the latter by assuming that:

$$\Delta_{kt} = E[m_{t+1}F_{kt+1}|U_t] = 1, \quad \text{for } k = 1, \dots, K. \quad (4.22)$$

By (4.21), the factor F_{kt+1} can be replaced by its scaled value F_{kt+1}/Δ_{kt} to get (4.22) without loss of generality. Each factor can then be interpreted as a return on a portfolio (a payoff of unit price) even though we do not assume that there exists a feasible mimicking portfolio ($F_{kt+1} \in P_{t+1}$). This normalization rule allows us to prove that the coefficients in the multibeta model of expected returns (P3) are given by:

$$v_{kt} = E[F_{kt+1}|U_t] - v_{0t} \quad \text{for } k = 1, \dots, K. \quad (4.23)$$

Since, on the other hand, it is easy to check that:

$$v_{0t} = \frac{1}{E[m_{t+1}|U_t]} \quad (4.24)$$

coincides with the risk-free return when there exists a risk-free asset, the multibeta model (P3) of expected returns can be rewritten in the more standard form:

$$E[r_{t+1}|U_t] - v_{0t} = \sum_{k=1}^K \beta_k^r(U_t)[E[F_{kt+1}|U_t] - v_{0t}], \quad (4.25)$$

which gives the risk premium of the asset as a linear combination of the risk premia of the various factors, with weights defined by the beta coefficients viewed as risk quantities. Moreover, (4.25) is very useful for statistical inference in factor models (see in particular Subsection 4.3) since it means that the beta pricing formula is characterized by the nullity of the intercept term in the conditional regression of net returns on net factors, given U_t .

4.3 Conditional factor analysis

Factor analysis with a cross-sectional point of view has been popularized by Ross (1976) to provide some foundations to multibeta models of expected returns. The basic idea is to start, for a countable sequence of assets $i = 1, 2, \dots$ with the

decomposition of their payoffs or returns into systematic and idiosyncratic parts with respect to K variables F_{kt+1} , $1, 2, \dots, K$, considered as candidate factors:

$$\begin{aligned} r_{it+1} &= \beta_{i0}^r(U_t) + \sum_{k=1}^K \beta_{ik}^r(U_t) F_{kt+1} + \varepsilon_{it+1} \\ E[\varepsilon_{it+1}|U_t] &= 0 \\ \text{Cov}[F_{kt+1}, \varepsilon_{it+1}|U_t] &= 0 \quad \forall k = 1, 2, \dots, K, \text{ for } i = 1, 2, \dots \end{aligned} \quad (4.26)$$

Since, as already explained, the multibeta model (P3) of expected returns amounts to assume that idiosyncratic risks are not compensated, that is:

$$E[m_{t+1}\varepsilon_{it+1}|U_t] = 0 \quad \text{for } i = 1, 2, \dots, \quad (4.27)$$

a natural way to look for foundations of this pricing model is to ask why idiosyncratic risk should not be compensated. Ross (1976) provides the following explanation. For a portfolio in the n assets defined by shares θ_{in} , $i = 1, 2, \dots, n$ of wealth invested:

$$\sum_{i=1}^n \theta_{in} = 1, \quad (4.28)$$

the unsystematic risk is measured by:

$$\text{Var}\left[\sum_{i=1}^n \theta_{in} \varepsilon_{it+1} | U_t\right] = \sum_{i=1}^n \theta_{in}^2 \sigma_i^2(U_t), \quad (4.29)$$

if we assume that the individual idiosyncratic risks are mutually uncorrelated:

$$\text{Cov}[\varepsilon_{it+1} \varepsilon_{jt+1} | U_t] = 0 \quad \text{if } i \neq j, \quad (4.30)$$

and we denote the asset idiosyncratic conditional variances by: $\sigma_i^2(U_t) = \text{Var}[\varepsilon_{it+1} | U_t]$.

Therefore, if it is possible to find a sequence $(\theta_{in})_{1 \leq i \leq n}$, $n = 1, 2, \dots$ conformable to (4.28) and (4.31) below:

$$P \lim_{n \rightarrow \infty} \sum_{i=1}^n \theta_{in}^2 \sigma_i^2(U_t) = 0, \quad (4.31)$$

the idiosyncratic risk can be diversified and should not be compensated by a simple no-arbitrage argument. Typically, this result will be valid with bounded conditional variances and equally-weighted portfolios ($\theta_{in} = 1/n$ for $i = 1, 2, \dots$).

In other words, according to Ross (1976), factors have as a basic property to define idiosyncratic risks which are mutually uncorrelated. This justifies beta pricing

with respect to them and provides the following decomposition of the conditional covariance matrix of returns:

$$\Sigma_t = \beta_t \phi_t \beta_t' + D_t \quad (4.32)$$

where Σ_t , β_t , ϕ_t , D_t are matrices of respective sizes $n \times n$, $n \times k$, $k \times k$ and $n \times n$ defined by:

$$\begin{aligned} \Sigma_t &= (\text{Cov}(r_{it+1}, r_{jt+1} | U_t))_{1 \leq i \leq n, 1 \leq j \leq n} \\ \beta_t &= (\beta_{ik}^r(U_t))_{1 \leq i \leq n, 1 \leq k \leq K} \\ \phi_t &= (\text{Cov}(F_{kt+1}, F_{lt+1} | U_t))_{1 \leq k \leq K, 1 \leq l \leq K} \\ D_t &= (\text{Cov}(\varepsilon_{it+1}, \varepsilon_{jt+1} | U_t))_{1 \leq i \leq n, 1 \leq j \leq n} \end{aligned} \quad (4.33)$$

with the maintained assumption that D_t is a diagonal matrix.

In the particular case where returns and factors are jointly conditionally gaussian given U_t , the returns are mutually independent knowing the factors in the conditional probability distribution given U_t . We have therefore specified a Factor Analysis model in a conditional setting. Moreover, if one adopts in such a setting some well-known results in the Factor Analysis methodology, one can claim that the model is fully defined by the decomposition (4.32) of the covariance matrix of returns with the diagonality assumption¹⁵ about the idiosyncratic variance matrix D_t . In particular, this decomposition defines by itself the set of K -dimensional variables F_{t+1} conformable to it with the interpretation (4.33) of the matrices:

$$F_{t+1} = E[F_{t+1} | U_t] + \phi_t \beta_t' \Sigma_t^{-1} (r_{t+1} - E[r_{t+1} | U_t]) + z_{t+1}, \quad (4.34)$$

where $r_{t+1} = (r_{it+1})_{1 \leq i \leq n}$ and z_{t+1} is a K -dimensional variable assumed to be independent of r_{t+1} given J_t and such that:

$$\begin{aligned} E[z_{t+1} | J_t] &= 0 \\ \text{Var}[z_{t+1} | J_t] &= \phi_t - \phi_t \beta_t' \Sigma_t^{-1} \beta_t \phi_t. \end{aligned} \quad (4.35)$$

It means that, up to an independent noise z_t (which represents factor indeterminacy), the factors are rebuilt by the so-called ‘‘Thompson Factor scores’’:

$$\widehat{F}_{t,t+1} = E[F_{t+1} | U_t] + \phi_t \beta_t' \Sigma_t^{-1} (r_{t+1} - E(r_{t+1} | U_t)), \quad (4.36)$$

which correspond to the conditional expectation: $\widehat{F}_{t,t+1} = E[F_{t+1} | U_t, r_{t+1}]$ in the particular case where returns and factors are jointly gaussian given U_t .

To summarize, according to Ross (1976) adapted in a conditional setting with latent variables, the question of specifying a multibeta model of expected returns

¹⁵ Chamberlain and Rothschild (1983) have proposed to take advantage of the sequence model ($n \rightarrow \infty$) to weaken the diagonality assumption on D_t by defining an approximate factor structure. We consider here a factor structure for fixed n .

can be addressed in two steps. In a first step, one should identify a factor structure for the family of returns:

$$\begin{aligned}\Sigma_t &= \beta_t \phi_t \beta_t' + D_t, \\ D_t &\text{ diagonal.}\end{aligned}\tag{4.37}$$

In a second step, the issue of a multibeta model for expected returns is addressed:¹⁶

$$E[r_{t+1}|U_t] = \beta_t E[F_{t+1}|U_t].\tag{4.38}$$

Due to the difficulty of disentangling the dynamics of the beta coefficients in β_t from the one of the factors, both at first order $E[F_{t+1}|U_t]$ in (4.38) and at second order $\phi_t = \text{Var}[F_{t+1}|U_t]$ in (4.37), a common solution in the literature is to add the quite restrictive assumption that the matrix β_t of conditional factor loadings is deterministic and time invariant:

$$\beta_t = \beta \text{ for every } t.\tag{4.39}$$

It should be noticed that assumption (4.39) does not imply *per se* that conditional betas coincide with unconditional ones since unconditional betas are not unconditional expectations of conditional ones. However, since by (4.39):

$$r_{t+1} = E(r_{t+1}|U_t) - \beta E(F_{t+1}|U_t) + \beta F_{t+1} + \varepsilon_{t+1},\tag{4.40}$$

it can be seen that β will coincide with the matrix of unconditional betas if and only if:

$$\text{Cov}[E(r_{t+1}|U_t) - \beta E(F_{t+1}|U_t), F_{t+1}|U_t] = 0.\tag{4.41}$$

In particular, if the conditional multibeta model (4.38) of expected returns and the assumption (4.39) of constant conditional betas are maintained simultaneously, the unconditional multibeta model of expected returns can be deduced:

$$Er_{t+1} = \beta EF_{t+1}.\tag{4.42}$$

Moreover, this joint assumption guarantees that the conditional factor analytic model (4.40) can be identified by a standard procedure of static factor analysis since:

$$\text{Var}(\varepsilon_{t+1}) = E(\text{Var}(\varepsilon_{t+1}|U_t)) = E(D_t)\tag{4.43}$$

will be a diagonal matrix as D_t . This remark has been fully exploited by King, Sentana and Wadhwani (1994). However, a general inference methodology for the

¹⁶ According to the comments following Theorem 4.4, we assume that factors are suitably scaled in order to get the convenient interpretation for the coefficients of the multibeta model of expected returns. Such a scaling can be done without loss of generality since it does not modify the property (4.37). Moreover, in (4.38), returns and factors are implicitly considered in excess of the risk-free rate (net returns and factors).

conditional factor analytic model remains to be stated. First, the restrictive assumption of fixed conditional betas should be relaxed. Second, even with fixed betas, one would like to be able to identify the conditional factor analytic model (4.40) without maintaining the joint hypothesis (4.38) of a multibeta model of expected returns. In this latter case, a factor stochastic volatility approach (see e.g. Meddahi and Renault (1996) and Pitt and Shephard (1999)) should be well-suited. The narrow link between our general state variable setting and the nowadays widespread stochastic volatility model is discussed in the next section.

5 A dynamic asset pricing model with latent variables

In the last section, we analyzed the cross-sectional restrictions imposed by financial asset pricing theories in the context of factor models. While these factor models were conditioned on an information set, the emphasis was not put on the dynamic behavior of asset returns. In this section, we propose an intertemporal asset pricing model based on a conditioning on state variables. Using assumptions spelled out in Section 3, we will accommodate a rich intertemporal framework where the stochastic discount factor can represent nonseparable preferences such as recursive utility.¹⁷

5.1 An equilibrium asset pricing model with recursive utility

Many identical infinitely lived agents maximize their lifetime utility and receive each period an endowment of a single nonstorable good. We specify a recursive utility function of the form:

$$V_t = W(C_t, \mu_t), \quad (5.1)$$

where W is an aggregator function that combines current consumption C_t with $\mu_t = \mu(\tilde{V}_{t+1} | J_t)$, a certainty equivalent of random future utility $\tilde{V}_{t+1}$, given the information available to the agents at time t , to obtain the current-period lifetime utility V_t . Following Kreps and Porteus (1978), Epstein and Zin (1989) propose the CES function as the aggregator function, i.e.

$$V_t = [C_t^\rho + \beta \mu_t^\rho]^\frac{1}{\rho}. \quad (5.2)$$

The way the agents form the certainty equivalent of random future utility is based on their risk preferences, which are assumed to be isoelastic, i.e. $\mu_t^\alpha = E[\tilde{V}_{t+1}^\alpha | I_t]$,

¹⁷ In the proposed intertemporal asset pricing model, we will specify the stochastic discount factor in an equilibrium setting. We will therefore make our stochastic assumptions on economic fundamentals such as consumption and dividend growth rates. In Garcia, Luger and Renault (1999), we make the same types of assumptions directly on the pair SDF-stock returns without reference to an equilibrium model. Similar asset pricing formulas and implications of the presence of leverage effects are obtained in this less specific framework.

where $\alpha \leq 1$ is the risk aversion parameter ($1 - \alpha$ is the Arrow–Pratt measure of relative risk aversion). Given these preferences, the following Euler condition must be valid for any asset j if an agent maximizes his lifetime utility (see Epstein and Zin (1989)):

$$E \left[\beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma(\rho-1)} M_{t+1}^{\gamma-1} R_{j,t+1} | J_t \right] = 1, \quad (5.3)$$

where M_{t+1} represents the return on the market portfolio, $R_{j,t+1}$ the return on any asset j , and $\gamma = \frac{\rho}{\alpha}$. The stochastic discount factor is therefore given by:

$$m_{t+1} = \beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma(\rho-1)} M_{t+1}^{\gamma-1}. \quad (5.4)$$

The parameter ρ is associated with intertemporal substitution, since the elasticity of intertemporal substitution is $1/(1 - \rho)$. The position of α with respect to ρ determines whether the agent has a preference towards early resolution of uncertainty ($\alpha < \rho$) or late resolution of uncertainty ($\alpha > \rho$).¹⁸

Since the market portfolio price, say P_t^M at time t , is determined in equilibrium, it should also verify the first-order condition:

$$E \left[\beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma(\rho-1)} M_{t+1}^\gamma | J_t \right] = 1. \quad (5.5)$$

In this model, the payoff of the market portfolio at time t is the total endowment of the economy C_t . Therefore the return on the market portfolio M_{t+1} can be written as follows:

$$M_{t+1} = \frac{P_{t+1}^M + C_{t+1}}{P_t^M}.$$

Replacing M_{t+1} by this expression, we obtain:

$$\lambda_t^\gamma = E \left[\beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma\rho} (\lambda_{t+1} + 1)^\gamma | J_t \right], \quad (5.6)$$

where: $\lambda_t = P_t^M / C_t$. The pricing of assets with price S_t which pay dividends D_t such as stocks will lead us to characterize the joint probability distribution of the stochastic process (X_t, Y_t, J_t) where: $X_t = \log(C_t / C_{t-1})$ and $Y_t = \log(D_t / D_{t-1})$. As announced in Section 3, we define this dynamics through a stationary vector-process of state variables U_t so that:

$$J_t = \vee_{\tau \leq t} [X_\tau, Y_\tau, U_\tau]. \quad (5.7)$$

¹⁸ As mentioned in Epstein and Zin (1991), the association of risk aversion with α and intertemporal substitution with ρ is not fully clear, since at a given level α of risk aversion, changing ρ affects not only the elasticity of intertemporal substitution but also determines whether the agent will prefer early or late resolution of uncertainty.

Given this model structure (with $\log(C_t/C_{t-1})$ serving as a factor F_t), we can restate Assumptions 3.1 and 3.2 as:

Assumption 5.1 *The pairs $(X_t, Y_t)_{1 \leq t \leq T}$, $t = 1, \dots, T$ are mutually independent knowing $U_1^T = (U_t)_{1 \leq t \leq T}$.*

Assumption 5.2 *The conditional probability distribution of (X_t, Y_t) given $U_1^T = (U_t)_{1 \leq t \leq T}$ coincides, for any $t = 1, \dots, T$, with the conditional probability distribution given $U_1^t = (U_\tau)_{1 \leq \tau \leq t}$.*

As mentioned in Section 3, Assumptions 5.1 and 5.2 together with Assumption 3.3 and the Markovianity of state variables U_t allow us to characterize the joint probability distribution of the (X_t, Y_t) pairs, $t = 1, \dots, T$, given U_1^T , by:

$$\ell[(X_t, Y_t)_{1 \leq t \leq T} | U_1^T] = \prod_{t=1}^T \ell[X_t, Y_t | U_t]. \quad (5.8)$$

Proposition 5.3 below provides the exact relationship between the state variables and equilibrium prices.

Proposition 5.3 *Under Assumptions 5.1 and 5.2 we have:*

$$P_t^M = \lambda(U_t)C_t, \quad S_t = \varphi(U_t)D_t,$$

where $\lambda(U_t)$ and $\varphi(U_t)$ are respectively defined by:

$$\lambda(U_t)^\gamma = E \left[\beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma \rho} (\lambda(U_{t+1}) + 1)^\gamma | U_t \right],$$

and

$$\varphi(U_t) = E \left[\beta^\gamma \left(\frac{C_{t+1}}{C_t} \right)^{\gamma \rho - 1} \left(\frac{\lambda(U_{t+1}) + 1}{\lambda(U_t)} \right)^{\gamma - 1} (\varphi(U_{t+1}) + 1) \frac{D_{t+1}}{D_t} | U_t \right].$$

Therefore, the functions $\lambda(\cdot)$, $\varphi(\cdot)$ are defined on R^P if there are P state variables. Moreover, the stationarity property of the U process together with assumptions 5.1, 5.2 and a suitable specification of the density function (3.6) allow us to make the process (X, Y) stationary by a judicious choice of the initial distribution of (X, Y) . In this setting, a contraction mapping argument may be applied as in Lucas (1978) to characterize the functions $\lambda(\cdot)$ and $\varphi(\cdot)$ according to Proposition 5.3. It should be stressed that this framework is more general than the Lucas one because the state variables U_t are given by a general multivariate Markovian process (while a Markovian dividend process is the only state variable in Lucas

(1978)). Using the return definition for the market portfolio and asset S_t , we can write:

$$\log M_{t+1} = \log \frac{\lambda(U_{t+1}) + 1}{\lambda(U_t)} + X_{t+1}, \text{ and} \quad (5.9)$$

$$\log R_{t+1} = \log \frac{\varphi(U_{t+1}) + 1}{\varphi(U_t)} + Y_{t+1}.$$

Hence, the return processes (M_{t+1}, R_{t+1}) are stationary as U , X and Y , but, contrary to the stochastic setting in the Lucas (1978) economy, are not Markovian due to the presence of unobservable state variables U .

Given this intertemporal model with latent variables, we will show how standard asset pricing models will appear as particular cases under some specific configurations of the stochastic framework. In particular, we will analyze the pricing of bonds, stocks and options and show under which conditions the usual models such as the CAPM or the Black–Scholes model are obtained.

5.2 Revisiting asset pricing theories for bonds, stocks and options through the leverage effect

In this section, we introduce an additional assumption on the probability distribution of the fundamentals X and Y given the state variables U .

Assumption 5.4

$$\begin{pmatrix} X_{t+1} \\ Y_{t+1} \end{pmatrix} | U_t^{t+1} \sim \mathfrak{N} \left[\begin{pmatrix} m_{Xt+1} \\ m_{Yt+1} \end{pmatrix}, \begin{bmatrix} \sigma_{Xt+1}^2 & \sigma_{XYt+1} \\ \sigma_{XYt+1} & \sigma_{Yt+1}^2 \end{bmatrix} \right],$$

where $m_{Xt+1} = m_X(U_1^{t+1})$, $m_{Yt+1} = m_Y(U_1^{t+1})$, $\sigma_{Xt+1}^2 = \sigma_X^2(U_1^{t+1})$, $\sigma_{XYt+1} = \sigma_{XY}(U_1^{t+1})$, $\sigma_{Yt+1}^2 = \sigma_Y^2(U_1^{t+1})$. In other words, these mean and variance covariance functions are time-invariant and measurable functions with respect to U_t^{t+1} , which includes both U_t and U_{t+1} .

This conditional normality assumption allows for skewness and excess kurtosis in unconditional returns. It is also useful for recovering as a particular case the Black–Scholes formula.¹⁹

¹⁹ It can also be argued that, if one considers that the discrete-time interval is somewhat arbitrary and can be infinitely split, log-normality (conditional on state variables U) is obtained as a consequence of a standard central limit argument given the independence between consecutive (X, Y) given U .

5.2.1 The pricing of bonds

The price of a bond delivering one unit of the good at time T , $B(t, T)$, is given by the following formula:

$$B(t, T) = E_t[\tilde{B}(t, T)], \quad (5.10)$$

where:

$$\tilde{B}(t, T) = \beta^{\gamma(T-t)} a_t^T(\gamma) \exp((\alpha - 1) \sum_{\tau=t}^{T-1} m_{X_{\tau+1}} + \frac{1}{2}(\alpha - 1)^2 \sum_{\tau=t}^{T-1} \sigma^2_{X_{\tau+1}}),$$

with: $a_t^T(\gamma) = \prod_{\tau=t}^{T-1} \left[\frac{1+\lambda(U_1^{\tau+1})}{\lambda(U_1^\tau)} \right]^{\gamma-1}$.

This formula shows how the interest rate risk is compensated in equilibrium, and in particular how the term premium is related to preference parameters. To be more explicit about the relationship between the term premium and the preference parameters, let us first notice that we have a natural factorization:

$$\tilde{B}(t, T) = \prod_{\tau=t}^{T-1} \tilde{B}(\tau, \tau + 1). \quad (5.11)$$

Therefore, while the discount parameter β affects the level of the $\tilde{B}$, the two other parameters α and γ affect the term premium (with respect to the return-to-maturity expectations hypothesis, Cox, Ingersoll, and Ross (1981)) through the ratio:

$$\frac{B(t, T)}{E_t \prod_{\tau=t}^{T-1} B(\tau, \tau + 1)} = \frac{E_t(\prod_{\tau=t}^{T-1} \tilde{B}(\tau, \tau + 1))}{E_t \prod_{\tau=t}^{T-1} E_\tau \tilde{B}(\tau, \tau + 1)}.$$

To better understand this term premium from an economic point of view, let us compare implicit forward rates and expected spot rates at only one intermediary period between t and T :

$$\frac{B(t, T)}{B(t, \tau)} = \frac{E_t \tilde{B}(t, \tau) \tilde{B}(\tau, T)}{E_t \tilde{B}(t, \tau)} = E_t \tilde{B}(\tau, T) + \frac{\text{Cov}_t[\tilde{B}(t, \tau), \tilde{B}(\tau, T)]}{E_t \tilde{B}(t, \tau)}. \quad (5.12)$$

Up to Jensen inequality, Equation (5.12) proves that a positive term premium is brought about by a negative covariation between present and future $\tilde{B}$. Given the expression for $\tilde{B}(t, T)$ above, it can be seen that for von-Neuman preferences ($\gamma = 1$) the term premium is proportional to the square of the coefficient of relative risk aversion (up to a conditional stochastic volatility effect). Another important observation is that even without any risk aversion ($\alpha = 1$), preferences still affect the term premium through the nonindifference to the timing of uncertainty resolution ($\gamma \neq 1$).

There is however an important sub-case where the term premium will be preference-free because the stochastic discount factor $\tilde{B}(t, T)$ coincides with the

observed rolling-over discount factor (the product of short-term future bond prices, $B(\tau, \tau+1)$, $\tau = t, \dots, T-1$). Taking Equation (5.11) into account, this will occur as soon as $\tilde{B}(\tau, \tau+1) = B(\tau, \tau+1)$, that is when $\tilde{B}(\tau, \tau+1)$ is known at time τ . From the expression of $\tilde{B}(t, T)$ above, it is easy to see that this last property stands if and only if the mean and variance parameters $m_{X_{\tau+1}}$ and $\sigma_{X_{\tau+1}}$ depend on $U_{\tau}^{\tau+1}$ only through U_{τ} .

This allows us to highlight the so-called “leverage effect” which appears when the probability distribution of (X_{t+1}) given U_t^{t+1} depends (through the functions m_X, σ_X^2) on the contemporaneous value U_{t+1} of the state process. Otherwise, the noncausality Assumption 5.2 can be reinforced by assuming no instantaneous causality from X to U .

In this case, $\ell(X_t|U_1^T) = \ell(X_t|U_1^{t-1})$; it is this property which ensures that short-term stochastic discount factors are predetermined, so the bond pricing formula becomes preference-free:

$$B(t, T) = E_t \prod_{\tau=t}^{T-1} B(\tau, \tau+1).$$

Of course this does not necessarily cancel the term premiums but it makes them preference-free in the sense that the role of preference parameters is fully hidden in short-term bond prices. Moreover, when there is no interest rate risk because the consumption growth rates X_t are serially independent, it is straightforward to check that constant $m_{X_{t+1}}$ and $\sigma_{X_{t+1}}^2$ imply constant $\lambda(\cdot)$ and in turn $\tilde{B}(t, T) = B(t, T)$, with zero term premiums.

5.2.2 The pricing of stocks

The stock price formula is given by:

$$S_t = E_t \left[\beta^{\gamma} \left(\frac{C_{t+1}}{C_t} \right)^{\alpha-1} \left[\frac{1 + \lambda(U_1^{t+1})}{\lambda(U_1^t)} \right]^{\gamma-1} (S_{t+1} + D_{t+1}) \right].$$

By a recursive argument, this Euler condition can be rewritten as follows:

$$E_t \left[\beta^{\gamma(T-t)} a_t^T(\gamma) b_t^T \left(\frac{C_T}{C_t} \right)^{\alpha-1} \frac{D_T}{D_t} \right] = 1, \quad (5.13)$$

with: $b_t^T = \prod_{\tau=t}^{T-1} (1 + \varphi(U_1^{\tau+1}))/\varphi(U_1^{\tau})$.

Under conditional log-normality Assumption 5.4, we obtain:

$$E_t \left[\tilde{B}(t, T) b_t^T \exp \left(\sum_{\tau=t+1}^T m_{Y\tau} + \frac{1}{2} \sum_{\tau=t+1}^T \sigma_{Y\tau}^2 + (\alpha - 1) \sum_{\tau=t+1}^T \sigma_{XY\tau} \right) \right] = 1. \quad (5.14)$$

With the definitional equation:

$$E \left[\frac{S_T}{S_t} | U_1^T \right] = \frac{\varphi(U_1^T)}{\varphi(U_1^t)} \exp \left(\sum_{\tau=t+1}^T m_{Y\tau} + \frac{1}{2} \sum_{\tau=t+1}^T \sigma_{Y\tau}^2 \right), \quad (5.15)$$

a useful way of writing the stock pricing formula is:

$$E_t [Q_{XY}(t, T)] = 1, \quad (5.16)$$

where:

$$Q_{XY}(t, T) = \tilde{B}(t, T) b_t^T \frac{\varphi(U_1^t)}{\varphi(U_1^T)} \exp \left((\alpha - 1) \sum_{\tau=t+1}^T \sigma_{XY\tau} \right) E \left[\frac{S_T}{S_t} | U_1^T \right]. \quad (5.17)$$

To understand the role of the factor $Q_{XY}(t, T)$, it is useful to notice that it can be factorized:

$$Q_{XY}(t, T) = \prod_{\tau=t}^{T-1} Q_{XY}(\tau, \tau + 1),$$

and that there is an important particular case where $Q_{XY}(\tau, \tau + 1)$ is known at time τ and therefore equal to one by (5.16). This is when there is no leverage effect in the sense that $\ell(X_t, Y_t | U_1^T) = \ell(X_t, Y_t | U_1^{t-1})$. This means that not only there is no leverage effect neither for X nor for Y , but also that the instantaneous covariance σ_{XY_t} itself does not depend on U_t . In this case, we have $Q_{XY}(t, T) = 1$. Since we also have $\tilde{B}(\tau, \tau + 1) = B(\tau, \tau + 1)$, we can express the conditional expected stock return as:

$$E \left[\frac{S_T}{S_t} | U_1^T \right] = \frac{1}{\prod_{\tau=t}^{T-1} B(\tau, \tau + 1)} \frac{1}{b_t^T} \frac{\varphi(U_1^T)}{\varphi(U_1^t)} \exp \left((1 - \alpha) \sum_{\tau=t+1}^T \sigma_{XY\tau} \right).$$

For pricing over one period (t to $t+1$), this formula provides the agent's expectation of the next period return (since in this case the only relevant information is U_1^t):

$$E \left[\frac{S_{t+1}}{S_t} \frac{1 + \varphi(U_1^{t+1})}{\varphi(U_1^{t+1})} | U_1^t \right] = \frac{1}{B(t, t+1)} \exp[(1 - \alpha)\sigma_{XY_{t+1}}],$$

that is:

$$E \left[\frac{S_{t+1} + D_{t+1}}{S_t} | U_1^t \right] = \frac{1}{B(t, t+1)} \exp[(1 - \alpha)\sigma_{XY_{t+1}}], \quad (5.18)$$

This is a particularly striking result since it is very close to a standard conditional CAPM equation, which remains true for any value of the preference parameters α and ρ . While Epstein and Zin (1991) emphasize that the CAPM obtains for $\alpha = 0$ (logarithmic utility) or $\rho = 1$ (infinite elasticity of intertemporal substitution), we stress here that the relation is obtained under a particular stochastic setting for any

values of α and ρ . Remarkably, the stochastic setting without leverage effect which produces this CAPM relationship will also produce most standard option pricing models (for example Black and Scholes (1973) and Hull and White (1987)), which are of course preference-free.²⁰

5.2.3 A generalized option pricing formula

The Euler condition for the price of a European option is given by:

$$\pi_t = E_t \left[\beta^{\gamma(T-t)} \left(\frac{C_T}{C_t} \right)^{\alpha-1} \prod_{\tau=t}^{T-1} \left[\frac{1 + \lambda(U_1^{\tau+1})}{\lambda(U_1^\tau)} \right]^{\gamma-1} \text{Max}[0, S_T - K] \right]. \quad (5.19)$$

It is worth noting that the option pricing formula (5.19) is path-dependent with respect to the state variables; it depends not only on the initial and terminal values of the process U_t but also on its intermediate values.²¹ Indeed, it is not so surprising that when preferences are not time-separable ($\gamma \neq 1$), the option price may depend on the whole past of the state variables.

Using Assumptions 5.2, 5.2 and 5.4, we arrive at an extended Black–Scholes formula:

$$\frac{\pi_t}{S_t} = E_t \left\{ Q_{XY}^*(t, T) \Phi(d_1) - \frac{K \tilde{B}(t, T)}{S_t} \Phi(d_2) \right\}, \quad (5.20)$$

where:

$$d_1 = \frac{\log \left[\frac{S_t Q_{XY}^*(t, T)}{K \tilde{B}(t, T)} \right]}{(\sum_{\tau=t+1}^T \sigma_{Y\tau}^2)^{1/2}} + \frac{1}{2} \left(\sum_{\tau=t+1}^T \sigma_{Y\tau}^2 \right)^{1/2},$$

$$d_2 = d_1 - \left(\sum_{\tau=t+1}^T \sigma_{Y\tau}^2 \right)^{1/2}, \text{ and}$$

$$Q_{XY}^*(t, T) = \frac{Q_{XY}(t, T)}{b_t^T} \frac{\varphi(U_1^T)}{\varphi(U_1^t)}. \quad (5.21)$$

To put this general formula in perspective, we will compare it to the three main approaches that have been used for pricing options: equilibrium option pricing, arbitrage-based option pricing, and GARCH option pricing. The latter pricing model can be set either in an equilibrium framework or in an arbitrage framework. Concerning the equilibrium approach, our setting is more general than

²⁰ A similar parallel is drawn in an unconditional two-period framework in Breeden and Litzenberger (1978).

²¹ Since we assume that the state variable process is Markovian, $\lambda(U_1^T)$ does not depend on the whole path of state variables but only on the last values U_T .

the usual expected utility framework since it accommodates non-separable preferences. The stochastic framework with latent variables could also accommodate state-dependent preferences such as habit formation based on state variables.

Of course, the most popular option pricing formulas among practitioners are based on arbitrage rather than on equilibrium in order to avoid in particular the specification of preferences. From the start, it should be stressed that our general formula (5.20) nests a large number of preference-free extensions of the Black–Scholes formula. In particular if $Q_{XY}(t, T) = 1$ and $\tilde{B}(t, T) = \prod_{\tau=t}^{T-1} B(\tau, \tau + 1)$, one can see that the option price (5.20) is nothing but the conditional expectation of the Black–Scholes price,²² where the expectation is computed with respect to the joint probability distribution of the rolling-over interest rate $\bar{r}_{t,T} = -\sum_{\tau=t}^{T-1} \log B(\tau, \tau + 1)$ and the cumulated volatility $\bar{\sigma}_{t,T} = \sqrt{\sum_{\tau=t+1}^T \sigma_{Y\tau}^2}$. This framework nests three well-known models. First, the most basic ones, the Black and Scholes (1973) and Merton (1973) formulas, when interest rates and volatility are deterministic. Second, the Hull and White (1987) stochastic volatility extension, since $\bar{\sigma}_{t,T}^2 = \text{Var} \left[\log \frac{S_T}{S_t} | U_1^T \right]$ corresponds to the cumulated volatility $\int_t^T \sigma_u^2 du$ in the Hull and White continuous-time setting.²³ Third, the formula allows for stochastic interest rates as in Turnbull and Milne (1991) and Amin and Jarrow (1992). However, the usefulness of our general formula (5.20) comes above all from the fact that it offers an explicit characterization of instances where the preference-free paradigm cannot be maintained. Usually, preference-free option pricing is underpinned by the absence of arbitrage in a complete market setting. However, our equilibrium-based option pricing does not preclude incompleteness and points out in which cases this incompleteness will invalidate the preference-free paradigm. The only cases of incompleteness which matter in this respect occur precisely when at least one of the two following conditions:

$$Q_{XY}(t, T) = 1 \quad (5.25)$$

²² We refer here to a BS option pricing formula where dividend flows arrive during the lifetime of the option and are accounted for in the definition of the risk neutral probability, while the option payoff does not include dividends. In other words, the BS option price is given by:

$$\pi_t^{BS} = e^{-r(T-t)} E_t[\text{Max}(0, S_T - K)] \quad (5.22)$$

$$= e^{-\delta(T-t)} S_t \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2), \quad (5.23)$$

since in the risk neutral world:

$$\log \frac{S_T}{S_t} \rightsquigarrow \mathcal{N}((r - \delta)(T - t), \sigma^2(T - t)), \quad (5.24)$$

where δ is the intensity of the dividend flow.

²³ See Subsection 5.3 for a detailed comparison between standard stochastic volatility models and our state variable framework.

$$\tilde{B}(t, T) = \prod_{\tau=t}^{T-1} B(\tau, \tau + 1) \quad (5.26)$$

is not fulfilled.

In general, preference parameters appear explicitly in the option pricing formula through $\tilde{B}(t, T)$ and $Q_{XY}(t, T)$. However, in so-called preference-free formulas, it happens that these parameters are eliminated from the option pricing formula through the observation of the bond price and the stock price. In other words, even in an equilibrium framework with incomplete markets, option pricing is preference-free if and only if there is no leverage effect in the general sense that $Q_{XY}(t, t + 1)$ and $\tilde{B}(t, t + 1)$ are predetermined. This result generalizes Amin and Ng (1993), who called this effect predictability.

It is worth noting that our results of equivalence between preference-free option pricing and no instantaneous causality between state variables and asset returns are consistent with another strand of the option pricing literature, namely GARCH option pricing. Duan (1995) derived it first in an equilibrium framework, but Kallsen and Taqqu (1998) have shown that it could be obtained with an arbitrage argument. Their idea is to complete the markets by inserting the discrete-time model into a continuous-time one, where conditional variance is constant between two integer dates. They show that such a continuous-time embedding makes possible arbitrage pricing which is *per se* preference-free. It is then clear that preference-free option pricing is incompatible with the presence of an instantaneous causality effect, since it is such an effect that prevents the embedding used by Kallsen and Taqqu (1998).

5.3 A comparison with stochastic volatility models

The typical stochastic volatility model (SV model hereafter) introduces a positive stochastic process such that its squared value h_t represents the conditional variance of the value at time $(t + 1)$ of a second-order stationary process of interest, given a conditioning information set J_t . In our setting, it is natural to define the conditioning information set J_t by (5.8). It means that the information available at time t is not summarized in general by the observation of past and current values of asset prices, since it also encompasses additional information through state variables U_t . Such a definition is consistent with the modern definition of SV processes (see Ghysels, Harvey and Renault, 1996, for a survey). It incorporates unobserved components that might capture well-documented evidence about conditional leptokurtosis and leverage effects of asset returns (given past and current returns). Moreover, such unobserved components are included in the relevant conditioning information set for option pricing models as in Hull and White (1987). The focus of interest in this subsection are the time series properties of asset returns implied

by the dynamic asset pricing model presented in Section 5.1. These time series of returns can be seen as stochastic volatility processes by Assumption 5.4 on the conditional probability distribution of the fundamentals (X_{t+1}, Y_{t+1}) given J_t . We focus on (X_{t+1}, Y_{t+1}) instead of asset returns since, by (5.9), the joint conditional probability distribution (given U_1^{t+1}) of returns for the two primitive assets is defined by Assumption 5.4 up to a shift in the mean.

Let us first consider the univariate dynamics in terms of the innovation process $\eta_{Y_{t+1}}$ of Y_{t+1} with respect to J_t defined as:

$$\eta_{Y_{t+1}} = Y_{t+1} - E[m_Y(U_1^{t+1})|U_1^t]. \quad (5.27)$$

The associated volatility and kurtosis dynamics are then characterized by:

$$\begin{aligned} h_t^Y &= \text{Var}[\eta_{Y_{t+1}}|U_1^t] \\ &= \text{Var}[m_Y(U_1^{t+1})|U_1^t] + E[\sigma_Y^2(U_1^{t+1})|U_1^t] \end{aligned} \quad (5.28)$$

and

$$\begin{aligned} \mu_{4t}^Y &= E[\eta_{Y_{t+1}}^4|U_1^t] \\ &= 3E[\sigma_Y^4(U_1^{t+1})|U_1^t] \\ &= 3[\text{Var}[\sigma_Y^2(U_1^{t+1})|U_1^t] + (E[\sigma_Y^2(U_1^{t+1})|U_1^t])^2]. \end{aligned} \quad (5.29)$$

As far as kurtosis is concerned, Equations (5.28) and (5.29) provide a representation of the fat-tail effect and its dynamics, sometimes termed the heterokurtosis effect. This extends the representation of the standard mixture model, first introduced by Clark (1973) and extended by Gallant, Hsieh and Tauchen (1991). Indeed, in the particular case where:

$$\text{Var}[m_Y(U_1^{t+1})|U_1^t] = 0, \quad (5.30)$$

we get the following expression²⁴ for the conditional kurtosis coefficient:

$$\frac{\mu_{4t}^Y}{(h_t^Y)^2} = 3[1 + (c_t^Y)^2] \quad (5.31)$$

with:

$$c_t^Y = \frac{(\text{Var}[\sigma_Y^2(U_1^{t+1})|U_1^t])^{\frac{1}{2}}}{E[\sigma_Y^2(U_1^{t+1})|U_1^t]}. \quad (5.32)$$

This expression emphasizes that the conditional normality assumption does not preclude conditional leptokurtosis with respect to a smaller set of conditioning information. It should be emphasized that formula (5.31) allows for even more

²⁴ It corresponds to the formula given by Gallant, Hsieh and Tauchen (1991) on page 204.

leptokurtosis than the standard formula since the probability distributions considered are still conditioned on a large information set, including possibly unobserved components. An additional projection on the reduced information set defined by past and current values of observed asset returns will increase the kurtosis coefficient. In other words, our model allows for innovation terms in asset returns that, even standardized by a genuine stochastic volatility (including a mixture effect), are still leptokurtic. Moreover, condition (5.30) is likely not to hold, providing an additional degree of freedom in our representation of kurtosis dynamics. If we consider the stock return itself instead of the dividend growth, the violation of (5.30) is even more likely since $m_Y(U_1^{t+1})$ is to be replaced by the “expected” return $m_Y(U_1^{t+1}) + \log(\varphi(U_1^{t+1}) + 1/\varphi(U_1^t))$. Condition (5.30) will be violated when this expected return differs from its expected value computed by investors according to our equilibrium asset pricing model, that is $E[m_Y(U_1^{t+1}) + \log(\varphi(U_1^{t+1}) + 1/\varphi(U_1^t))|U_1^t]$. We will show now that it is precisely this difference which can produce a genuine leverage effect in stock returns, as defined by Black (1976) and Nelson (1991) for conditionally heteroskedastic returns.²⁵ This justifies *a posteriori* the use of the expression leverage effect in Section 5.2 to account for the fact that the probability distribution of (X_{t+1}, Y_{t+1}) given U_1^{t+1} depends (through the functions $m_X, m_Y, \sigma_X, \sigma_Y$ and σ_{XY}) on the contemporaneous value U_{t+1} of the state process.²⁶

According to the standard terminology, the stochastic volatility dividend process exhibits a leverage effect if and only if:

$$\text{Cov}[\eta_{Y_{t+1}}, h_{t+1}^Y | U_1^t] = \text{Cov}[m_Y(U_1^{t+1}), h_{t+1}^Y | U_1^t] < 0. \quad (5.33)$$

Barring the restriction (5.30), if $m_Y(U_1^{t+1})$ is truly a function of U_{t+1} , the condition in (5.33) amounts to the negativity of the sum of two terms:

$$\text{Cov}[m_Y(U_1^{t+1}), \text{Var}[m_Y(U_1^{t+2}) | U_1^{t+1}] | U_1^t] \quad (5.34)$$

and:

$$\text{Cov}[m_Y(U_1^{t+1}), E[\sigma_Y^2(U_1^{t+2}) | U_1^{t+1}] | U_1^t]. \quad (5.35)$$

In other words, the leverage effect of the stochastic volatility process Y_{t+1} can be produced by any of the two following leverage effects or both.²⁷ The conditional

²⁵ We will conduct the discussion below in terms of $m_Y(U_1^{t+1})$ but it could be reinterpreted in terms of $m_Y(U_1^{t+1}) + \log(\varphi(U_1^{t+1}) + 1/\varphi(U_1^t))$.

²⁶ The key point is that the mean functions $m_X(U_1^{t+1})$ and $m_Y(U_1^{t+1})$ depend on U_{t+1} . However, if these functions are replaced by the shifted conditional expectations for asset returns according to (5.9), the functions $\sigma_X(U_1^{t+1})$, $\sigma_Y(U_1^{t+1})$ and $\sigma_{XY}(U_1^{t+1})$ will be reintroduced in these expected returns through the functions $\lambda(U_1^{t+1})$ and $\varphi(U_1^{t+1})$ defined by Proposition 5.3.

²⁷ This decomposition of the leverage effect in two terms is the exact analogue of the decomposition discussed in Fiorentini and Sentana (1998) and Meddahi (1999) for persistence.

mean process $m_Y(U_1^{t+1})$ may be a stochastic volatility process which features a leverage effect defined by the negativity of (5.34). Or the process Y_{t+1} itself may be characterized by a leverage effect and then (5.35) be negative, which means that bad news about expected returns (when $m_Y(U_1^{t+1})$ is smaller than its unconditional expectations) implies on average a higher expected volatility of Y , that is a value of $E[\sigma_Y^2(U_1^{t+2})|U_1^{t+1}]$ greater than its unconditional mean. To summarize, Assumption 5.4 not only allows us to capture the standard features of a stochastic volatility model (in terms of heavy tails and leverage effects) but also provides for a richer set of possible dynamics. Moreover, we can certainly extend these ideas to multivariate dynamics either for the joint behavior of market and stock returns or for any portfolio consideration. For instance, the dependence of $\sigma_{XY}(U_1^{t+1})$ on the whole set of state variables offers great flexibility to model the stochastic behavior of correlation coefficients, as recently put forward empirically by Andersen *et al.* (1999). This last feature is clearly highly relevant for asset allocation or conditional beta pricing models.

6 Conclusion

In this chapter, we provided a unifying analysis of latent variable models in finance through the concept of stochastic discount factor (SDF). We extended both the asset pricing factor models and the equilibrium dynamic asset pricing models through a conditioning on state variables. This conditioning enriches the dynamics of asset returns through instantaneous causality between the asset returns and the latent variables. Such correlation or leverage effects explain departures from usual CAPM pricing for stocks or Black and Scholes and Hull and White pricing for options. The dependence of conditional covariances on the state variables allows for a rich dynamic stochastic behavior of correlation coefficients which is important for asset allocation or value-at-risk strategies.

The enriched set of empirical implications from such dynamic latent variable models requires us to set up a general inference methodology which will account for the inobservability of both cross-sectional factors and longitudinal latent variables. Indirect inference, efficient method of moments or Markov chain Monte Carlo (MCMC) for Bayesian inference are all avenues that can prove useful in this context, since they have been used successfully in stochastic volatility models.

References

- Amin, K.I. and Jarrow, R. (1992), Pricing options in a stochastic interest rate economy, *Mathematical Finance*, **3**(3), 1–21.
- Amin, K.I. and Ng, V.K. (1993), Option Valuation with Systematic Stochastic Volatility, *Journal of Finance*, **XLVIII**, 3, 881–909.

- Andersen, T.B., Bollerslev, T., Diebold, F.X. and Labys, P. (1999), The distribution of exchange rate volatility, NBER Working Paper no. 6961.
- Bansal, R., Hsieh, D. and Viswanathan, S. (1993), No arbitrage and arbitrage pricing: a new approach, *Journal of Finance* **48**, 1231–62.
- Bartholomew, D.J. (1987), *Latent Variable Models and Factor Analysis*. Oxford University Press, Oxford.
- Black, F. (1976), Studies of stock market volatility Changes, *1976 Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pp. 177–81.
- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–59.
- Breeden, D. and Litzenberger, R. (1978), Prices of state-contingent claims implicit in option prices, *Journal of Business* **51**, 621–51.
- Burt, C. (1941), *The Factors of the Mind: An Introduction to factor Analysis in Psychology*. Macmillan, New York.
- Chamberlain, G. and Rothschild, M. (1983), Arbitrage and mean variance analysis on large asset markets, *Econometrica* **51**, 1281–304.
- Clark, P.K. (1973), A subordinated stochastic process model with variance for speculative prices, *Econometrica* **41**, 135–56.
- Cox, D.R. (1981), Statistical analysis of time series: some recent developments, *Scandinavian Journal of Statistics* **8**, 93–115.
- Cox, J., Ingersoll, J. and Ross, S. (1981), A reexamination of traditional hypotheses about the term structure of interest rates, *Journal of Finance* **36**, 769–99.
- Dai, Q. and Singleton, K.J. (1999), Specification analysis of term structure models, forthcoming in the *Journal of Finance*.
- Diebold, F.X. and Nerlove, M. (1989), The dynamics of exchange rate volatility: a multivariate latent factor ARCH model, *Journal of Applied Econometrics* **4**, 1–21.
- Duan, J.C. (1995), The GARCH option pricing model, *Mathematical Finance* **5**, 13–32.
- Duffie D. and Kan, R. (1996), A yield-factor model of interest rates, *Mathematical Finance*, 379–406.
- Engle, R.F., Ng, V. and Rothschild, M. (1990), Asset pricing with a factor arch covariance structure: empirical estimates with treasury bills, *Journal of Econometrics* **45**, 213–38.
- Epstein, L. and Zin, S. (1989), Substitution, risk aversion and the temporal behavior of consumption and asset returns I: a theoretical framework, *Econometrica* **57**, 937–69.
- Epstein, L. and Zin, S. (1991), Substitution, risk aversion and the temporal behavior of consumption and asset returns I: an empirical analysis, *Journal of Political Economy* **99**, 2, 263–86.
- Ferson, W.E. and Korajczyk, R.A. (1995), Do arbitrage pricing models explain the predictability of stock returns, *Journal of Business* **68**, 309–49.
- Fiorentini, G. and Sentana, E. (1998), Conditional means of time series processes and time series processes for conditional means, *International Economic Review* **39**, 1101–18.
- Florens, J.-P. and Mouchart, M. (1982), A note on noncausality, *Econometrica* **50**(3), 583–91.
- Florens, J.-P., Mouchart, M. and J.-Rollin, P. (1990), *Elements of Bayesian Statistics*. Dekker, New York.
- Gallant, A.R., Hsieh, D. and Tauchen, G. (1991), on fitting a recalcitrant series: the pound/dollar exchange rate 1974–1983, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, (eds. William Barnett, A., Jim Powell and

- Georges Tauchen), Cambridge University Press, Cambridge.
- Garcia R., Luger, R. and Renault, E. (1999), Asymmetric smiles, leverage effects and structural parameters, working paper, CIRANO, Montreal, Canada.
- Ghysels, E., Harvey, A. and Renault, E. (1996), *Stochastic Volatility, Statistical Methods in Finance* (C. Rao, R. and Maddala, G.S.). North-Holland, Amsterdam, pp. 119–91.
- Granger, C.W.J. (1969), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**, 424–38.
- Hamilton, J.D. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**, 357–84.
- Hansen, L. and Richard, S. (1987), The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models, *Econometrica* **55**, 587–614.
- Harrison, J.M. and Kreps, D. (1979), Martingale and Arbitrage in Multiperiod Securities Markets, *Journal of Economic Theory* **20**, 381–408.
- Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.
- Harvey, C.R. (1991), The world price of covariance risk, *Journal of Finance* **46**, 111–57.
- Hull, J. and White, A. (1987), The pricing of options on assets with stochastic volatilities, *Journal of Finance* **XLII**, 281–300.
- Kallsen, J. and Taqqu, M.S. (1998), Option pricing in ARCH-type models, *Mathematical Finance*, 13–26.
- King, M., Sentana, E. and Wadhwani, S. (1994), Volatility and links between national stock markets, *Econometrica* **62**, 901–33.
- Lintner, J. (1965), The Valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets, *Review of Economics and Statistics* **47**, 13–37.
- Kreps, D. and Porteus, E. (1978), Temporal resolution of uncertainty and dynamic choice theory, *Econometrica* **46**, 185–200.
- Lucas, R. (1978), Asset prices in an exchange economy, *Econometrica* **46**, 1429–45.
- Meddahi, N. (1999), Aggregation of long memory processes, unpublished paper, Université de Montréal.
- Meddahi, N. and Renault, E. (1996), Aggregation and marginalization of GARCH and stochastic volatility models, GREMAQ DP 96.30.433, Toulouse.
- Merton, R.C. (1973), Rational theory of option pricing, *Bell Journal of Economics and Management Science* **4**, 141–83.
- Nelson, D.B. (1991), Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–70.
- Pitt, M.K. and Shephard, N. (1999), Time-varying covariances: a factor stochastic volatility approach, *Bayesian Statistics* **6**, 547–70.
- Renault, E. (1999), *Dynamic Factor Models in Finance, Core Lectures*. Oxford University Press, Oxford, forthcoming.
- Ross, S. (1976), The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–60.
- Sharpe, W.F. (1964), Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**, 425–42.
- Sims, C.A. (1972), Money, income and causality, *American Economic Review* **62**, 540–52.
- Spearman, C. (1927), *The Abilities of Man*. Macmillan, New York.
- Turnbull, S. and Milne, F. (1991), A simple approach to interest-rate option pricing, *Review of Financial Studies* **4**, 87–121.

Monte Carlo Methods for Security Pricing*

Phelim Boyle, Mark Broadie and Paul Glasserman

1 Introduction

In recent years the complexity of numerical computation in financial theory and practice has increased enormously, putting more demands on computational speed and efficiency. Numerical methods are used for a variety of purposes of finance. These include the valuation of securities, the estimation of their sensitivities, risk analysis, and stress testing of portfolios. The Monte Carlo method is a useful tool for many of these calculations, evidenced in part by the voluminous literature of successful applications. For a brief sampling, the reader is referred to the stochastic volatility applications in Duan (1995), Hull and White (1987), Johnson and Shanno (1987), and Scott (1987);¹ the valuation of mortgage-backed securities in Schwartz and Torous (1989); the valuation of path-dependent options in Kemna and Vorst (1990); the portfolio optimization in Worzel et al. (1994); and the valuation of interest-rate derivative claims in Carverhill and Pang (1995). In this paper we focus on recent methodological developments. We review the Monte Carlo approach and describe some recent applications in the finance area.

In modern finance, the prices of the basic securities and the underlying state variables are often modelled as continuous-time stochastic processes. A derivative security, such as a call option, is a security whose payoff depends on one or more of the basic securities. Using the assumption of no arbitrage, financial economists have shown that the price of a generic derivative security can be expressed as the expected value of its discounted payouts. This expectation is taken with respect to a transformation of the original probability measure known as the *equivalent martingale measure* or the *risk-neutral measure*. The book by Duffie (1996) provides an excellent account of this material.

The Monte Carlo method lends itself naturally to the evaluation of security prices represented as expectations. Generically, the approach consists of the following

* Reprinted from the *Journal of Economic Dynamics and Control* **21** (1977) 1267–1321.

¹ Wiggins (1987) also studies pricing under stochastic volatility but does not use Monte Carlo simulation.

steps:

- Simulate sample paths of the underlying state variables (e.g., underlying asset prices and interest rates) over the relevant time horizon. Simulate these according to the risk-neutral measure.
- Evaluate the discounted cash flows of a security on each sample path, as determined by the structure of the security in question.
- Average the discounted cash flows over sample paths.

In effect, this method computes a multi-dimensional integral – the expected value of the discounted payouts over the space of sample paths. The increase in the complexity of derivative securities in recent years has led to a need to evaluate high dimensional integrals.

Monte Carlo becomes increasingly attractive compared to other methods of numerical integration as the dimension of the problem increases. Consider the integral of the function $f(x)$ over the d -dimensional unit hypercube. The simple (or crude) Monte Carlo estimate of the integral is equal to the average value of the function f over n points selected at random² from the unit hypercube. From the strong law of large numbers this estimate converges to the true value of the integrand as n tends to infinity. In addition, the central limit theorem assures us that the standard error³ of the estimate tends to zero as $1/\sqrt{n}$. Thus the error convergence rate is independent of the dimension of the problem and this is the dominant advantage of the method over classical numerical integration approaches. The only restriction on the function f is that it should be square integrable, and this is a relatively mild restriction.

Furthermore, the Monte Carlo method is flexible and easy to implement and modify. In addition, the increased availability of powerful computers has enhanced the attractiveness of the method. There are some disadvantages of the method but in recent years progress has been made in overcoming them. One drawback is that for very complex problems a large number of replications may be required to obtain precise results. Different variance reduction techniques have been developed to increase precision. Two of the classical variance reduction techniques are the control variate approach and the antithetic variate method. More recently, moment matching, importance sampling, and conditional Monte Carlo methods have been introduced in finance applications.

Another technique for speeding up the valuation of multidimensional integrals uses deterministic sequences rather than random sequences. These deterministic

² In standard Monte-Carlo application the n points are usually not truly random but are generated by a deterministic algorithm and are described as *pseudorandom* numbers.

³ We can readily estimate the variance of the Monte Carlo estimate by using the same set of n random numbers to estimate the expected value of f^2 .

sequences are chosen to be more evenly dispersed throughout the region of integration than random sequences. If we use these sequences to estimate multidimensional integrals we can often improve the convergence. Deterministic sequences with this property are known as low-discrepancy sequences or quasi-random sequences. Using this approach one can in theory derive deterministic error bounds, though the practical use of the bounds is problematic. In contrast, standard Monte Carlo yields simple, useful probabilistic error bounds. Although low-discrepancy sequences are well known in computational physics they have only recently been applied in finance problems. There are different procedures for generating such low-discrepancy sequences and these procedures are generally based on number theoretic methods. We describe some of the recent developments in this area. We also discuss applications of this approach to problems in finance and conduct some rough comparisons between standard Monte Carlo methods and two different quasi-random approaches.

Until recently, the valuation of American style options was widely considered outside the scope of Monte Carlo. However Tilley (1993), Barraquand and Martineau (1995), and Broadie and Glasserman (1997), and have proposed approaches to this problem, and there has been other related work as well. We provide a brief survey of the recent research progress in this area.

The layout of the paper is as follows. Variance reduction techniques are described in the next section. The ideas behind the use of low-discrepancy sequences and brief numerical comparisons with standard Monte Carlo methods are given in Section 3. Price sensitivity estimation using simulation is discussed in Section 4. Various approaches to pricing American options using simulation are briefly described in Section 5. Other issues are touched on briefly in Section 6.

2 Variance reduction techniques

In this section, we first discuss the role of variance reduction in meeting the broader objective of improving the computational efficiency of Monte Carlo simulations. We then discuss specific variance reduction techniques and illustrate their application to pricing problems.

2.1 Variance reduction and efficiency improvement

The reduction of variance seems so obviously desirable that the precise argument for its benefit is sometimes overlooked. We briefly review the underlying justification for variance reduction and examine it from the perspective improving computational efficiency.

Suppose we want to compute a parameter θ – for example, the price of a derivative security. Suppose we can generate by Monte Carlo an i.i.d. sequence $\{\hat{\theta}_i, i = 1, 2, \dots\}$, where each $\hat{\theta}_i$ has expectation θ and variance σ^2 . A natural estimator of θ based on n replications is then the sample mean

$$\frac{1}{n} \sum_{i=1}^n \hat{\theta}_i.$$

By the central limit theorem, for large n this sample mean is approximately normally distributed with mean θ and variance σ^2/n . Probabilistic error bounds in the form of confidence intervals follow readily from the normal approximation, and indicate that the error in the estimator is proportional to $\sigma/\sqrt{n}$. Thus, decreasing the variance σ^2 by a factor of 10, say, while leaving everything else unchanged, does as much for error reduction as increasing the number of samples by a factor of 100.

Suppose, now, that we have a choice between two types of Monte Carlo estimates which we denote by $\{\hat{\theta}_i^{(1)}, i = 1, 2, \dots\}$ and $\{\hat{\theta}_i^{(2)}, i = 1, 2, \dots\}$. Suppose that both are unbiased, so that $E[\hat{\theta}_i^{(1)}] = E[\hat{\theta}_i^{(2)}] = \theta$, but $\sigma_1 < \sigma_2$, where $\sigma_j^2 = \text{Var}[\hat{\theta}^{(j)}]$, $j = 1, 2$. From our previous observations it follows that a sample mean of n replications of $\hat{\theta}^{(1)}$ gives a more precise estimate of θ than does a sample mean of n replications of $\hat{\theta}^{(2)}$. But this analysis oversimplifies the comparison because it fails to capture possible differences in the computational effort required by the two estimators. Generating n replications of $\hat{\theta}^{(1)}$ may be more time-consuming than generating n replications of $\hat{\theta}^{(2)}$; smaller variance is not sufficient grounds for preferring one estimator over another.

To compare estimators with different computational requirements as well as different variances, we argue as follows. Suppose the work required to generate one replication of $\hat{\theta}^{(j)}$ is a constant b_j , $j = 1, 2$. (In some problems, the work per replication is stochastic; assuming it is constant simplifies the discussion.) With computing time t , the number of replications of $\hat{\theta}^{(j)}$ that can be generated is $\lfloor t/b_j \rfloor$; for simplicity, we drop the $\lfloor \cdot \rfloor$ and treat the ratios t/b_j as though they were integers. The two estimators available with computing time t are therefore

$$\frac{b_1}{t} \sum_{i=1}^{t/b_1} \hat{\theta}_i^{(1)} \quad \text{and} \quad \frac{b_2}{t} \sum_{i=1}^{t/b_2} \hat{\theta}_i^{(2)}.$$

For large t , these are approximately normally distributed with mean θ and with standard deviations

$$\sigma_1 \sqrt{\frac{b_1}{t}} \quad \text{and} \quad \sigma_2 \sqrt{\frac{b_2}{t}}.$$

Thus, for large t , the first estimator should be preferred over the second if

$$\sigma_1^2 b_1 < \sigma_2^2 b_2. \quad (1)$$

Equation (1) provides a sound basis for trading-off estimator variance and computational requirements. In light of the discussion leading to (1), it is reasonable to take the product of variance and work per run as a measure of *efficiency*. Using efficiency as a basis for comparison, the lower-variance estimator should be preferred only if the variance ratio σ_1^2/σ_2^2 is smaller than the work ratio b_2/b_1 . By the same argument, a higher-variance estimator may actually be preferable if it takes much less time to generate.

In its simplest form, the principle expressed in (1) dates at least to Hammersley and Handscomb (1964, p.22). More recently, the idea has been substantially extended by Glynn and Whitt (1992). They allow the work per run to be random (in which case each b_j is the *expected* work per run) and also consider efficiency in the presence of bias.

2.2 Antithetic variates

Equipped with a basis for evaluating potential efficiency improvements, we can now consider specific variance reduction techniques. One of the simplest and most widely used techniques in financial pricing problems is the method of *antithetic variates*. We introduce it with a simple example, then generalize.

Consider the problem of computing the Black–Scholes price of a European call option on a no-dividend stock. Of course, there is no need to evaluate this price by simulation, but the example serves as a useful introduction. In the Black–Scholes model, the stock price follows a lognormal diffusion. Independent replications of the terminal stock price under the risk-neutral measure can be generated from the formula

$$S_T^{(i)} = S_0 e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z_i}, \quad i = 1, \dots, n, \quad (2)$$

where S_0 is the current stock price, r is the riskless interest rate, σ is the stock's volatility, T is the option's maturity, and the $\{Z_i\}$ are independent samples from the standard normal distribution. See, e.g., Hull (2000) for background on this model, and see Devroye (1986) for methods of sampling from the normal distribution. Based on n replications, a moment-matched estimator of the price of an option with strike K is given by

$$\hat{C} = \frac{1}{n} \sum_{i=1}^n C_i \equiv \frac{1}{n} \sum_{i=1}^n e^{-rT} \max\{0, S_T^{(i)} - K\}. \quad (3)$$

In this context, the method of antithetic variates⁴ is based on the observation that if Z_i has a standard normal distribution, then so does $-Z_i$. The price $\tilde{S}_T^{(i)}$ obtained from (2) with Z_i replaced by $-Z_i$ is thus a valid sample from the terminal stock price distribution. Similarly, each

$$\tilde{C}_i = e^{-rT} \max\{0, \tilde{S}_T^{(i)} - K\}$$

is an unbiased estimator of the option price, as is therefore

$$\hat{C}_{AV} = \frac{1}{n} \sum_{i=1}^n \frac{C_i + \tilde{C}_i}{2}.$$

A heuristic argument for preferring $\hat{C}_{AV}$ notes that the random inputs obtained from the collection of antithetic pairs $\{(Z_i, -Z_i)\}$ are more regularly distributed than a collection of $2n$ independent samples. In particular, the sample mean over the antithetic pairs always equals the population mean of 0, whereas the mean over finitely many independent samples is almost surely different from 0. If the inputs are made more regular, it may be hoped that the outputs are more regular as well. Indeed, a large value of $S_T^{(i)}$ resulting from a large Z_i will be paired with a small value of $\tilde{S}_T^{(i)}$ obtained from $-Z_i$.

A more precise argument compares efficiencies. Because C_i and $\tilde{C}_i$ have the same variance,

$$\text{Var} \left[\frac{C_i + \tilde{C}_i}{2} \right] = \frac{1}{2} (\text{Var}[C_i] + \text{Cov}[C_i, \tilde{C}_i]). \quad (4)$$

Thus, we have $\text{Var}[\hat{C}_{AV}] \leq \text{Var}[\hat{C}]$ if $\text{Cov}[C_i, \tilde{C}_i] \leq \text{Var}[C_i]$. However, $\hat{C}_{AV}$ uses twice as many replications as $\hat{C}$, so we must account for differences in computational requirements. If generating the Z_i takes a negligible fraction of the work per replication (which would typically be the case in the pricing of a more elaborate option), then the work to generate $\hat{C}_{AV}$ is roughly double the work to generate $\hat{C}$. Thus, for antithetics to increase efficiency, we require

$$2 \text{Var}[\hat{C}_{AV}] \leq \text{Var}[\hat{C}],$$

which, in light of (4), simplifies to the requirement that $\text{Cov}[C_i, \tilde{C}_i] \leq 0$.

That this condition is met is easily demonstrated. Define ϕ so that $C_i = \phi(Z_i)$; ϕ is the composition of the mappings from Z_i to the stock price and from the stock price to the discounted option payoff. As the composition of two increasing functions, ϕ is monotone, so by a standard inequality (e.g., Section 2.2 of Barlow

⁴ This method was introduced to option pricing in Boyle (1977), where its use was illustrated in the pricing of a European call on a dividend-paying stock.

and Proschan 1975)

$$E[\phi(Z_i)\phi(-Z_i)] \leq E[\phi(Z_i)]E[\phi(-Z_i)], \quad (5)$$

i.e., $\text{Cov}[C_i, \tilde{C}_i] \equiv E[\phi(Z_i)\phi(-Z_i)] - E[\phi(Z_i)]E[\phi(-Z_i)] \leq 0$, and we may conclude that antithetics help.

This argument can be adapted to show that the method of antithetic variates increases efficiency in pricing a European put and other options that depend monotonically on inputs (e.g., Asian options). The notable departure from monotonicity in some barrier options (e.g., a down-and-in call) suggests that the use of antithetics in pricing these options may sometimes be less effective.

In computing confidence intervals with antithetic variates, it is essential that the standard error be estimated using the sample standard deviation of the n averaged pairs $(C_i + \tilde{C}_i)/2$ and not the $2n$ individual observations $C_1, \tilde{C}_1, \dots, C_n, \tilde{C}_n$. The averaged pairs are independent but the individual observations are not. This is a case (we will see others shortly) in which the use of a variance reduction technique affects the estimation of the standard error and, in particular, requires some “batching” of observations to deal with dependence.

It is worth noting that the method of antithetic variates is by no means restricted to simulations whose only stochastic inputs are standard normal variates. The most primitive stochastic input in most simulations is a sequence $\{U_n\}$ of independent variates uniformly distributed on the unit interval. In this case, $1 - U_n$ has the same distribution as U_n , and the pair $(U_n, 1 - U_n)$ are called antithetic because they exhibit negative dependence. If the simulation output depends monotonically on the input random numbers, then the output obtained from $\{1 - U_1, 1 - U_2, \dots\}$ will be negatively correlated with that obtained from $\{U_1, U_2, \dots\}$, resulting in increased efficiency compared with independent replications.

For further general background on antithetic variates and other methods based on correlation induction, see Bratley, Fox, and Schrage (1987), Hammersley and Handscomb (1964), Glynn and Iglehart (1988), and references there. For some examples of application in finance, see Boyle (1977), Clewlow and Carverhill (1994), and Hull and White (1987).

2.3 Control variates

The method of *control variates* is among the most widely applicable, easiest to use, and effective of the variance reduction techniques.⁵ Simply put, the principle underlying this technique is “use what you know.”

The most straightforward implementation of control variates replaces the evaluation of an unknown expectation with the evaluation of the *difference* between

⁵ The earliest application of this technique to option pricing is Boyle (1977).

the unknown quantity and another expectation whose value is known. A specific illustration can be found in the analysis of Boyle and Emanuel (1985) and Kemna and Vorst (1990) of Asian options. Let P_A be the price of an option whose payoff depends on the arithmetic average of the underlying asset. Let P_G be the price of an option equivalent in every respect except that a geometric average replaces the arithmetic average. Most options based on averages use arithmetic averaging, so P_A is of much greater practical value; but whereas P_A is analytically intractable, P_G can often be evaluated in closed form. Can knowledge of P_G be leveraged to compute P_A ?

It can, through the control variate method. Write $P_A = E[\hat{P}_A]$ and $P_G = E[\hat{P}_G]$, where $\hat{P}_A$ and $\hat{P}_G$ are the discounted option payoffs for a single simulated path of the underlying asset. Then

$$P_A = P_G + E[\hat{P}_A - \hat{P}_G];$$

in other words, P_A can be expressed as the known price P_G plus the expected difference between $\hat{P}_A$ and $\hat{P}_G$. An unbiased estimator of P_A is thus provided by

$$\hat{P}_A^{\text{cv}} = \hat{P}_A + (P_G - \hat{P}_G). \quad (6)$$

This representation⁶ suggests a slightly different interpretation: $\hat{P}_A^{\text{cv}}$ adjusts the straightforward estimator $\hat{P}_A$ according to the difference between the known value P_G and the observed value $\hat{P}_G$. The known error $(P_G - \hat{P}_G)$ is used as a *control* in the estimation of P_A .

If most of the computational effort goes to generating paths of the underlying asset, then the additional work required to evaluate $\hat{P}_G$ along with $\hat{P}_A$ is minor. It therefore seems reasonable to compare variances alone. Since

$$\text{Var}[\hat{P}_A^{\text{cv}}] = \text{Var}[\hat{P}_A] + \text{Var}[\hat{P}_G] - 2\text{Cov}[\hat{P}_A, \hat{P}_G],$$

this method is effective if the covariance between $\hat{P}_A$ and $\hat{P}_G$ is large. The numerical results of Kemna and Vorst indicate that this is indeed the case. Fu, Madan, and Wang (1998) have investigated the use of other control variates for Asian options, based on Laplace transform values. These appear to be less strongly correlated with the option price.

A closer examination of (6) reveals that this estimator does not make optimal use of the relation between the two option prices. Consider the family of unbiased estimators

$$\hat{P}_A^\beta = \hat{P}_A + \beta(P_G - \hat{P}_G), \quad (7)$$

⁶ To go from (6) to Boyle's (1977) example, let P_G be the price of a European call option on a no-dividend stock and let P_A be the corresponding option price in the presence of dividends.

parameterized by the scalar β . We have

$$\text{Var}[\hat{P}_A^\beta] = \text{Var}[\hat{P}_A] + \beta^2 \text{Var}[\hat{P}_G] - 2\beta \text{Cov}[\hat{P}_A, \hat{P}_G].$$

The variance-minimizing β is therefore

$$\beta^* = \frac{\text{Cov}[\hat{P}_A, \hat{P}_G]}{\text{Var}[\hat{P}_G]}.$$

Depending on the application, β^* may or may not be close to 1, the implicit value in (6). In using an estimator of the form (6), we forgo an opportunity for greater variance reduction. Indeed, whereas (6) may increase or decrease variance, an estimator based on β^* is guaranteed not to increase variance, and will result in a strict decrease in variance so long as $\hat{P}_A$ and $\hat{P}_G$ are not uncorrelated.

In practice, of course, we rarely know β^* because we rarely know $\text{Cov}[\hat{P}_A, \hat{P}_G]$. However, given n independent replications $\{(P_{Ai}, P_{Gi}), i = 1, \dots, n\}$ of the pairs $(\hat{P}_A, \hat{P}_G)$ we can estimate β^* via regression. At this point we face a choice. Using all n replications to compute an estimate $\hat{\beta}$ of β^* introduces a bias in the estimator

$$\frac{1}{n} \sum_{i=1}^n P_{Ai} + \hat{\beta} \left(P_G - \frac{1}{n} \sum_{i=1}^n P_{Gi} \right),$$

and its estimated standard error because of the dependence between $\hat{\beta}$ and the P_{Gi} . Reserving n_1 replications for the estimation of β^* and the remaining $n - n_1$ replications for the sample mean of the P_{Gi} (typically with $n_1 \ll n$) eliminates the bias but may deteriorate the estimate of β^* . Neither issue significantly limits the applicability of the method, because the possible bias vanishes as n increases and because the estimate of β^* need not be very precise to achieve a reduction in variance.

The advantage of working with (7) over (6) becomes even more pronounced when further controls are introduced. For example, when the asset price is simulated under risk-neutral probabilities, the present value $e^{-rT} E[S_T]$ of the terminal price must equal the current price S_0 . We can therefore form the estimator

$$\hat{P}_A + \beta_1(P_G - \hat{P}_G) + \beta_2(S_0 - e^{-rT} S_T).$$

The variance-minimizing coefficients (β_1^*, β_2^*) are easily found by multiple regression. This optimization step seems particularly crucial in this case; for whereas one might guess that β_1^* is close to 1, it seems unlikely that β_2^* would be. Optimizing over the β s also allows us to exploit controls that are negatively correlated with the option payoff.

For further general background on control variates see Bratley, Fox, and Schrage (1987), Glynn and Iglehart (1988), and Lavenberger and Welch (1981). For examples of control variate applications in finance, see Boyle (1977), Boyle and

Emanuel (1985), Broadie and Glasserman (1996), Carverhill and Pang (1995), Clewlow and Carverhill (1994), Duan (1995), and Kemna and Vorst (1990).

2.4 Moment matching methods

Next we describe a variance reduction technique proposed by Barraquand (1995), who termed it *quadratic resampling*. His technique is based on *moment matching*. As before, we introduce it with the simple example of estimating the European call option price on a single asset and then generalize.

Let Z_i , $i = 1, \dots, n$, denote independent standard normals used to drive a simulation. The sample moments of the n Z 's will not exactly match those of the standard normal. The idea of moment matching is to transform the Z 's to match a finite number of the moments of the underlying population. For example, the first moment of the standard normal can be matched by defining

$$\tilde{Z}_i = Z_i - \bar{Z}, \quad i = 1, \dots, n, \quad (8)$$

where $\bar{Z} = \sum_{i=1}^n Z_i/n$ is the sample mean of the Z 's. Note that the $\tilde{Z}_i$'s are normally distributed if the Z_i 's are normal. However, the $\tilde{Z}_i$'s are not independent. As before, terminal stock prices are generated from the formula

$$\tilde{S}_T(i) = S_0 e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}\tilde{Z}_i}, \quad i = 1, \dots, n.$$

An unbiased estimator of the call option price is the average of the n values $\tilde{C}_i = e^{-rT} \max(\tilde{S}_T(i) - K, 0)$.

In the standard Monte Carlo method, confidence intervals for the true value C could be estimated from the sample mean and variance of estimator. This cannot be done here since the n values of $\tilde{Z}$ are no longer independent, and hence the values $\tilde{C}_i$ are not independent. This points out one drawback of the moment matching method: confidence intervals are not as easy to obtain.⁷ Indeed, for confidence intervals it appears to be necessary to apply moment matching to independent batches of runs and estimate the standard error from the batch means. This reduces the efficacy of the method compared with matching moments across all runs.

Equation (8) showed one way to match the first moment of a distribution with mean zero. If the underlying population does not have a zero mean, transformed Z 's could be generated using $\tilde{Z}_i = Z_i - \bar{Z} + \mu_Z$, where μ_Z is the population mean. The idea can easily be extended to match two moments of a distribution. In this case, an appropriate transformation is

$$\tilde{Z}_i = (Z_i - \bar{Z}) \frac{\sigma_Z}{s_Z} + \mu_Z, \quad i = 1, \dots, n, \quad (9)$$

⁷ The point is not merely a minor technical issue. The sample variance of the $\tilde{C}_i$'s is usually a poor estimate of $\text{Var}[\tilde{C}_i]$.

where s_Z is the sample standard deviation of the Z_i 's and σ_Z is the population standard deviation. Of course, for a standard normal, $\mu_Z = 0$ and $\sigma_Z = 1$. An estimator of the call option price is the average of the n values $\tilde{C}_i$.

Using the transformation (9), the $\tilde{Z}_i$'s are not normally distributed even if the Z_i 's are normal. Hence, the corresponding $\tilde{C}_i$ are biased estimators of the true option value. For most financial problems of practical interest, this bias is likely to be small. However, the bias can be arbitrarily large in extreme circumstances (even when only the first moment of the distribution is matched).⁸ The dependence and bias in the moment matching method makes it difficult to quantify the improvement in general analytical terms.

The moment matching method is another example of the idea to "use what you know." In this simple European option example, the mean and variance of the terminal stock price S_T is also known. So the moment matching idea could be applied to the simulated terminal stock values $S_T(i)$. In this case, to match the first moment, define

$$\tilde{S}_T(i) = S_T(i) - \bar{S}_T + \mu_{S_T}, \quad (10)$$

where $\mu_{S_T} = S_0 e^{rT}$ and $\bar{S}_T$ is the sample mean of the $S_T(i)$'s. To match the first two moments, define

$$\tilde{S}_T(i) = (S_T(i) - \bar{S}_T) \frac{\sigma_{S_T}}{s_{S_T}} + \mu_{S_T}, \quad (11)$$

where $\sigma_{S_T} = S_0 \sqrt{e^{2rT}(e^{\sigma^2 T} - 1)}$ and s_{S_T} is the sample standard deviation of the $S_T(i)$'s. Duan and Simonato (1998) use a related method. They apply a multiplicative transformation to asset prices to enforce the martingale property over a finite set of paths.⁹ They apply their method to GARCH option pricing.

Comparisons of various moment matching strategies are given in Table 1. For this comparison, $n = 100$ simulation trials were used to estimate the European call option price. Standard errors were estimated by re-simulation. That is, $m = 10\,000$ simulation trials were conducted, each one based on n replications of the estimator. The sample standard deviation of the m simulation estimates gives an estimate of the standard error of a single simulation estimate. Root-mean-squared errors are not reported because they are identical to the standard errors for the number of digits reported.

⁸ For example, let Z take the values $+1$ or -1 with probability one-half. Consider a security which pays $+\$1$ if $Z = 1$ and $-\$x$ if $Z \neq 1$. The expected payoff of the security is $(1 - x)/2$. To estimate this expected payoff by Monte Carlo simulation, draw n samples Z_i according to the prescribed distribution. Then use equation (8) to define $\tilde{Z}_i$'s which match the first moment. For almost all samples for any large n , the estimated expected payoff is $-x$ and the bias is $(1 + x)/2$. This bias does not decrease as n increases. Care must be taken when using equation (8) or (9) when the support of the random variable is not the entire real line. For example, applying (8) or (9) to uniform or exponential random variables could cause the transformed values to fall outside of the relevant domain.

⁹ This is equivalent to enforcing put-call parity.

Table 1. *Standard errors for European call options.*

σ	S_0/K	No variance reduction	MM1 Equation (8)	MM2 Equation (9)	MM1 Equation (10)	MM2 Equation (11)
0.2	0.9	0.24	0.19	0.11	0.19	0.09
	1.0	0.62	0.29	0.09	0.26	0.10
	1.1	0.93	0.19	0.09	0.15	0.11
0.4	0.9	0.80	0.55	0.24	0.51	0.17
	1.0	1.22	0.66	0.19	0.56	0.23
	1.1	1.61	0.63	0.17	0.48	0.28
0.6	0.9	1.40	0.95	0.38	0.84	0.28
	1.0	1.93	1.10	0.31	0.91	0.39
	1.1	2.38	1.13	0.25	0.85	0.49

All results are based on $n = 100$ simulation trials. The option parameters are: $K = 100$, $r = 0.10$, $T = 0.2$, with S_0 and σ varying as indicated. Standard error estimates are based on $m = 10\,000$ simulations.

The results in Table 1 show that matching two moments can reduce the simulation error by a factor ranging from 2 to 10. Matching two moments dominates matching one moment, but there is not a clear choice between transforming the original standard normals using (9) or the terminal stock prices using (11). Further computational results, not included in Table 1, indicate that the improvement factor with moment matching is *essentially constant* as n increases. This may seem counterintuitive, since the moment matching adjustments converge to zero as n increases. But the progressively smaller adjustments are equally important in reducing the estimation error as the number of simulation trials increases. For example, the standard error for $n = 10\,000$ simulation trials is one-tenth of the corresponding number for $n = 100$ reported in Table 1.

The moment matching method can be extended to match covariances. For options that depend on multiple assets, the entire covariance structure is typically a simulation input. Barraquand (1995) suggests a method to match the entire covariance structure and reports error reduction factors ranging from two to several hundred for this method applied to pricing options on the maximum of k assets.

The moment matching procedure could be applied to matching higher order moments as well. In addition to different methods for transforming random outcomes to match specified moments, additional points could be added as another way to match moments.

Whenever a moment is known, it can be used as a control rather than for moment matching. In an appendix, we give a theoretical argument favoring the use of moments as controls rather than for matching.

2.5 Stratified and Latin hypercube sampling

Like many variance reduction techniques, stratified sampling seeks to make the inputs to simulation more regular than random inputs. In particular, it forces certain empirical probabilities to match theoretical probabilities, just as moment matching forces empirical moments to match theoretical moments.

Consider, for example, the generation of 100 normal random variates as inputs to a simulation. The empirical distribution of an independent sample $Z_1, \dots, Z_{100}$ will look only roughly like the normal density; the tails of the distribution – often the most important part – will inevitably be underrepresented. Stratified sampling can be used to force exactly one observation to lie between the $(i - 1)^{\text{th}}$ and i^{th} percentile, $i = 1, \dots, 100$, and thus produce a better match to the normal distribution. One way to implement this generates 100 independent random variates $U_1, \dots, U_{100}$, uniform on $[0, 1]$ and set $\tilde{Z}_i = N^{-1}((i + U_i - 1)/100)$, $i = 1, \dots, 100$, where N^{-1} is the inverse of the cumulative normal distribution. This works because $(i + U_i - 1)/100$ falls between the $(i - 1)^{\text{th}}$ and i^{th} percentiles of the uniform distribution, and percentiles are preserved by the inverse transform.

Of course, $\tilde{Z}_1, \dots, \tilde{Z}_{100}$ are highly dependent, complicating the estimation of standard errors. Computing confidence intervals with stratified sampling typically requires batching the runs. For example, with a budget of 100 000 replications we might run 100 independent stratified samples each of size 1000, rather than a single stratified sample of size 100 000. To estimate standard errors we must therefore sacrifice some variance reduction, just as with moment matching.

In principle, this approach applies in arbitrary dimensions. To generate a stratified sample from the d -dimensional unit hypercube, with n strata in each coordinate, we could generate a sequence of vectors $U_j = (U_j^{(1)}, \dots, U_j^{(d)})$, $j = 1, 2, \dots$, and then set

$$V_j = \frac{U_j + (i_1, \dots, i_d)}{n}, \quad i_k = 0, \dots, n - 1, \quad k = 1, \dots, d.$$

Exactly one V_j will lie in each of the n^d cubes defined by the product of the n strata in each coordinate.

The difficulty in high dimensions is that generating even a single stratified sample of size n^d may be prohibitive unless n is very small. *Latin hypercube sampling* can be viewed as a way of randomly sampling n points of a stratified sample while preserving some of the regularity from stratification. The method was introduced by McKay, Conover, and Beckman (1979) and further analyzed in Stein (1987). It works as follows. Let $\pi_1, \dots, \pi_d$ be independent random permutations of $\{1, \dots, n\}$, each uniformly distributed over all $n!$ possible permutations. Set

$$V_j^{(k)} = \frac{U_j^{(k)} + \pi_k(j) - 1}{n}, \quad k = 1, \dots, d, \quad j = 1, \dots, n.$$

The randomization ensures that each vector V_j is uniformly distributed over the d -dimensional hypercube. At the same time, the coordinates are perfectly stratified in the sense that exactly one of $V_1^{(k)}, \dots, V_n^{(k)}$ falls between $(j-1)/n$ and j/n , $j = 1, \dots, n$, for each dimension $k = 1, \dots, d$. As before, the dependence introduced by this method implies that standard errors can be estimated only through batching.

These methods can be viewed as part of a hierarchy of methods introducing additional levels of regularity in inputs at the expense of complicating the estimation of errors. Some, like stratified sampling, fix the size of the sample while others leave flexibility. The extremes of this hierarchy are straightforward Monte Carlo (completely random) and the low-discrepancy methods (completely deterministic) discussed in Section 3. Owen (1995a, 1995b) discusses these and other methods and introduces a hybrid that combines the regularity of low-discrepancy methods with the simple error estimation of standard Monte Carlo. Shaw (1995) uses an extension proposed by Stein (1987) to handle dependent inputs in a novel approach to estimating value at risk.

2.6 Some numerical comparisons

The variance reduction methods discussed thus far are fairly generic, in the sense that they do not rely on the detailed structure of the security to be priced. This contrasts with the remaining two methods that we discuss – importance sampling and conditional Monte Carlo. These methods must be carefully tailored to each application. It therefore seems appropriate to digress briefly into a numerical comparison of the generic methods on some option pricing problems.

We first examine the performance of these methods in pricing Asian options. The payoff of a discretely sampled arithmetic average Asian option is $\max(\bar{S} - K, 0)$, where $\bar{S} = \sum_{i=1}^k S_i/k$, S_i is the asset price at time $t_i = iT/k$, and T is the option maturity. The value of the option is $E[e^{-rT} \max(\bar{S} - K, 0)]$. There is no easily evaluated closed-form expression for this option value. Various formulas to approximate the Asian option price have been developed, but simulation is usually used to test the accuracy of the approximations.

For this Asian option, k random numbers are needed to simulate one option payoff, and nk random numbers are needed in total. Moment matching (MM2, for two moments) was applied k times to the n numbers used to generate each S_i at time t_i . Latin hypercube sampling (LHS) was applied to sample n points from the k -dimensional unit cube. The discretely sampled geometric average Asian price was used as a control variate (see Turnbull and Wakeman 1991 for a closed-form solution for this price). Results appear in Table 2.

The results in Table 2 indicate that matching two moments can reduce the simulation error by a factor ranging from 1 to 10. Using the geometric average Asian

Table 2. *Standard errors for arithmetic average Asian options.*

σ	K/S_0	No variance reduction	Antithetic method	Control variate	MM2	LHS
0.2	0.9	0.053	0.052	0.003	0.048	0.049
	1.0	0.344	0.231	0.004	0.162	0.161
	1.1	0.566	0.068	0.006	0.052	0.058
0.4	0.9	0.308	0.297	0.014	0.240	0.248
	1.0	0.694	0.506	0.017	0.352	0.354
	1.1	1.017	0.388	0.021	0.281	0.289
0.6	0.9	0.632	0.583	0.032	0.451	0.455
	1.0	1.052	0.817	0.038	0.566	0.578
	1.1	1.443	0.759	0.047	0.539	0.560

All results are based on $n = 100$ simulation trials with $k = 50$ prices in the average. The option parameters are: $K = 100$, $r = 0.10$, $T = 0.2$, with S_0 and σ varying as indicated. Standard error estimates based on $m = 10\,000$ simulations. The geometric average Asian option is used as the control variate. Moment matching (MM2) was applied to the i^{th} price in the average, $i = 1, \dots, 5$, across replications.

option price as a control variate reduces error by a factor ranging from 20 to 100, and is consistently the most effective method. LHS and MM2 perform similarly. Antithetics are consistently dominated by the other methods.

Next we compare these variance reduction techniques in pricing down-and-out call options with discrete barriers. The payoff of this option at expiration is the standard call option payoff if the asset price S_i exceeds the barrier H at all times $t_i = iT/k$, $i = 1, \dots, k$, otherwise the payoff is zero. The option is knocked out if $S_i \leq H$ at any time t_i . As a control we use the Black–Scholes price of a standard call. Moment matching and LHS are implemented as with the Asian option. Results are given in Table 3. These are consistent with the pattern in Table 2, except that the superiority of the control variate method is less pronounced.

Although it is always risky to draw conclusions from limited numerical evidence, we suggest the following broad conclusions. The antithetic method is easy to implement, but often leads to only modest error reductions. Moment matching is similarly easy to implement and often leads to significant error reductions, but the error estimation is more difficult and bias is a potential problem. LHS suffers from the same error estimation difficulty but does not introduce bias. The control variate technique can lead to very substantial error reductions, but its effectiveness hinges on finding a good control for each problem.

Table 3. *Standard errors for down-and-out call options with discrete barriers.*

σ	K/S_0	No variance reduction	Antithetic method	Control variate	MM2	LHS
0.2	0.9	0.96	0.44	0.37	0.43	0.39
	1.0	0.62	0.44	0.13	0.31	0.30
	1.1	0.30	0.28	0.03	0.22	0.22
0.4	0.9	1.59	1.15	0.73	0.95	0.88
	1.0	1.22	1.00	0.45	0.76	0.74
	1.1	0.88	0.82	0.26	0.61	0.61
0.6	0.9	2.19	1.83	1.07	1.44	1.36
	1.0	1.86	1.62	0.80	1.25	1.23
	1.1	1.54	1.40	0.58	1.09	1.09

All results are based on $n = 100$ simulation trials. There are $k = 5$ points in the discrete barrier at 95. The other option parameters are: $S_0 = 100$, $r = 0.10$, $T = 0.2$, with K and σ varying as indicated. Standard error estimates are based on $m = 10\,000$ simulations. The standard European call option (Black–Scholes formula) is used as the control variate. Moment matching (MM2) was applied to the i^{th} return, $i = 1, \dots, 5$, across replications.

2.7 Importance sampling

This technique builds on the observation that an expectation under one probability measure can be expressed as an expectation under another through the use of a *likelihood ratio* or *Radon–Nikodym derivative*. This idea is familiar in finance because it underlies the representation of prices as expectations under a martingale measure. In Monte Carlo, the change of measure is used to try to obtain a more efficient estimator. We present some examples using this technique; for general background see Bratley et al. (1987) or Hammersley and Handscomb (1964).

As a simple example, consider the evaluation of the Black–Scholes price of a call option – i.e., the computation of $e^{-rT} E[\max\{S_T - K, 0\}]$ with S_T as in (2). A straightforward approach generates samples of the terminal value S_T consistent with a geometric Brownian motion having drift r and volatility σ , just as in (2). But we are in fact free to generate S_T consistent with any other drift μ , provided we weight the result with a likelihood ratio. For emphasis, we subscript the expectation operator with the drift parameter. Then

$$E_r[\max\{S_T - K, 0\}] = E_\mu[\max\{S_T - K, 0\}L],$$

where the likelihood ratio L is the ratio of the lognormal densities with parameters

r and μ evaluated at S_T , given by

$$L = \left(\frac{S_T}{S_0} \right)^{\frac{r-\mu}{\sigma^2}} \exp\left(\frac{(\mu^2 - r^2)T}{2\sigma^2} \right).$$

Indeed, S_T need not even be sampled from a lognormal distribution. The only requirement is that the support of the importance sampling measure contain the support of the original measure so that the likelihood ratio is well-defined; this is an *absolute continuity* requirement. In the example above, this means that any distribution for S_T whose support includes $(0, \infty)$ is admissible.

Ideally, one would like to choose the importance sampling distribution to reduce variance. In the example above, one obtains a *zero-variance* estimator by sampling S_T from the density

$$f(x) = c^{-1} \max\{x - K, 0\} e^{-rT} g(x),$$

where g is the (lognormal) density of S_T and c is a normalizing constant that makes f integrate to 1. The difficulty is that c is the Black–Scholes price itself, so this method requires knowledge of the solution for its implementation. Nevertheless, it gives some indication of the potential gain from importance sampling.

Reider (1993) has investigated the impact of importance sampling based on a change of drift and volatility. (Changing the volatility is consistent with absolute continuity in a discrete-time approximation of a diffusion though not in the continuous-time limit.) He finds that choosing the importance sampling distribution to have higher drift and volatility provides substantial variance reduction in pricing deep out-of-the-money options. He also investigates the combination of importance sampling with antithetic variates and control variates, and the use of put-call parity for indirect estimation. Nielsen (1994) has explored some related importance sampling ideas in sampling from a binomial tree.

Andersen (1995) has developed a powerful application of importance sampling for simulating interest rates and has applied it to nonlinear stochastic differential equation models. We briefly describe his approach. Let r_t be the instantaneous short rate described, e.g., by a diffusion model. Then

$$B(T) = E \left[\exp \left(- \int_0^T r_t dt \right) \right]$$

is the price today of a zero-coupon bond with face value \$1, maturing at time T . In, for example, the Cox–Ingersoll–Ross and Vasicek models,¹⁰ $B(T)$ is available

¹⁰ See, e.g., Hull (1993, Chapter 15) for background on these models.

in closed form. We may therefore define a new probability measure $\bar{P}$ by setting

$$\bar{P}(A) = E \left[\exp \left(- \int_0^T r_t dt - \log B(T) \right) \mathbf{1}_A \right]$$

for any event A , where $\mathbf{1}_A$ denotes the indicator of the event A . Let $\bar{E}$ denote expectation with respect to $\bar{P}$. Then for any random variable X , $E[X] = \bar{E}[X L_T]$ where the likelihood ratio L_T is given by

$$L_T = \exp \left(\int_0^T r_t dt + \log B(T) \right).$$

In particular, if we take $X = \exp(-\int_0^T r_t dt)$, we know that $E[X] = B(T)$ and therefore $B(T)$ is the expectation under $\bar{E}$ of $X L_T$; i.e., of

$$\exp \left(- \int_0^T r_t dt \right) \times \exp \left(\int_0^T r_t dt + \log B(T) \right).$$

But this simplifies to $B(T)$ itself, meaning that we obtain a zero-variance estimator of the bond price by switching to the new probability measure. Moreover, Andersen shows that sample paths of r_t can be generated under $\bar{P}$ simply by applying a change of drift to the original process.

As described above, the method would appear to require knowledge of the solution for its implementation. Nevertheless, the method has two important applications. The first is in the pricing of contingent claims. Because $\bar{P}$ eliminates the variance of bond prices, it should be effective in reducing variance for pricing, e.g., European bond options expiring at time T . Andersen's numerical results bear this out. A second application is in the pricing of bond models with no closed-form solutions: Andersen's results show that the change of drift derived from a tractable model (like CIR or Vasicek) remains effective when applied to an intractable model, and this significantly expands the scope of the method.

Importance sampling is frequently used to make rare events less rare; this is already suggested in Reider's (1994) application to out-of-the-money options. Our next example further highlights this aspect through a new application to barrier options. We consider a knock-in option far from the barrier and use importance sampling to increase the probability of a payout.

Suppose the barrier is monitored at discrete times $n\Delta t$, $n = 0, 1, \dots, m$, with $\Delta T = T/m$. Set the barrier at $H = S_0 e^{-b}$ and the strike at $K = S_0 e^c$, with $b, c > 0$. A down-and-in call pays $S_T - K$ at time T if $S_T > K$ and $S_{n\Delta t} < H$ for some $n = 1, \dots, m$. We can write the price of the underlying at monitoring instants as

$$S_{n\Delta t} = S_0 e^{U_n}, \quad U_n = \sum_{i=1}^n X_i,$$

with the X_i i.i.d. normal having mean $(r - \frac{1}{2}\sigma^2)\Delta t$ and variance $\sigma^2\Delta t$. Let τ be the first time U_n drops below $-b$; then the probability of a payout is $P(\tau < m, U_m > c)$. If b and c are large, this probability is small, and most simulation runs return zero. Through importance sampling, we can increase this probability and thus get more information out of each run.

Consider alternative probability measures P_{μ_1, μ_2} that give U_n a drift of $\mu_1\Delta t$ until τ and then switch the drift to $\mu_2\Delta t$. Intuitively, we would like to make $\mu_1 < 0$ to drive the asset price to the barrier and then make $\mu_2 > 0$ to drive it above the strike. For any μ_1, μ_2 , we have

$$P(\tau < m, U_m > c) = E_{\mu_1, \mu_2}[L_{\mu_1, \mu_2} \mathbf{1}_{\{\tau < m, U_m > c\}}].$$

The likelihood ratio is given by

$$L_{\mu_1, \mu_2} = \exp(-\theta_1 U_\tau + \psi(\theta_1)\tau - \theta_2(U_m - U_\tau) + \psi(\theta_2)(m - \tau)),$$

where $\theta_i = (\mu_i - r + \frac{1}{2}\sigma^2)/\sigma^2$, $i = 1, 2$, and $\psi(\theta) = (r - \frac{1}{2}\sigma^2)\Delta t\theta + \frac{1}{2}\sigma^2\Delta t\theta^2$. This follows from algebraic simplification of the product of the ratios of the densities of the X_i under the original and new means.

It remains to choose μ_1, μ_2 . Intuitively, most of the variability in L_{μ_1, μ_2} comes from τ (the time of the barrier crossing): for large b, c , in the event of a payout we expect to have $U_\tau \approx -b$ and $U_m \approx c$ so these terms should contribute less variability. If we choose μ_1, μ_2 so that $\psi(\theta_1) = \psi(\theta_2)$, the likelihood ratio simplifies to

$$L_{\mu_1, \mu_2} = \exp(-(\theta_1 - \theta_2)U_\tau - \theta_2 U_m + m\psi(\theta_2)),$$

which depends on τ only through $U_\tau \approx -b$. The condition $\psi(\theta_1) = \psi(\theta_2)$ translates to $\mu_1 = -\mu_2 \equiv -\mu$, so it only remains to choose this drift parameter. We choose it so that the time to traverse the straight line path from 0 to $-b$ and then to c at rate μ equals the number of steps m :

$$\frac{b}{\mu\Delta t} + \frac{(b+c)}{\mu\Delta t} = m;$$

i.e., $\mu = (2b+c)/T$. Interestingly, this change of drift does not depend on the original mean increment $(r - \frac{1}{2}\sigma^2)\Delta t$.

Table 4 illustrates the performance of this method. The computational effort with and without importance sampling is essentially the same, so the efficiency improvement is just the ratio of the variances. The improvement varies widely but shows the potential for dramatic gains from importance sampling, particularly when the barrier is far from the current price of the underlying.¹¹

¹¹ The standard errors in the table are all quite small, but so are the associated option values. Hence, the *relative* error without importance sampling is quite significant.

Table 4. *Standard errors for down-and-in calls: importance sampling.*

H	K	No variance reduction	Importance sampling	Efficiency ratio
92	100	0.003 09	0.000 69	20
92	105	0.001 29	0.000 14	85
88	96	0.001 10	0.000 11	96
85	90	0.000 84	0.000 08	116
92	105	0.014 18	0.005 41	7
85	105	0.003 28	0.000 38	75
75	96	0.000 30	0.000 01	1124
75	85	0.001 48	0.000 10	222

All results are based on $n = 100\,000$ simulation trials. The parameters are: $S_0 = 95$, $\sigma = 0.15$, and $r = 0.05$, with the barrier H and strike K varying as indicated. The first four cases have $T = 0.25$ and $m = 50$; the last four have $T = 1$ and $m = 250$.

In recent work, Andersen and Brotherton-Ratcliffe (1996) and Beaglehole, Dybvig, Zhou (1997) show how to eliminate the bias caused by using a simulation at a discrete set of times to price continuous options on extrema, e.g., barrier or lookback options.

2.8 Conditional Monte Carlo

This approach to efficiency improvement exploits the variance reducing property of conditional expectation: for any random variables X and Y , $\text{Var}[E[X|Y]] \leq \text{Var}[X]$, with strict inequality except in trivial cases.¹² In replacing an estimator by its conditional expectation we reduce variance essentially because we are doing part of the integration analytically and leaving less to be done by Monte Carlo.

Hull and White (1987) use this idea to price options with stochastic volatilities. Consider a model in which an asset price and its volatility evolve as follows:

$$\begin{aligned} dS &= rS dt + \nu S dW_1 \\ d\nu^2 &= \alpha \nu^2 dt + \xi \nu^2 dW_2, \end{aligned}$$

with W_1, W_2 independent. Suppose we want to price a standard European call on S . A straightforward approach simulates sample paths of ν and S up to time T and averages $\max\{S_T - K, 0\}$ over all paths. An alternative notes that, conditional on the path of ν_t in $[0, T]$, the asset price S_t may be treated as having a time-varying

¹² This is a direct consequence of Jensen's inequality for conditional expectations.

but deterministic volatility. Thus, conditional on the volatility path, the option can be priced by the Black–Scholes formula:

$$e^{-rT} E[\max\{S_T - K, 0\} | v_t, 0 \leq t \leq T] = \text{BS}(S_0, K, r, T, \sqrt{V_T}),$$

where

$$V_T = \frac{1}{T} \int_0^T v_t^2 dt$$

is the average squared volatility over the path, and $\text{BS}(S, K, T, r, \sigma)$ is the Black–Scholes price of a call with constant volatility σ and the other parameters as indicated. Using this conditional expectation as the estimator is sure to reduce variance and may even reduce computational effort since it obviates simulation of S . It is worth emphasizing that both straightforward Monte Carlo and conditional Monte Carlo would have to be applied to discrete-time approximations of the continuous processes above. Also, the applicability of conditional Monte Carlo in this setting relies on the fact that the evolution of the asset price does not influence the volatility path. See Willard (1997) for an extension to the case of correlated W_1 and W_2 .

As a further illustration of the use of conditional Monte Carlo, we give a new illustration in the pricing of a down-and-in call with a discretely monitored barrier. Let $0 = t_0 < t_1 < \dots < t_m = T$ be the monitoring instants and S_{t_i} the price of the underlying at the i^{th} such instant. The option price is $E[e^{-rT} \max\{S_T - K, 0\} \mathbf{1}_{\{\tau_H \leq T\}}]$, where H is the barrier and τ_H is the first monitoring time at which the barrier is breached.

Straightforward simulation generates paths of the underlying and evaluates the estimator

$$e^{-rT} \max\{S_T - K, 0\} \mathbf{1}_{\{\tau_H \leq T\}}.$$

Our first alternative conditions on $\{S_0, \dots, S_{\tau_H}\}$, the path of the underlying until the barrier crossing; i.e.,

$$\begin{aligned} E[e^{-rT} \max\{S_T - K, 0\} \mathbf{1}_{\{\tau_H \leq T\}}] \\ &= e^{-rT} E[E[\max\{S_T - K, 0\} \mathbf{1}_{\{\tau_H \leq T\}} | S_0, \dots, S_{\tau_H}]] \\ &= e^{-rT} E[\text{BS}(S_{\tau_H}, K, r, T - \tau_H, \sigma) \mathbf{1}_{\{\tau_H \leq T\}}]. \end{aligned}$$

This yields the estimator

$$\text{CMC}_1 = e^{-rT} \text{BS}(S_{\tau_H}, K, r, T - \tau_H, \sigma) \mathbf{1}_{\{\tau_H \leq T\}}$$

This says: simulate until the barrier is crossed or the option expires; if the barrier was crossed, return the Black–Scholes price starting from price S_{τ_H} with maturity $T - \tau_H$.

Our second alternative conditions one step earlier, at each monitoring instant evaluating the probability that the barrier will be breached for the first time at the next monitoring instant:

$$\begin{aligned}
 E[e^{-rT} \max(S_T - K, 0) \mathbf{1}_{\{\tau_H \leq T\}}] &= e^{-rT} E \left[\max\{S_T - K, 0\} \sum_{n=1}^m \mathbf{1}_{\{\tau_H = t_n\}} \right] \\
 &= e^{-rT} E \left[\sum_{n=1}^m E[\max\{S_T - K, 0\} \mathbf{1}_{\{\tau_H = t_n\}} | S_{t_0}, \dots, S_{t_{n-1}}] \right] \\
 &= e^{-rT} E \left[\sum_{n=0}^{\tau_H-1} \text{BS2}(S_{t_n}, K, H, r, t_{n+1} - t_n, T - t_n, \sigma) \right]
 \end{aligned}$$

where $\text{BS2}(S, K, H, r, t, T, \sigma)$ is the price of a down-and-in call that knocks in only if the underlying is below H at time t . We thus arrive at the estimator

$$\text{CMC}_2 = e^{-rT} \sum_{n=0}^{\tau_H-1} \text{BS2}(S_{t_n}, K, H, r, t_{n+1} - t_n, T - t_n, \sigma),$$

with

$$\text{BS2}(S, K, H, r, t, T, \sigma) = SN_2(a_1, b_1, \rho) - e^{-rT} K N_2(a_2, b_2, \rho)$$

where $\rho = -\sqrt{t/T}$, N_2 is the bivariate cumulative normal distribution with correlation ρ , and

$$\begin{aligned}
 a_1 &= \frac{\log(S/K) + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}, & a_2 &= a_1 - \sigma\sqrt{T} \\
 b_1 &= \frac{\log(H/S) - (r + \frac{1}{2}\sigma^2)t}{\sigma\sqrt{t}}, & b_2 &= b_1 + \sigma\sqrt{t}.
 \end{aligned}$$

(The derivation of this formula is fairly standard and therefore omitted.) The CMC_2 estimator can be expected to have lower variance than the CMC_1 estimator because it conditions on *less* information and thus does *more* integration analytically. In fact, CMC_2 is not a conditional Monte Carlo estimator in the strict sense because it conditions on different information at different times, making it more precisely a *filtered* Monte Carlo estimator in the sense of Glasserman (1996).

Because the two estimators above have the same expectation, their difference has mean 0 and can be used as a control variate to form a further estimator

$$\text{CMC}' = \text{CMC}_1 + \beta(\text{CMC}_2 - \text{CMC}_1).$$

With β optimized, this has lower variance than either individual estimator.

Numerical results appear in Table 5. As expected, each level of conditioning further reduces variance, and the combined estimator achieves the lowest standard

Table 5. Comparison of CMC estimators for down-and-in call.

Method	Standard Error (s)	Computation Time (t)	$s\sqrt{t}$
Base	0.108	0.133	0.039
CMC ₁	0.034	0.117	0.012
CMC ₂	0.021	3.233	0.038
CMC'	0.014	3.367	0.026

Results based on $n = 10\,000$ replications with $\sigma = 0.4$, $r = 0.10$, $S_0 = K = 100$, $H = 95$, $T = 0.5$, and 10 equally spaced monitoring times.

error of all. However, repeated evaluation of the function BS2 turns out to be time-consuming, making CMC₁ overall the most efficient estimator.

3 Low-discrepancy sequences

For complex problems the performance of the basic Monte Carlo approach may be rather unsatisfactory because the error is $O(1/\sqrt{n})$. We can sometimes improve convergence by using pre-selected deterministic points to evaluate the integral. The accuracy of this approach depends on the extent to which these deterministic points are evenly dispersed throughout the domain of integration. *Discrepancy* measures the extent to which the points are evenly dispersed throughout a region: the more evenly dispersed the points are the lower the discrepancy. Low-discrepancy sequences are often called quasi-random sequences even though they are not at all random.¹³ We shall use both terms in this paper.

Low-discrepancy methods have recently been used to tackle a number of problems in finance. These applications are more fully described in papers by Birge (1994), Joy, Boyle, and Tan (1996) and Paskov and Traub (1995); the use of quasi-Monte Carlo is also proposed in Cheyette (1992). In this section we describe how the approach works and review some of the recent applications. The book by Press et al. (1992) provides an intuitive introduction to low-discrepancy sequences and quasi-Monte Carlo methods. Spanier and Maize (1994) provide a recent overview of quasi-random methods and how they can be used to evaluate integrals with medium sized samples. Niederreiter (1992) and Tezuka (1995) provide in-depth analyses of low-discrepancy sequences. Moskowitz and Caflisch (1996) discuss recent developments in improving the convergence of quasi-random Monte Carlo methods. In earlier work, Haselgrove (1961) describes a method for multi-

¹³ Thus the name quasi-random is very misleading since these sequences are deterministic. However, it seems to be sanctioned by usage.

variate integration that can be applied to security pricing. Haselgrove's method is developed for problems of eight dimensions or less and our numerical experiments suggest that it is competitive with the low-discrepancy sequences investigated in this section for problems of this size.

The basic idea behind the approach is quite intuitive and is readily explained in the one-dimensional case. Suppose we wish to integrate a function $f(x)$ over the interval $[0, 1]$ using a sequence of n points. Rather than pick a random sequence suppose we pick a deterministic sequence of points that are, in some sense, evenly distributed. With this choice, the accuracy of the estimate will be higher than that obtained using the crude Monte Carlo approach. If we use an equally spaced grid we obtain the trapezoidal method of numerical integration which has an error of $O(n^{-1})$. However, the more challenging task is to evaluate multi-dimensional integrals. Without loss of generality we can assume that the domain of integration is contained in the d -dimensional unit hypercube. The advantages of the uniformly spaced grid in the one-dimensional case do not carry over to higher dimensions. The principal reason is that the error bound for the d -dimensional trapezoidal rule is $O(n^{-2/d})$. In addition, if we use an evenly spaced Cartesian grid, we would have to decide the number of points in advance to achieve uniformity. This is restrictive because, in numerical applications, we would like to be able to add points sequentially until some termination criterion is met.

Low-discrepancy sequences have the property that as successive points are added the entire sequence of points still remains more or less evenly dispersed throughout the region. Niederreiter (1992) gives a detailed analysis of the discrepancy of a sequence. Here, we just briefly recall the definition. Suppose we have a sequence of n points $\{x_1, x_2, \dots, x_n\}$ in the d -dimensional half-open unit cube, $I^d = [0, 1)^d$ and a subset J of I^d . We define

$$D(J; n) = \frac{A(J; n)}{n} - V(J),$$

where $A(J; n)$ is the number of k , $1 \leq k \leq n$, with $x_k \in J$ and $V(J)$ is the volume of J . The discrepancy, D_n , of the sequence is defined to be the supremum of $|D(J; n)|$ over all J . The star discrepancy D_n^* , is obtained by taking the supremum over sets J of the form

$$\prod_{i=1}^d [0, u_i).$$

In the one-dimensional case there is a simple explicit form for the (star)¹⁴ discrepancy of a sequence of n points. If we label the points so that, $0 \leq x_1 \leq \dots \leq$

¹⁴ For the rest of the paper we simply use the term discrepancy rather than star discrepancy to refer to D_n^* .

$x_n \leq 1$, then the discrepancy of this sequence is

$$D_n^* = \frac{1}{2n} + \max_{k=1, \dots, n} \left| x_k - \frac{2k-1}{2n} \right|.$$

We can see that the star discrepancy is at least $1/(2n)$ and that the lowest value is attained when

$$x_k = \frac{2k-1}{2n}, \quad 1 \leq k \leq n.$$

In higher dimensions there is no simple form for the discrepancy of a sequence.

There are several examples of low-discrepancy sequences, including the sequences proposed by Halton (1960), Sobol' (1967), Faure (1982), and Niederreiter (1988).¹⁵ For these sequences the asymptotic form of the star discrepancy has been shown to be

$$D_n^* = O\left(\frac{(\log n)^d}{n}\right).$$

This bound for the discrepancy involves a constant which in general depends on the dimension d of the sequence. These constants are very difficult to estimate accurately in high dimensions. For large values of d the constants “are often ridiculously large for reasonable values of n ” according to Spanier and Maize (1994, p. 23). Furthermore for high dimensions it may take a long time before the discrepancy reaches its asymptotic level. Morokoff and Caflisch (1995) note that for intermediate values of n the discrepancy may be $O(\sqrt{n})$. They suggest that the transition to $O(n^{-1}(\log n)^d)$ occurs at around values of $n = e^d$. For large d this will be an enormous number.

The error in numerical integration using a low-discrepancy sequence admits a deterministic bound. The bound reflects both the discrepancy of the sequence of points used to evaluate the integral as well as the regularity of the function. The result is contained in the following theorem.

Theorem (Koksma–Hlawka) *Let $I^d = [0, 1]^d$ and let f have bounded variation $V(f)$ on $[0, 1]^d$ in the Hardy–Krause¹⁶ sense. Then for any $x_1, x_2, \dots, x_n \in I^d$ we have*

$$\left| \frac{1}{n} \sum_{k=1}^n f(x_k) - \int_{I^d} f(u) du \right| \leq V(f) D_n^*.$$

¹⁵ Interestingly, linear congruential generators – frequently used to generate the pseudo-random numbers that drive ordinary Monte Carlo – produce sets of points with low-discrepancy over the entire period of the generator; see Niederreiter (1976). This suggests the possibility of choosing such a generator with period roughly equal to the total number of points required as a type of quasi-Monte Carlo method. In ordinary Monte Carlo, one prefers instead that the period be many orders of magnitude larger than the number of points required. We thank Peter Hellekalek of the University of Salzburg for this observation.

¹⁶ For a more complete discussion of the Hardy–Krause definition of variation and details on this theorem see Niederreiter (1992).

The error bound provided by this theorem, while it is of theoretical interest, is of little help in most practical situations. The theoretical bound normally overestimates the actual error by a wide margin and $V(f)$ may be difficult to evaluate or even approximate. We have noted that the constants buried in the bounds for the discrepancy are large. Another reason for the coarseness of the bound is that the Koksma–Hlawka theorem does not reflect additional smoothness in f . Intuitively we would expect the approximation to be better as f becomes smoother. In finance applications the payoffs are normally continuous functions of the variables (with some important exceptions – payoffs on digital and barrier options are discontinuous), but may not be sufficiently smooth to have finite variation because of functions like “max” embedded in the payoffs. Hlawka (1971) provides an alternative bound under weaker smoothness requirements.

To date, studies using low-discrepancy sequences in finance applications find that the errors produced are substantially lower than the corresponding errors generated by crude Monte Carlo. Joy, Boyle, and Tan (1996) used Faure sequences to price several complex derivative securities. They found that the quasi-Monte Carlo approach resulted in significantly smaller errors than the standard Monte Carlo approach. They confirmed that the actual error bound (for cases in which it could be computed precisely) was dramatically less than the bound computed from the Koksma–Hlawka inequality. Paskov and Traub (1995) used both Sobol’ sequences and Halton sequences to evaluate mortgage-backed security prices. Their work involves the evaluation of integrals with dimensions up to 360; they find that Sobol’ sequences are more efficient than Halton sequences and that the quasi-random approach outperforms the standard Monte Carlo approach for these types of problems.¹⁷ Paskov and Traub’s results stand in contrast to the claim that is sometimes found in the literature¹⁸ that the superiority of low-discrepancy algorithms vanishes for intermediate values of d around 30. Bratley, Fox, and Niederreiter (1992) conducted practical numerical experiments using low-discrepancy sequences and conclude that standard Monte Carlo is superior to quasi-Monte Carlo for high dimensions, say greater than 12. They used Sobol’ and Niederreiter sequences in their tests. They conclude that in high dimensions, “quasi-Monte Carlo seems to offer no practical advantage over pseudo-Monte Carlo because the discrepancy bound for the former is far larger than $\sqrt{n}$ for $n = 2^{30}$, say.” (In a personal communication, Fox adds that the crossover probably depends a lot on the sequence.) The reason for the difference between this verdict and the results of the finance applications may be that the integrands typically found in finance applica-

¹⁷ Bratley et al. (1992) note that the Niederreiter sequence they tested theoretically beats Sobol’ sequences in dimensions higher than seven.

¹⁸ See, for example, Rensburg and Torrie (1993) or Morokoff and Caflisch (1995).

tions behave better than those used by numerical analysts¹⁹ to compare different algorithms. Another important consideration is that financial applications typically involve discounting, and this may effectively reduce dimensionality; for example, some of the 360 months in the life of a mortgage may have little influence on the value of a mortgage-backed security. Nevertheless, the experience of Bratley et al. (1992) serves as a useful caution against assuming that quasi-Monte Carlo will outperform standard Monte Carlo in all situations.

Some theoretical differences among low-discrepancy sequences can be understood through the concepts of (t, m, s) -nets and (t, s) -sequences; these are discussed in detail in Niederreiter (1992). Briefly, an elementary interval in base b in dimension s is a set of the form

$$\prod_{j=1}^s \left[\frac{a_j}{b^{k_j}}, \frac{a_j + 1}{b^{k_j}} \right),$$

with k_j, a_j nonnegative integers and $a_j < b^{k_j}$. A (t, m, s) -net (with $0 \leq t \leq m$) is a set of b^m points in the s -dimensional hypercube such that every elementary interval of volume b^{t-m} contains b^t points. Speaking loosely, this means that the proportion of points in each sufficiently large box equals the volume of the box. Smaller t implies greater uniformity. An infinite sequence forms a (t, s) -sequence if for all $m \geq t$ certain finite subsequences of length b^m form (t, m, s) -nets in base b . Sobol' points are (t, s) -sequences in base 2 and Faure points are $(0, s)$ sequences in prime bases not less than s . Thus, Faure points achieve the smallest value of t , but at the expense of a large base. A smaller base implies that uniformity holds over shorter subsequences.

An important issue in the use of quasi-Monte Carlo concerns the termination criterion, since the Koksma–Hlawka bound is often of little practical value. Various heuristics are available. Birge (1994) suggests that a rough bound may be obtained by tracking the maximum and minimum values over a period that shows equal numbers of increases and decreases. For instance the criterion could be to stop at the first set of two thousand observations in which the number of increases and decreases are within ten percent of each other. He suggests that the maximum and minimum realized values could be used as bounds on the true value. Fox (1986) suggests that we compare the estimate of the integral based on a sample of $2n$ points with the estimate based on n points and stop if the answer lies within some tolerance level. Paskov and Traub (1995) use a similar termination criterion based

¹⁹ For example, one of the integrals used by Bratley, Fox, and Niederreiter (1992) was

$$\int_0^1 \cdots \int_0^1 \prod_{k=1}^d k \cos(kx_k) dx_1 \cdots dx_d.$$

This integrand is highly periodic for large values of d .

on successive errors: stop when the difference between two consecutive approximations using $10\,000i$, $i = 1, 2, \dots, 1000$, sample points falls below some threshold. Owen (1995a, 1995b) proposes a hybrid of Monte Carlo and low-discrepancy methods which provides error estimates and has good convergence properties. In addition to these approaches, one can also run standard Monte Carlo at the outset and use the probabilistic error term to assess when enough low-discrepancy points have been used in the quasi-random calculation. This benchmarking with standard Monte Carlo would be useful if the same set of calculations were being carried out frequently with only slightly different input values. This situation is common in finance applications. There is often a need to perform the same set of calculations frequently; e.g., the risk analysis of a book of business at the end of each day. In these cases one can conduct experiments to see which sets of low-discrepancy sequences provide the best results. The *right* number of low-discrepancy points could be determined just once at the outset.

Before leaving this section, we should mention some recent advances and new techniques to improve the performance of quasi-random Monte Carlo. Niederreiter and Xing (1996), Tezuka (1994), and Ninomiya and Tezuka (1996) have proposed new low-discrepancy sequences that appear to have the potential to perform substantially better than previous methods. We have noted that the efficiency of quasi-random Monte Carlo improves as the integrand becomes smoother. Moskowitz and Caflisch (1996) illustrate procedures that can be used for this purpose. It is sometimes possible to enhance the performance of quasi-random sequences by reducing the effective dimension of the problem. Moskowitz and Caflisch also indicate how this can be accomplished in the discretization of a Wiener process and in the solution of the Feynman–Kac equation. This is relevant for finance applications since the prices of derivative securities have a Feynman–Kac representation. See Acworth, Broadie, and Glasserman (1997), Berman (1996), and Caflisch, Morokoff, and Owen (1998) for recent work applying low-discrepancy sequences with alternative constructions of Wiener processes. Spanier and Maize (1994) discuss a battery of techniques that can be used to improve the performance of quasi-Monte Carlo methods for relatively small sample sizes.

Next we compare the Monte Carlo method using pseudo-random numbers with the Faure, Halton, and Sobol' low-discrepancy methods.

3.1 Numerical results

For an initial comparison, we test the methods on the problem of pricing a European option on a single underlying asset with the usual Black–Scholes assumptions. In this framework, the Black–Scholes formula can be evaluated to give the true option values in order to compare alternative methods. Rather than using

a single option, we evaluate the methods on a random sample of 500 options. The probability distribution of the parameters is chosen to represent a reasonable range of values in practical applications.²⁰ The error measure that we use is root-mean-squared (RMS) relative error defined by

$$\text{RMS} = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(\frac{\hat{C}_i - C_i}{C_i} \right)^2}, \quad (12)$$

where i is the index of the $m = 500$ options in the test set, C_i is the true option value, and $\hat{C}_i$ is the estimated option value. The results are given in Figure 1.

Figure 1 plots RMS relative error against the number of points, n . The Monte Carlo method (i.e., using pseudo-random numbers) displays the expected $O(1/\sqrt{n})$ convergence: e.g., increasing n by a factor of 100 decreases the RMS error by a factor of 10. The low-discrepancy method using Faure sequences dominates the Monte Carlo method. Indeed, 129 Faure points gives an error lower than 1000 Monte Carlo points. The Sobol' method is the best of the three methods tested. Using 192 Sobol' points gives an error lower than 10 000 Monte Carlo points.

A major consideration in the comparison of methods is the overall computation time, not just the number of points. The Sobol' sequence numbers can be generated significantly faster than Faure numbers (see, e.g., Bratley and Fox 1988) and as fast as most pseudo-random number methods. Hence, in the important RMS error versus computation time comparison, the relative advantage of the Sobol' method increases.

A low-discrepancy sequence will often have additional uniformity properties at certain points in the sequence (see, e.g., Fox 1986 and Bratley and Fox 1988). For example, in the Sobol' sequence the running average returns to 0.5 at the points $n = 2^k - 1$ for $k = 1, 2, \dots$. One might expect that choosing n to be one of these "favorable" points would lead to better option price estimates. For large values of n , the advantage of using favorable points becomes negligible, but for small n the effect can be quite significant. Indeed, in the experiment above, using the Sobol' points 1 through 254 gives an RMS error of 10%, while using the points 1 through 255 gives an RMS error of 4%.²¹ Better results are often obtained by ignoring an initial portion of a low-discrepancy sequence. For example, using the Sobol' points 1 through 63 gives an RMS error of 13%, while using the Sobol' points 64 through 127 gives an RMS error of 2%. In the results in Figure 1, the Sobol' sequence was always started at point 64, so the label 192 in Figure 1 corresponds to the 192 Sobol' points from 64 to 255. Similarly, the Faure sequence was always started at

²⁰ The details of the distribution are given in Broadie and Detemple (1996).

²¹ We take the first point of the Sobol' sequence to be 0.5, not 0.0.

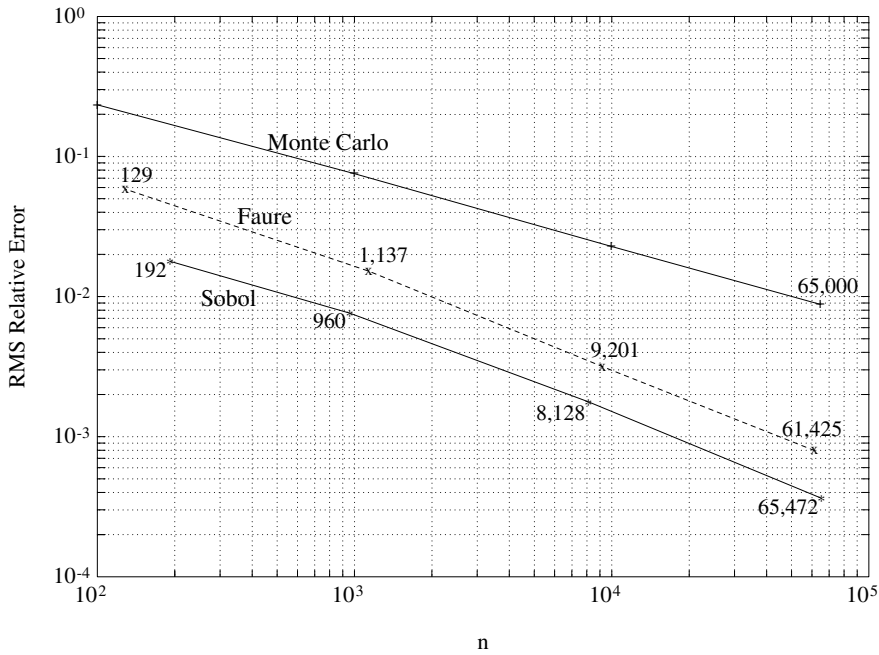


Fig. 1. RMS relative error vs. number of points.

point 16, so the label 129 in Figure 1 corresponds to the 129 Sobol' points from 16 to 144.

3.2 One-dimensional vs. higher dimensional sequences

It is sometimes asserted that low-discrepancy methods can be implemented in existing simulation programs by simply replacing the pseudo-random number generator with a low-discrepancy sequence generator. This naive approach can lead to disastrous results as the following example shows.

Consider pricing a European option on the maximum of two non-dividend paying assets with the parameters: $S_1 = S_2 = K = 100$, $\sigma_1 = \sigma_2 = 0.2$, $\rho = 0.3$, $r = 0.05$, and $T = 1$. Under the usual Black-Scholes assumptions, a formula for the price of the option can be derived (see, e.g., Johnson 1987 or Stulz 1982) and gives a price of 16.442. Running one Monte Carlo simulation with 1000 points (hence 2000 random numbers) gave an estimated price of 16.279 with a standard error of 0.533. Using 2000 one-dimensional low-discrepancy values gave a price estimate of 4.320 using the Sobol' sequence and an estimate of 1.909 using the

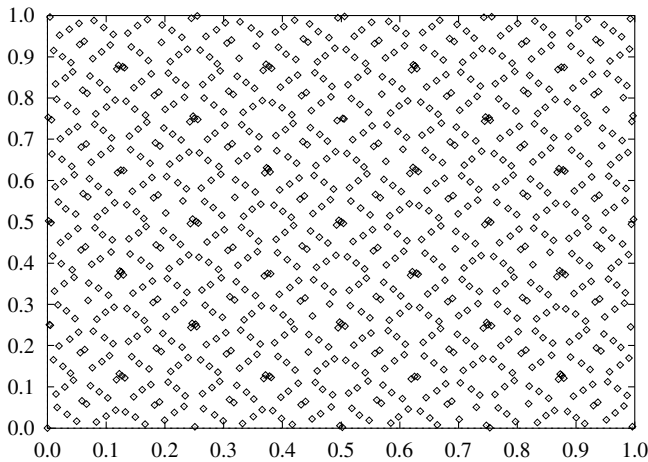


Fig. 2. 1000 two-dimensional Faure points.

Faure sequence (starting at point 16). The cause of the problem can be seen by examining Figures 2–5.

Figures 2 and 3 show 1000 two-dimensional Faure and Sobol’ points, respectively. The figures illustrate how the sequences fill the two-dimensional space in regular but different ways. By contrast, Figures 4 and 5 show 2000 one-dimensional Faure and Sobol’ points, respectively, plotted in two dimensions. The plots are created by taking successive points in the one-dimensional sequence to be the (x, y) coordinates in two-dimensional space. In neither figure are the points filling the two-dimensional space (note that the axes do not extend from 0 to 1) and this explains why the price estimates do not converge to the correct values. Even in the quarter of the unit square where the points fall, the points do not uniformly fill the space. This problem is reminiscent of the well-known “collinearity” or “hyperplane” problem of some pseudo-random number generators, but is even more serious with these low-discrepancy sequences.

A similar problem can occur if a high-dimensional low-discrepancy sequence is used for a problem of low dimension. Figure 6 shows the 49th and 50th dimension of 1000 50-dimensional Faure points. Using the last two dimensions of the 50-dimensional sequence to price a two-dimensional option will give very poor results.

3.3 Higher dimensional test

To test the effect of problem dimension, we price options in dimensions $d = 10, 50$, and 100. We price discretely sampled geometric average Asian options, because the problem dimension is easily varied and a closed form solution for the price

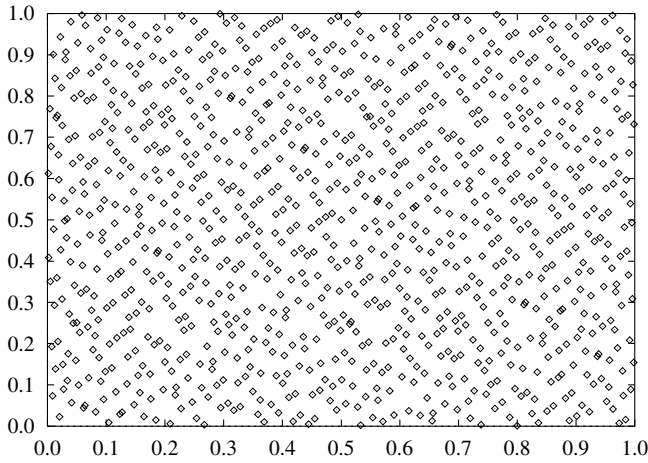


Fig. 3. 1000 two-dimensional Sobol' points.

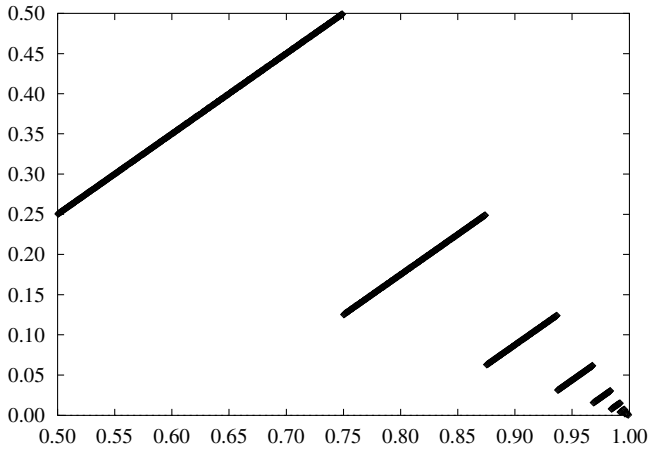


Fig. 4. 2000 one-dimensional Faure points.

is available (see Turnbull and Wakeman 1991). The price of a geometric average Asian option is given by

$$C = E[e^{-rT}(\tilde{S} - K)^+],$$

where $\tilde{S} = (\prod_{i=1}^d S_i)^{1/d}$ and S_i is the asset price at time iT/d .

We test standard Monte Carlo, Monte Carlo with antithetic variates, and the low-discrepancy sequences of Faure, Sobol', and Halton.²² For each dimension, we select 500 option parameters at random, and compute RMS relative error (see

²² We thank Spassimir Paskov and Joseph Traub for providing their code for the Sobol' sequences.

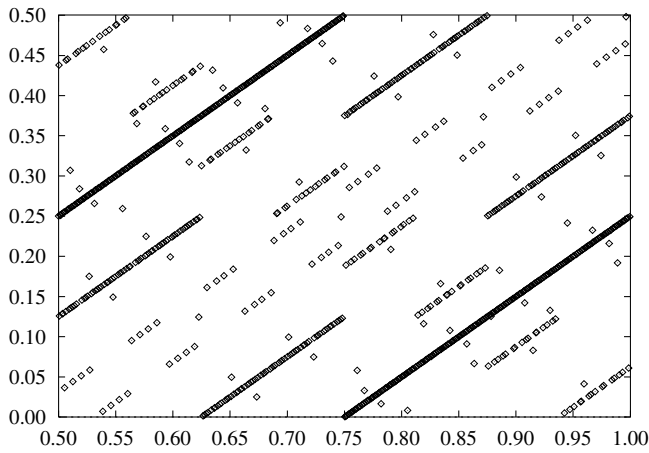


Fig. 5. 2000 one-dimensional Sobol' points.

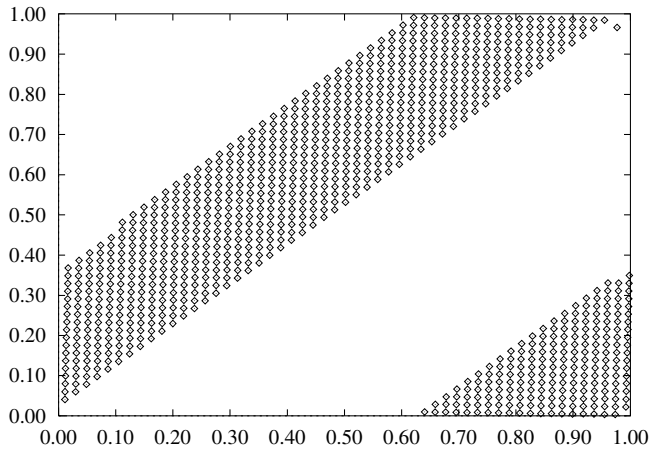


Fig. 6. Coordinates 49 and 50 of 1000 50-dimensional Faure points.

equation 12) for each method.²³ Results for 50 000 and 200 000 sample points are given in Figures 7 and 8, respectively. (The antithetic method uses 25 000 and 100 000 independent pairs of points, respectively.)

Results for the Halton sequence were not competitive and are suppressed. RMS error for standard Monte Carlo is nearly independent of the problem dimension. The antithetic method gives minimal variance reduction. The relative advantage, in terms of RMS error, of the low-discrepancy sequences decreases with the problem dimension. For this test problem, the crossover point is beyond dimension 100.

²³ The details of the distribution are given in Broadie and Detemple (1996).

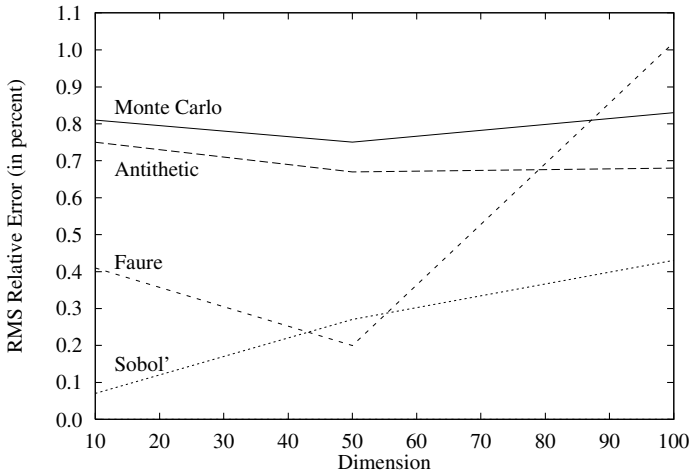


Fig. 7. Results with 50 000 points.

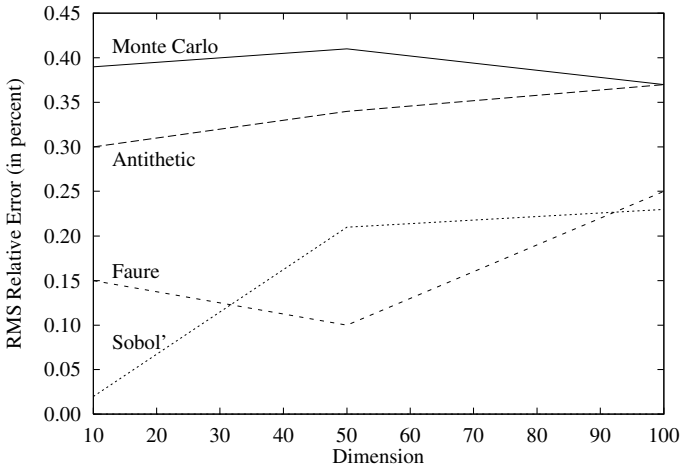


Fig. 8. Results with 200 000 points.

4 Estimating price sensitivities

Most of the discussion in this paper centers on the use of Monte Carlo for pricing securities. In practice, the evaluation of price *sensitivities* is often as important as the evaluation of the prices themselves. Indeed, whereas prices for some securities

can be observed in the market, their sensitivities to parameter changes typically cannot and must therefore be computed. Since price sensitivities are important measures of risk, the growing emphasis on risk management systems suggests a greater need for their efficient computation.

The derivatives of a derivative security's price with respect to various model parameters are collectively referred to as *Greeks*, because several of these are commonly referred to with the names of Greek letters.²⁴ Perhaps the most important of these – and the one to which we give primary attention – is *delta*: the derivative of the price of a contingent claim with respect to the current price of an underlying asset. The delta of a stock option, for example, is the derivative of the option price with respect to the current stock price. An option involving multiple underlying assets has multiple deltas, one for each underlying asset.

In the rest of this section, we discuss various approaches to estimating price sensitivities, especially delta. We begin by examining finite-difference approximations and show that these can be improved through the use of *common random numbers*. We then discuss *direct methods* that estimate derivatives without requiring resimulation at perturbed parameter values.

4.1 Finite-difference approximations

Consider the problem of computing the delta of the Black–Scholes price of a European call; i.e., computing

$$\Delta = \frac{dC}{dS_0},$$

where C is the option price and S_0 is the current stock price. There is, of course, an explicit expression for delta, so simulation is not required, but the example is useful for purposes of illustration. A crude estimate of delta is obtained by generating a terminal stock price

$$S_T = S_0 e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z} \quad (13)$$

(see (2) for notation) from the current stock price S_0 and a second, independent terminal stock price

$$S_T(\epsilon) = (S_0 + \epsilon) e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z'} \quad (14)$$

from the perturbed initial price $S_0 + \epsilon$, with Z and Z' independent. For each terminal price, a discounted payoff can be computed like this:

$$\hat{C}(S_0) = e^{-rT} \max\{0, S_T - K\}, \quad \hat{C}(S_0 + \epsilon) = e^{-rT} \max\{0, S_T(\epsilon) - K\}$$

²⁴ See, e.g., Chapter 13 of Hull (2000) for background.

(see (3) for notation). A crude estimate of delta is then provided by the finite-difference approximation

$$\tilde{\Delta} = \epsilon^{-1}[\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)]. \quad (15)$$

By generating n independent replications of S_T and $S_T(\epsilon)$ we can calculate the sample mean of n independent copies of $\tilde{\Delta}$. As $n \rightarrow \infty$, this sample mean converges to the true finite-difference ratio

$$\epsilon^{-1}[C(S_0 + \epsilon) - C(S_0)], \quad (16)$$

where $C(\cdot)$ is the option price as a function of the current stock price.

This discussion suggests that to get an accurate estimate of Δ we should make ϵ small. However, because we generated S_T and $S_T(\epsilon)$ independently of each other, we have

$$\text{Var}[\tilde{\Delta}] = \epsilon^{-2}(\text{Var}[\hat{C}(S_0 + \epsilon)] + \text{Var}[\hat{C}(S_0)]) = O(\epsilon^{-2}),$$

so the variance of $\tilde{\Delta}$ becomes very large if we make ϵ small. To get an estimator that converges to Δ we must let ϵ decrease slowly as n increases, resulting in slow overall convergence. A general result of Glynn (1989) shows that the best possible convergence rate using this approach is typically $n^{-1/4}$. Replacing the forward difference estimator in (15) with the central difference $(2\epsilon)^{-1}[\hat{C}(S_0 + \epsilon) - \hat{C}(S_0 - \epsilon)]$ typically improves the optimal convergence rate to $n^{-1/3}$. These rates should be compared with $n^{-1/2}$, the rate ordinarily expected from Monte Carlo.

Better estimators can generally be improved using the method of *common random numbers*, which, in this context, simply uses the same Z in (13) and (14). Denote by $\hat{\Delta}$ the finite-difference approximation thus obtained. For fixed ϵ , the sample mean of independent replications of $\hat{\Delta}$ also converges to (16). The variance parameter is given by

$$\text{Var}[\hat{\Delta}] = \epsilon^{-2}(\text{Var}[\hat{C}(S_0)] + \text{Var}[\hat{C}(S_0 + \epsilon)] - 2\text{Cov}[\hat{C}(S_0), \hat{C}(S_0 + \epsilon)]),$$

because $\hat{C}(S_0)$ and $\hat{C}(S_0 + \epsilon)$ are no longer independent. Indeed, if they are positively correlated, then $\hat{\Delta}$ has smaller variance than $\tilde{\Delta}$. That they are in fact positively correlated follows from the monotonicity of the function mapping Z to $\hat{C}$ by the argument used in our discussion of antithetics in Section 3. Thus, the use of common random numbers reduces the variance of the estimate of delta.

The impact of this variance reduction is most dramatic when ϵ is small. A simple calculation shows that, using common random numbers,

$$\begin{aligned} |\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)| &\leq |S_T(\epsilon) - S_T| \\ &\leq \epsilon e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z}. \end{aligned}$$

Because this upper bound has finite second moment, we may conclude that

$$E[|\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)|^2] = O(\epsilon^2), \quad (17)$$

and therefore that

$$\text{Var}[\epsilon^{-1}\{\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)\}] = O(1);$$

i.e., the variance of $\hat{\Delta}$ remains *bounded* as $\epsilon \rightarrow 0$, whereas we saw previously that the variance of $\tilde{\Delta}$ increases at rate ϵ^{-2} . Thus, the more precisely we try to estimate Δ (by making ϵ small) the greater the benefit of common random numbers. Moreover, this indicates that to get an estimator that converges to Δ we may let ϵ decrease faster as n increases than was possible with $\tilde{\Delta}$, resulting in faster overall convergence. An application of Proposition 2 of L'Ecuyer and Perron (1994) shows that a convergence rate of $n^{-1/2}$ can be achieved in this case, and that is the best that can ordinarily be expected from Monte Carlo. For more on convergence rates using common random numbers see Glasserman and Yao (1992), Glynn (1989), and L'Ecuyer and Perron (1994).

The dramatic success of common random numbers in this example relies on the fast rate of mean-square convergence of $\hat{C}(S_0 + \epsilon)$ to $\hat{C}(S_0)$ evidenced by (17). This rate does not apply in all cases. It fails to hold, for example, in the case of a *digital option*²⁵ paying a fixed amount B if $S_T > K$ and 0 otherwise. The price of this option is $C = e^{-rT} B P(S_T > K)$; the obvious simulation estimator is

$$\hat{C}(S_0) = \mathbf{1}_{\{S_T > K\}} e^{-rT} B.$$

Because $\hat{C}(S_0)$ and $\hat{C}(S_0 + \epsilon)$ differ only when $S_T \leq K < S_T(\epsilon)$, we have

$$\begin{aligned} E[|\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)|^2] &= B^2 e^{-2rT} P(S_T \leq K < S_T(\epsilon)) \\ &= B^2 e^{-2rT} P(S_T \leq K < (1 + \epsilon/S_0)S_T) = O(\epsilon), \end{aligned}$$

compared with $O(\epsilon^2)$ for a standard call. As a result, delta estimation is more difficult for the digital option, and a similar argument applies to barrier options generally. Even in these cases, the use of common random numbers can result in substantial improvement compared with differences based on independent runs.

Table 6 compares the performance of four types of delta estimates: forward and central finite-differences with and without common random numbers. The methods are compared at four values of the perturbation parameter ϵ , and applied to the two options discussed above. The values in the table are estimated root mean square errors. The numerical results substantiate the analysis above. Much lower errors are obtained for the standard call than for the digital option, allowing for smaller ϵ ; central differences beat forward differences; common random numbers helps, but

²⁵ Also called a “binary” or “cash-or-nothing” option; see Hull (2000, p. 464).

Table 6. *RMS errors for various delta estimation methods.*

		Independent		Common	
ϵ		Forward	Central	Forward	Central
Standard	10	0.10	0.01	0.100	0.009
Call	1	0.18	0.09	0.012	0.006
Option	0.1	1.78	0.87	0.006	0.006
	0.01	7.47	8.98	0.006	0.006
Digital	20	0.51	0.37	0.51	0.37
Option	10	0.22	0.11	0.21	0.10
	5	0.16	0.07	0.11	0.05
	1	0.67	0.34	0.14	0.10

Root mean square error of delta estimates for two options using four methods with various values of ϵ . Both options have $S_0 = 100$, $K = 100$, $\sigma = 0.40$, $r = 0.10$, and $T = 0.2$. The digital option has $B = 100$. Each entry is computed from 1000 delta estimates, each estimate based on 10 000 replications. The value of delta is 0.580 for the first option and 2.185 for the second.

it helps the standard call more than the digital option. In several cases, the minimal error is obtained using a fairly large ϵ . This reflects the fact that the bias resulting from a large ϵ is sometimes overwhelmed by the large variance resulting from a small ϵ .

Although we have discussed common random numbers in only a limited context, it can easily be applied to a wide range of problems. If all stochastic inputs to a simulation are samples from the normal distribution, then common random numbers can be implemented by using the same samples at two different parameter settings. More generally, if the stochastic inputs are all drawn from a sequence of uniform random variates, then common random numbers can be implemented by using these variates at two different parameter settings.

4.2 *Direct estimates*

Even with the improvements in performance obtained from common random numbers, derivative estimates based on finite differences still suffer from two shortcomings. They are biased (since they compute difference ratios rather than derivatives) and they require multiple resimulations: estimating sensitivities to d parameter changes requires repeatedly running one simulation with all parameters at their base values and d additional simulations with each of the parameters perturbed.

The computation of 10–50 Greeks²⁶ for a single security is not unheard of, and this represents a significant computational burden when multiple resimulations are required.

Over the last decade, a variety of *direct* methods have been developed for estimating derivatives by simulation. Direct methods compute a derivative estimate from a single simulation, and thus do not require resimulation at a perturbed parameter value. Under appropriate conditions, they result in unbiased estimates of the derivatives themselves, rather than of a finite-difference ratio. Our discussion focuses on the use of *pathwise* derivatives as direct estimates, based on a technique generally called *infinitesimal perturbation analysis* (see, e.g., Glasserman 1991).

The pathwise estimate of the true delta dC/dS_0 is the derivative of the sample price $\hat{C}$ with respect to S_0 . More precisely, it is

$$\frac{d\hat{C}}{dS_0} = \lim_{\epsilon \rightarrow 0} \epsilon^{-1} [\hat{C}(S_0 + \epsilon) - \hat{C}(S_0)],$$

provided the limit exists with probability 1. If $\hat{C}(S_0)$ and $\hat{C}(S_0 + \epsilon)$ are computed from the same Z , then provided $S_T \neq K$, we have

$$\begin{aligned} \frac{d\hat{C}}{dS_0} &= \frac{d\hat{C}}{dS_T} \frac{dS_T}{dS_0} \\ &= e^{-rT} \mathbf{1}_{\{S_T > K\}} \frac{S_T}{S_0}. \end{aligned} \tag{18}$$

We have used (13) to get

$$\frac{dS_T}{dS_0} = e^{(r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z} = \frac{S_T}{S_0},$$

and

$$\frac{d\hat{C}}{dS_T} = e^{-rT} \frac{d}{dS_T} \max\{0, S_T - K\} = \begin{cases} e^{-rT}, & S_T > K; \\ 0, & S_T < K. \end{cases}$$

At $S_T = K$, C fails to be differentiable; however, since this occurs with probability zero, the random variable $d\hat{C}/dS_0$ is almost surely well defined.

The pathwise derivative $d\hat{C}/dS_0$ can be thought of as a limiting case of the common random numbers finite-difference estimator in which we evaluate the limit analytically rather than numerically. It is a direct estimator of the option delta because it can be computed directly from a simulation starting at S_0 without the need for a separate simulation at a perturbed value S_0 . This is evident from the expression in (18). The question remains whether this estimator is unbiased; that

²⁶ Sensitivities to various changes in the yield curve often account for several of these.

is, whether

$$E \left[\frac{d\hat{C}}{dS_0} \right] = \frac{dC}{dS_0} \equiv \frac{d}{dS_0} E[\hat{C}].$$

The unbiasedness of the pathwise estimate thus reduces to the interchangeability of derivative and expectation. The interchange is easily justified in this case; see Broadie and Glasserman (1996) for this example and conditions for more general cases. Applying the same reasoning used above, we obtain the following pathwise estimators of three other Greeks for the Black–Scholes price:

$$\begin{aligned} \text{Rho } (dC/dr): & \quad K T e^{-rT} \mathbf{1}_{\{S_T \geq K\}} \\ \text{Vega } (dC/d\sigma): & \quad e^{-rT} \mathbf{1}_{\{S_T \geq K\}} \frac{S_T}{\sigma} \left(\ln(S_T/S_0) - (r - \tfrac{1}{2}\sigma^2)T \right) \\ \text{Theta } (-dC/dT): & \quad r e^{-rT} \max(S_T - K, 0) - \mathbf{1}_{\{S_T \geq K\}} e^{-rT} \frac{S_T}{2T} \left(\ln(S_T/S_0) \right. \\ & \quad \left. + (r - \tfrac{1}{2}\sigma^2)T \right). \end{aligned}$$

Each of these estimators is unbiased.

Of course, Monte Carlo estimators are not required for these derivatives because closed-form expressions are available for each. The Black–Scholes setting is useful for illustration, but the utility of the technique rests on its applicability to more general models. In Broadie and Glasserman (1996), pathwise estimates are derived and studied (both theoretically and numerically) for Asian options and a model with stochastic volatility. For example, the Asian-option delta estimate is simply

$$e^{-rT} \frac{\bar{S}}{S_0} \mathbf{1}_{\{\bar{S} > K\}},$$

where $\bar{S}$ is the average asset price used to determine the option payoff. Evaluating this expression takes *negligible* time compared with resimulating to estimate the option price from a perturbed initial stock price. The pathwise estimate is thus both more accurate and faster to compute than the finite-difference approximation. These advantages extend to a wide class of problems.

As already noted, the unbiasedness of pathwise derivative estimates depends on an interchange of derivative and expectation. In practice, this generally means that the security payoff should be a pathwise continuous function of the parameter in question. The standard call option payoff $e^{-rT} \max\{0, S_T - K\}$ is continuous in each of its parameters. An example where continuity fails is a digital option with payoff $e^{-rT} \mathbf{1}_{\{S_T > K\}} B$, with B the amount received if the stock finishes in the

money.²⁷ Because of the discontinuity at $S_T = K$, the pathwise method (in its simplest form) cannot be applied to this type of option.

The problem of discontinuities often arises in the estimation of *gamma*, the second derivative of an option price with respect to the current price of an underlying asset. Consider, again, the standard European call option. We have an expression for $d\hat{C}/dS_0$ in (18) involving the indicator $\mathbf{1}_{\{S_T > K\}}$. This shows that $d\hat{C}/dS_0$ is discontinuous in S_T , preventing us from differentiating pathwise a second time to get a direct estimator of gamma.

To address the problem of discontinuities, Broadie and Glasserman (1996) construct *smoothed* estimators. These estimators are unbiased, but not as simple to derive and implement as ordinary pathwise estimators. Broadie and Glasserman also investigate another technique for direct derivative estimation called the *likelihood ratio method*. This method differentiates the probability density of an asset price, rather than the outcome of the asset price itself.²⁸ The domains of this method and the pathwise method overlap, but neither contains the other. When both apply, the pathwise method generally has lower variance.

Overviews of these methods can be found in Glasserman (1991), Glynn (1987), and Rubinstein and Shapiro (1993). For discussions specific to financial applications see Broadie and Glasserman (1996) and Fu and Hu (1995).

5 Pricing American options by simulation

European contingent claims have cash flows that cannot be influenced by decisions of the owner. Examples include European options, barrier options, and many types of swaps. By contrast, the cash flows of *American* contingent claims depend both on the price path of the underlying asset or assets and the decisions of the owner. Many types of American contingent claims trade on exchanges and in the over-the-counter market. Examples include American options, American swaptions, shout options, and American Asian options. They also arise in other contexts, for example as “real options” in the theory of economic investment described in Dixit and Pindyck (1994).

To be concrete, suppose that we wish to estimate the quantity $\max_{\tau} E[e^{-r\tau} h(S_{\tau})]$, where r is the constant riskless interest rate, $h(S_{\tau})$ is the payoff at time τ in state S_{τ} , and the max is taken over all stopping times $\tau \leq T$. This formulation of the American pricing problem will suffice to illustrate the major points. First, note that the state can be vector-valued and hence

²⁷ We used this example at the end of Section 3. The settings are related: problems for which common random numbers is particularly effective are generally problems to which the pathwise method can be applied even more effectively.

²⁸ Though not presented in a Monte Carlo context, the expressions in Carr (1993) are potentially relevant to this approach.

applies to pricing American options on multiple assets. Second, since simulation algorithms are discrete in nature, the continuous-time exercise decision must be approximated by restricting the exercise opportunities to lie in a finite set of times $0 = t_0 < t_1 < \dots < t_d = T$. This is not always a serious restriction. For example, for a call option on a stock which pays dividends at discrete points in time, it can be shown that early exercise is only optimal just prior to the ex-dividend dates. In other cases, Richardson or other extrapolation techniques can be used to better approximate the price with exercise in continuous time from a finite set of exercise opportunities.²⁹ However, we now restrict attention to estimating the quantity

$$P \equiv \max_{\tau} E[e^{-r\tau} h(S_{\tau})], \quad (19)$$

where the max is taken over all stopping times τ in the set t_i , for $i = 0, \dots, d$. The need to estimate an optimal stopping time is the crucial distinction between American and European pricing problems.

If the state space is of low dimension, say three or less, a discretization scheme together with a dynamic programming algorithm can often be used to numerically approximate the value in (19). Even in these cases, simulation can be used to estimate the expectation in the recursive step. Simulation-based methods become essential when the dimension of the state space is large.

An obvious simulation-based algorithm for estimating the quantity P in equation (19) is to generate a random path of states S_{t_i} , for $i = 1, \dots, d$, and form the path estimate

$$\hat{P} = \max_{i=0, \dots, d} e^{-rt_i} h(S_{t_i}).$$

However, this estimator corresponds to using *perfect foresight*, and so it is biased high. That is, $E[\hat{P}] \geq P$, which follows immediately from the inequality $\max_{i=0, \dots, d} e^{-rt_i} h(S_{t_i}) \geq e^{-r\tau} h(S_{\tau})$. A natural goal would be to develop an alternative unbiased estimator. A negative result in this regard is provided in Broadie and Glasserman (1997): among a large class of estimators, there is no unbiased estimator of P . In particular, the estimators proposed in Tilley (1993), Grant, Vora, and Weeks (1997), and Barraquand and Martineau (1995) are all biased. Unfortunately, they provide no way to estimate the extent of the bias or to correct for the bias in a general setting. Broadie and Glasserman (1997) circumvent this problem by developing two estimators, one biased high and one biased low (but both asymptotically unbiased), which can be used together to form a valid confidence interval for the quantity P . In the remainder of this section, we give brief descriptions of the four methods mentioned and describe some strengths and weaknesses of each.

²⁹ Geske and Johnson (1984) gave the first financial application of Richardson extrapolation. An extensive treatment of extrapolation techniques is given in Marchuk and Shaidurov (1983).

5.1 Tilley's bundling algorithm

Tilley (1993) sparked considerable interest by demonstrating the potential practicality of applying simulation to pricing American contingent claims. Tilley describes a “bundling procedure” for pricing an American option on a single underlying asset. To estimate P he suggests simulating n paths of asset prices denoted $S_{t_i}(j)$ for $i = 1, \dots, d$ and $j = 1, \dots, n$ in the usual way. Next, partition the asset price space and call the paths which fall into a given partition at a fixed time a “bundle.” A dynamic programming algorithm is applied to bundles to estimate C . In particular, the estimated option price $P_{t_i}(j)$ at time t_i for path j is the maximum of the immediate exercise value, $h(S_{t_i}(j))$, and the present value of continuing. The latter value is defined to be the average of $e^{-r(t_{i+1}-t_i)} P_{t_{i+1}}(k)$ over all paths k which fall in the bundle containing path j at time t_i . Details of the partitioning are given in Tilley (1993).

In order to implement the algorithm, all paths must be stored so they can be sorted into bundles at each time step. Since simulation typically requires a large number of paths for good estimates, the storage and sorting requirements can be significant. More importantly, the algorithm does not easily generalize to multiple state variables. In higher dimensions, it is not clear how to define the bundles. Even then it is likely that most partitions will contain very few paths and lead to a large bias, or the partitions will be so large that the continuation values are poorly estimated.

Because Tilley's algorithm uses the same paths to estimate the optimal decisions and the value, the estimator tends to be biased high (although the bundling induces an approximation which is difficult to analyze). Tilley introduces a “sharp boundary” variant which reduces the bias, but this variant does not easily generalize to higher dimensions. Carriere (1996) contains further analysis of Tilley's algorithm and suggests a procedure based on spline functions to reduce the bias. It remains to be seen whether the spline procedure is practical for higher dimensional problems. Nevertheless, for single state variable problems, Tilley demonstrated the potential practicality of applying simulation to American-style pricing problems.

5.2 Barraquand and Martineau's stratified state aggregation (SSA) algorithm

Barraquand and Martineau (1995) propose a partitioning algorithm, but unlike Tilley's bundling algorithm, they partition the payoff space instead of the state space. Hence, only a one dimensional space is partitioned at each time step, independent of the number of state variables.³⁰ Their algorithm works as follows.

³⁰ In fact, they distinguish between partitioning the state space, which they term “stratified state aggregation,” and partitioning the payoff space, which they term “stratified state aggregation along the payoff.” The latter method is the only one that they test or specify in detail. Hence we focus our discussion on this variant of their method.

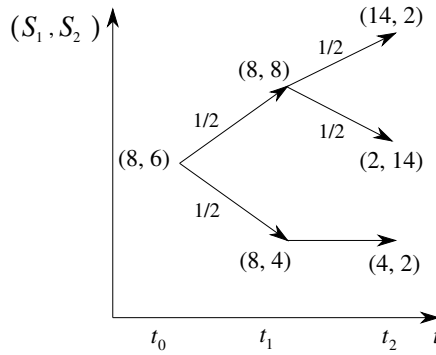


Fig. 9. State evolution.

First, partition the payoff space into K disjoint cells. Then simulate n paths of asset prices denoted $S_{t_i}(j)$ for $i = 1, \dots, d$ and $j = 1, \dots, n$ in the usual way. For each payoff cell k at time t_i , record the number of paths, $a_{t_i}(k)$, which fall into the cell. For each pair of cells k and l at consecutive times t_i and t_{i+1} , record the number of paths, $b_{t_i}(k, l)$, which fall into both cells. Also, for each cell k at time t_i , record the sum of the payoff values, $c_{t_i}(k) = \sum h(S_{t_i}(j))$, where the sum is over all paths j which fall into cell k at time t_i . The transition probability from (t_i, k) to (t_{i+1}, l) is approximated by $p_{t_i}(k, l) = b_{t_i}(k, l)/a_{t_i}(k)$. The estimated option price $P_{t_i}(k)$ at time t_i in cell k is the maximum of the immediate exercise value and the present value of continuing. The immediate exercise value is approximated by $c_{t_i}(k)/a_{t_i}(k)$. The present value of continuing is approximated by $e^{-r(t_{i+1}-t_i)} \sum_{l=1}^K p_{t_i}(k, l) P_{t_{i+1}}(l)$. This procedure can be applied backwards in time to determine the simulation estimate of the price P .

Details of a payoff space partitioning scheme are given in Barraquand and Martineau (1995). Once a single path is generated and the summary information a , b , and c is recorded, the path can be discarded. Hence the storage requirements with this method are modest: on the order of $K^2 d$. One drawback of this method is a possible lack of convergence, as the following example illustrates.

Figure 9 shows the evolution of two asset prices (S_1, S_2) . The option payoff is $h(S_1, S_2) = \max(S_1, S_2)$ and for convenience the riskless rate is taken to be zero. Using the risk-neutral probabilities in Figure 9, the true value of the option at time t_0 is 11, which at time t_1 involves exercise in state $(8, 4)$ but continuing in state $(8, 8)$. When the states are partitioned by their payoffs, these two states are indistinguishable. As seen in the payoff evolution in Figure 10, the best strategy at time t_1 in payoff state 8 is to continue. The apparent value of the option in Figure 10 is 9 ($= (1/2)14 + (1/2)4$). In this example, partitioning the payoff

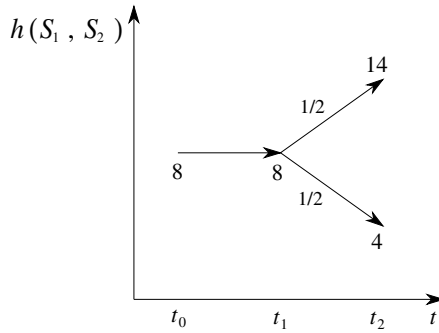


Fig. 10. Payoff evolution.

space leads to a significant underestimate of the option value. Hence, a simulation algorithm based on partitioning the payoff space cannot converge to the correct value. Although this example may seem contrived, Broadie and Detemple (1997) show that the payoff value is not a sufficient statistic for determining the optimal exercise decision for options on the maximum of several assets. Indeed, the payoff process $h(S_t)$ is hardly ever Markovian.

There is currently no way to bound the error in the Barraquand and Martineau method. Without an error estimate, it is difficult to determine the appropriate number of paths to simulate or the appropriate number of partitions to use. Their method can be slightly modified to generate an option price estimate which is biased low as follows. Their procedure gives an exercise strategy based on the immediate exercise payoff. Using this strategy, a new (independent) set of paths can be simulated, and an option value can be estimated under the exercise strategy previously estimated. The resulting option price estimate will be biased low because the exercise policy is not, in general, the optimal policy. With this modification, the average direction of the error is known. Raymar and Zwecher (1997) extend the Barraquand and Martineau approach by basing the exercise decision on a partition of two state-variables, rather than one.

5.3 Broadie and Glasserman's random tree algorithm

Broadie and Glasserman (1997) propose an algorithm based on simulated trees. In order to handle the bias problem, they develop two estimators, one biased high and one biased low, but both convergent and asymptotically unbiased as the computational effort increases. A valid confidence interval for the true value P is obtained by taking the upper confidence limit from the "high" estimator and the lower confidence limit from the "low" estimator. Briefly, their algorithm works as follows.

First, simulate a tree of asset prices (or, more generally, state variables) using b branches at each node. Two paths emanating from a node evolve as independent copies of the state process. The high estimator, Θ , is defined to be the value obtained by the usual dynamic programming algorithm applied to the simulated tree. Then repeat the process for n trees, and compute a point estimate and confidence interval for $E[\Theta]$. A low estimator is obtained by modifying the dynamic programming algorithm at each node. Instead of using all b branches to determine the decision and value, b_1 branches are used to determine the exercise decision, and the remaining $b_2 = b - b_1$ branches are used to determine the continuation value. Their actual low estimator, θ , includes another modification of this procedure which reduces the variance of the estimate. As before, estimates from n trees are combined to give a point estimate and confidence interval for $E[\theta]$. Details of the procedure can be found in Broadie and Glasserman (1997).

For the Θ estimator, all of the branches at a given node are used to determine the optimal decision and the corresponding node value, and this leads to an upward bias, i.e., $E[\Theta] \geq P$. For the θ estimator, the decision and the continuation value are determined from independent information sets. This eliminates the upward bias, but a downward bias occurs, i.e., $E[\theta] \leq P$. The intuition for this result follows. If the correct decision is inferred at a node, the node value estimate would be unbiased. If the incorrect decision is inferred at a node, the node value estimate would be biased low because of the suboptimality of the decision. The expected node value is a weighted average of an unbiased estimate (based on the correct decision) and an estimate which is biased low (based on the incorrect decision). The net effect is an estimate which is biased low. Both estimators are consistent and asymptotically unbiased as b increases.

The computational effort with this algorithm is order nb^d and its main drawback is that d cannot be too large for practical computations. Broadie and Glasserman (1997) give numerical results for options with $d = 4$. As mentioned earlier, to approximate option values with continuous exercise opportunities, some type of extrapolation procedure is required. Special care is necessary to implement extrapolation procedures within a simulation context because of the randomness in the estimates.

5.4 Other developments³¹

Grant, Vora, and Weeks (1997) describe a method specially designed to price American arithmetic Asian options on a single underlying asset. In this application the optimal exercise decision depends on the current asset price and the current

³¹ More recent developments in pricing American options by simulation include Broadie and Glasserman (1997), Broadie, Glasserman and Ha (2000) and Longstaff and Schwartz (2001).

value of the average. Using repeated simulation runs, they attempt to identify the form of an optimal exercise policy based on these two pieces of information. Once an exercise policy is specified, simulation is used to estimate the option value under this fixed policy. Since the fixed policy is a suboptimal approximation to the optimal stopping rule, their procedure leads to a simulation estimator which is biased low.

GVW perform extensive sensitivity analysis which indicates that their option value estimate is relatively insensitive to deviations in the chosen exercise policy. So it may be that their method gives good option price estimates relative to some accuracy level, but it is not clear how to quantify their error. It is not clear how to improve their estimates to an arbitrary accuracy level as the simulation effort increases. Their procedure is specific to the case of American Asian options and does not at this point constitute a general approach to pricing American contingent claims.

Bossaerts (1989) proposes two estimators of optimal early exercise, a *moment estimator* and a *smooth optimization estimator*, and studies their convergence properties. His method appears to require a parametric representation of the exercise boundary and may therefore face difficulties in higher dimension. The optimization approach described in Fu and Hu (1995) also requires a parametric representation.

Rust (1997)³² studies the general problem of solving *discrete decision problems*, which include optimal stopping problems as a special case. He develops a Monte Carlo method and shows that it succeeds in breaking the “curse of dimensionality” in these problem. Rust’s focus is on computational complexity, but his approach appears to provide a promising direction for finance applications.

5.5 Summary

The valuation of securities with American-type features requires the determination of optimal decisions. High dimension versions of these problems arise from multiple state variables and/or path dependencies. Although simulation is a powerful tool for solving some higher dimensional problems, conventional wisdom was that simulation could not be applied to American-style pricing problems. The algorithms described here represent the first attempts to solve these problems that were long thought to be computationally intractable.

6 Further topics

We conclude this paper with a brief mention of two important areas of current work in the application of Monte Carlo methods to finance, not discussed in this article.

³² We thank A. Dixit for pointing us to this reference.

A central numerical issue in simulating interest rates, asset prices with stochastic volatilities, and other complex diffusions is the accurate approximation of stochastic differential equations by discrete-time processes. Kloeden and Platen (1992) discuss a variety of methods for constructing discrete-time approximations with different orders of convergence. Andersen (1995) applies some of these to interest-rate models. In general, decreasing the time increment in a discrete approximation can be expected to give more accurate results, but at the expense of greater computational effort. Duffie and Glynn (1995) analyze this trade-off and characterize asymptotically optimal time steps as the overall computational effort grows.

In this article we have focused almost exclusively on the use of Monte Carlo for pricing. A related, growing area of application is risk management – in particular, the use of Monte Carlo to assess value at risk, credit risk, and related measures. For some examples of recent applications in these areas see Iben and Brotherton-Ratcliffe (1994), Lawrence (1994), Beckström and Campbell (1995) and Glasserman, Heidelberger and Shahabuddin (2000).

Appendix: Moment controls beat moment matching asymptotically

As mentioned in Section 2.4, any time a moment is available for use with moment matching, it can alternatively be used as a control variate. In this appendix, we argue that moment matching is asymptotically equivalent to a control variate technique with suboptimal coefficients, and is therefore dominated by the optimal use of moments as controls. This asymptotic link applies in large samples. A related link between linear and nonlinear control variates is made in Glynn and Whitt (1989), but the current setting does not fit their framework.

Let $Z_1, Z_2, \dots$ be i.i.d. (not necessarily normal) with mean μ and variance σ^2 . Let s denote the sample standard deviation of $Z_1, \dots, Z_n$ and $\bar{Z}$ their sample mean. Suppose we want to estimate $E[f(Z)]$ for some function f . The standard estimator is $n^{-1} \sum_{i=1}^n f(Z_i)$ and the moment matching estimator is $n^{-1} \sum_{i=1}^n f(\tilde{Z}_i)$ with $\tilde{Z}_i$ defined in (9). For each i , the scaled difference

$$\sqrt{n}(\tilde{Z}_i - Z_i) = \sqrt{n} \left(\frac{\sigma - s}{s} \right) Z_i - \sqrt{n}[(\sigma \bar{Z}/s) - \mu]$$

converges in distribution, by the central limit theorem for $\bar{Z}$ and s . Thus, $(\tilde{Z}_i - Z_i) = O_p(n^{-1/2})$ (see, e.g., Appendix A of Pollard 1984 for O_p, o_p notation).

Suppose now that, with probability one, f is differentiable at Z_i . Then

$$f(\tilde{Z}_i) = f(Z_i) + f'(Z_i)[\tilde{Z}_i - Z_i] + o_p(n^{-1/2}),$$

suggesting that up to terms $o_p(n^{-1/2})$ the moment matching estimator and standard

estimator are related via

$$\begin{aligned}
 \frac{1}{n} \sum_{i=1}^n f(\tilde{Z}_i) &\approx \frac{1}{n} \sum_{i=1}^n f(Z_i) + \frac{1}{n} \sum_{i=1}^n f'(Z_i)[\tilde{Z}_i - Z_i] \\
 &= \frac{1}{n} \sum_{i=1}^n f(Z_i) + \frac{1}{n} \sum_{i=1}^n f'(Z_i) \left[\left(\frac{\sigma}{s} - 1 \right) Z_i - \frac{\sigma}{s} \bar{Z} + \mu \right] \\
 &= \frac{1}{n} \sum_{i=1}^n f(Z_i) + \left(\frac{1}{n} \sum_{i=1}^n f'(Z_i) Z_i \right) \left(\frac{\sigma}{s} - 1 \right) \\
 &\quad + \left(\frac{1}{n} \sum_{i=1}^n f'(Z_i) \right) \left(\mu - \frac{\sigma}{s} \bar{Z} \right) \\
 &\equiv \frac{1}{n} \sum_{i=1}^n f(Z_i) + \hat{\beta}_1 \left(\frac{\sigma}{s} - 1 \right) + \hat{\beta}_2 \left(\mu - \frac{\sigma}{s} \bar{Z} \right)
 \end{aligned}$$

where $\hat{\beta}_i \rightarrow \beta_i$, $i = 1, 2$, as $n \rightarrow \infty$, with

$$\beta_1 = E[f'(Z)Z], \quad \text{and} \quad \beta_2 = E[f'(Z)].$$

Thus, moment matching is asymptotically equivalent to using

$$\left(\frac{\sigma}{s} - 1 \right) \quad \text{and} \quad \left(\mu - \frac{\sigma}{s} \bar{Z} \right) \quad (20)$$

as controls (both quantities converge to zero almost surely) with estimates of coefficients β_1, β_2 . In general, these do not coincide with the optimal coefficients β_1^*, β_2^* , so moment matching is asymptotically dominated by the control variate method. In addition, the controls in (20) introduce some bias (as does moment matching itself) because though they converge to zero they do not have mean zero for finite n . In contrast, the more natural moment control variates $(s^2 - \sigma^2)$ and $(\bar{Z} - \mu)$ have mean zero for all n and thus introduce no bias.

References

- Acworth, P., M. Broadie, and P. Glasserman, 1997, A Comparison of Some Monte Carlo and Quasi Monte Carlo Methods for Option Pricing, in *Monte Carlo and Quasi Monte Methods for Scientific Computing*, G. Larcher, P. Hellekalek, H. Niederreiter, and P. Zinterhof (eds.), Springer-Verlag, Berlin.
- Andersen, L., 1995, Efficient Techniques for Simulation of Interest Rate Models Involving Non-Linear Stochastic Differential Equations, Working paper (General Re Financial Products, New York, NY).
- Andersen, L., and R. Brotherton-Ratcliffe, 1996, Exact Exotics, *Risk* 9, October, 85–89.
- Barlow, R.E. and F. Proschan, 1975, *Statistical Theory of Reliability and Life Testing* (Holt, Reinhart and Winston, New York).
- Barraquand, J., 1995, Numerical Valuation of High Dimensional Multivariate European Securities, *Management Science* 41, 1882–1891.

- Barraquand, J. and D. Martineau, 1995, Numerical Valuation of High Dimensional Multivariate American Securities, *Journal of Financial and Quantitative Analysis* 30, 383–405.
- Beaglehole, D., P. Dybvig, and G. Zhou, 1997, Going to Extremes: Correcting Simulation Bias in Exotic Option Valuation, *Financial Analysts Journal* (Jan/Feb) 62–68.
- Beckström, R. and A. Campbell, 1995, *An Introduction to VAR* (CATS Software, Palo Alto, California).
- Berman, L., 1996, Comparison of Path Generation Methods for Monte Carlo Valuation of Single Underlying Derivative Securities, Research Report RC-20570, IBM Research, Yorktown Heights, New York.
- Birge, J.R., 1994, Quasi-Monte Carlo Approaches to Option Pricing, Technical Report 94–119 (Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109).
- Bossaerts, P., 1989, Simulation Estimators of Optimal Early Exercise, Working paper (Carnegie-Mellon University, Pittsburgh, PA, 15213).
- Boyle, P., 1977, Options: A Monte Carlo Approach, *Journal of Financial Economics* 4, 323–338.
- Boyle, P. and D. Emanuel, 1985, The Pricing of Options on the Generalized Mean, Working paper (University of Waterloo).
- Bratley, P. and B. Fox, 1988, ALGORITHM 659: Implementing Sobol's Quasirandom Sequence Generator, *ACM Transactions on Mathematical Software* 14, 88–100.
- Bratley, P., B.L. Fox, and H. Niederreiter, 1992, Implementation and Tests of Low-Discrepancy Sequences, *ACM Transactions on Modelling and Computer Simulation* 2, 195–213.
- Bratley, P., B.L. Fox, and L. Schrage, 1987, *A Guide to Simulation*, 2nd Ed. (Springer-Verlag, New York).
- Broadie, M. and J. Detemple, 1997, The Valuation of American Options on Multiple Assets, *Mathematical Finance* 7, 241–286.
- Broadie, M. and J. Detemple, 1996, American Option Valuation: New Bounds, Approximations, and a Comparison of Existing Methods, *Review of Financial Studies* 9, 1211–1250.
- Broadie, M. and P. Glasserman, 1996, Estimating Security Price Derivatives by Simulation, *Management Science* 42, 269–285.
- Broadie, M. and P. Glasserman, 1997, Pricing American-Style Securities Using Simulation, *Journal of Economic Dynamics and Control* 21, 1323–1352.
- Broadie, M. and P. Glasserman, 1997, A Stochastic Mesh Method for Pricing High-Dimensional American Options, Working paper, Columbia Business School, New York.
- Broadie, M., P. Glasserman, and Z. Ha, 2000, Pricing American Options by Simulation Using a Stochastic Mesh with Optimized Weights, in *Probabilistic Constrained Optimization*, S. Uryasev, ed., 26–44 (Kluwer, Norwell, Mass.)
- Caffisch, R.E., W. Morokoff, and A. Owen, 1998, Valuation of Mortgage Backed Securities Using Brownian Bridges to Reduce Effective Dimension, in *Monte Carlo: Methodologies and Applications for Pricing and Risk Management*, 301–314 (Risk Publications, London).
- Carr, P., 1993, Deriving Derivatives of Derivative Securities, Working paper (Johnson Graduate School of Business, Cornell University).
- Carriere, J.F., 1996, Valuation of the Early-Exercise Price for Derivative Securities using Simulations and Splines, *Insurance: Mathematics and Economics* 19, 19–30.
- Carverhill, A. and K. Pang, 1995, Efficient and Flexible Bond Option Valuation in the

- Heath, Jarrow and Morton Framework, *Journal of Fixed Income* 5, September, 70–77.
- Cheyette, O., 1992, Term Structure Dynamics and Mortgage Valuation, *Journal of Fixed Income* 2, March, 28–41.
- Clewlow, L. and A. Carverhill, 1994, On the Simulation of Contingent Claims, *Journal of Derivatives* 2, Winter, 66–74.
- Devroye, L., 1986, *Non-Uniform Random Variate Generation* (Springer-Verlag, New York).
- Dixit, A. and R. Pindyck, 1994, *Investment Under Uncertainty* (Princeton University Press).
- Duan, J.-C., 1995, The GARCH Option Pricing Model, *Mathematical Finance* 5, 13–32.
- Duan, J.-C. and J.-G. Simonato, 1998, Empirical Martingale Simulation for Asset Prices, *Management Science* 44, 1218–1233.
- Duffie, D., 1996, *Dynamic Asset Pricing Theory*, 2nd ed. (Princeton University Press, Princeton, New Jersey).
- Duffie, D. and P. Glynn, 1995, Efficient Monte Carlo Simulation of Security Prices, *Annals of Applied Probability* 5, 897–905.
- Faure H., 1982, Discrépance de Suites Associées à un Système de Numération (en Dimension s), *Acta Arithmetica* 41, 337–351.
- Fox, B.L., 1986, ALGORITHM 647: Implementation and Relative Efficiency of Quasi-Random Sequence Generators, *ACM Transactions on Mathematical Software* 12, 362–376.
- Fu, M. and J.Q. Hu, 1995, Sensitivity Analysis for Monte Carlo Simulation of Option Pricing, *Probability in the Engineering and Information Sciences* 9, 417–446.
- Fu, M., D. Madan, and T. Wong, 1998, Pricing Continuous Time Asian Options: A Comparison of Analytical and Monte Carlo Methods, *Journal of Computational Finance* 2, 49–74.
- Geske, R. and H.E. Johnson, 1984, The American Put Options Valued Analytically, *Journal of Finance* 39, 1511–1524.
- Glasserman, P., 1991, *Gradient Estimation via Perturbation Analysis* (Kluwer Academic Publishers, Norwell, Mass).
- Glasserman, P., 1993, Filtered Monte Carlo, *Mathematics of Operations Research* 18, 610–634.
- Glasserman, P., P. Heideberger, and P. Shahabuddin, 2000, Variance Reduction Techniques for Estimating Value-at-Risk, *Management Science* 46, 1349–1365.
- Glasserman, P. and D.D. Yao, 1992, Some Guidelines and Guarantees for Common Random Numbers, *Management Science* 38, 884–908.
- Glynn, P.W., 1987, Likelihood Ratio Gradient Estimation: An Overview, in: *Proceedings of the Winter Simulation Conference* (The Society for Computer Simulation, San Diego, California) 366–374.
- Glynn, P.W., 1989, Optimization of Stochastic Systems via Simulation, in: *Proceedings of the Winter Simulation Conference* (The Society for Computer Simulation, San Diego, California) 90–105.
- Glynn, P.W. and D.L. Iglehart, 1988, Simulation Methods for Queues: An Overview, *Queueing Systems* 3, 221–255.
- Glynn, P.W. and W. Whitt, 1989, Indirect Estimation via $L = \lambda W$, *Operations Research* 37, 82–103.
- Glynn, P.W. and W. Whitt, 1992, The Asymptotic Efficiency of Simulation Estimators, *Operations Research* 40, 505–520.
- Grant, D., G. Vora, and D. Weeks, 1997, Path-Dependent Options: Extending the Monte

- Carlo Simulation Approach, *Management Science* 43, 1589–1602.
- Halton, J.H., 1960, On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals, *Numerische Mathematik* 2, 84–90.
- Hammersley, J.M. and D.C. Handscomb, 1964, *Monte Carlo Methods* (Chapman and Hall, London).
- Haselgrove, C.B., 1961, A Method for Numerical Integration, *Mathematics of Computation* 15, 323–337.
- Hlawka, E., 1971, Discrepancy and Riemann Integration, in: L. Mirsky, ed., *Studies in Pure Mathematics* (Academic Press, New York).
- Hull, J., 2000, *Options, Futures, and Other Derivative Securities*, 4th ed. (Prentice-Hall, Englewood Cliffs, New Jersey).
- Hull, J. and A. White, 1987, The Pricing of Options on Assets with Stochastic Volatilities, *Journal of Finance* 42, 281–300.
- Iben, B. and R. Brotherton-Ratcliffe, 1994, Credit Loss Distributions and Required Capital for Derivatives Portfolios, *Journal of Fixed Income* 4, June, 6–14.
- Johnson, H., 1987, Options on the Maximum or the Minimum of Several Assets, *Journal of Financial and Quantitative Analysis* 22, 227–283.
- Johnson, H. and D. Shanno, 1987, Option Pricing When the Variance is Changing, *Journal of Financial and Quantitative Analysis* 22, 143–151.
- Joy C., P.P. Boyle, and K.S. Tan, 1996, Quasi-Monte Carlo Methods in Numerical Finance, *Management Science* 42, 926–938.
- Kemna, A.G.Z. and A.C.F. Vorst, 1990, A Pricing Method for Options Based on Average Asset Values, *Journal of Banking and Finance* 14, 113–129.
- Kloeden, P. and E. Platen, 1992, *Numerical Solution of Stochastic Differential Equations* (Springer-Verlag, New York).
- L'Ecuyer, P. and G. Perron, 1994, On the Convergence Rates of IPA and FDC Derivative Estimators, *Operations Research* 42, 643–656.
- Lavenberg, S.S. and P.D. Welch, 1981, A Perspective on the Use of Control Variables to Increase the Efficiency of Monte Carlo Simulations, *Management Science* 27, 322–335.
- Lawrence, D., 1994, Aggregating Credit Exposures: The Simulation Approach, in: *Derivative Credit Risk* (Risk Publications, London).
- Longstaff, F.A. and E.S. Schwartz, 2001, Valuing American Options by Simulation: A Simple Least Squares Approach, *Review of Financial Studies* 14, 113–148.
- Marchuk, G. and V. Shaidurov, 1983, *Difference Methods and Their Extrapolations* (Springer Verlag, New York).
- McKay, M.D., W.J. Conover, and R.J. Beckman, 1979, A Comparison of Three Methods for Selecting Input Variables in the Analysis of Output from a Computer Code, *Technometrics* 21, 239–245.
- Morokoff, W.J. and R.E. Caflisch, 1995, Quasi-Monte Carlo Integration, *Journal of Computational Physics*, 122, 218–230.
- Moskowitz B. and R.E. Caflisch, 1996, Smoothness and Dimension Reduction in Quasi-Monte Carlo Methods, *Mathematical and Computer Modeling* 23, 37–54.
- Niederreiter, H., 1988, Low Discrepancy and Low Dispersion Sequences, *Journal of Number Theory* 30, 51–70.
- Niederreiter, H., 1976, On the Distribution of Pseudo-Random Numbers Generated by the Linear Congruential Method. III, *Mathematics of Computation* 30, 571–597.
- Niederreiter, H., 1992, *Random Number Generation and Quasi-Monte Carlo Methods* (CBMS-NSF 63, SIAM, Philadelphia, Pa).
- Niederreiter, H. and C. Xing, 1996, Low-Discrepancy Sequences and Global Function

- Fields with Many Rational Places, Finite Fields and their Applications 2, 241–273.
- Nielsen, S., 1994, Importance Sampling in Lattice Pricing Models, Working paper (Management Science and Information Systems, University of Texas at Austin).
- Ninomiya, S., and S. Tezuka, 1996, Toward Real-Time Pricing of Complex Financial Derivatives, *Applied Mathematical Finance* 3, 1–20.
- Owen, A., 1995a, Monte Carlo Variance of Scrambled Equidistribution Quadrature, in: H. Niederreiter and P.J.S. Shiue, eds., *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (Springer-Verlag, Berlin).
- Owen, A., 1995b, Randomly Permuted (t, m, s) -Nets and (t, s) -Sequences, in *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, H. Niederreiter and P. Shiue (eds.), 299–317 (Springer-Verlag, New York).
- Paskov, S. and J. Traub, 1995, Faster Valuation of Financial Derivatives, *Journal of Portfolio Management* 22, Fall, 113–120.
- Pollard, D., 1984, *Convergence of Stochastic Processes*, Springer-Verlag, New York.
- Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, 1992, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press).
- Raymar, S., and M. Zwecher, 1997, A Monte Carlo Valuation of American Call Options On the Maximum of Several Stocks, *Journal of Derivatives* 5 (Fall), 7–24.
- Reider, R., 1993, An Efficient Monte Carlo Technique for Pricing Options, Working paper (Wharton School, University of Pennsylvania).
- Rubinstein, R. and A. Shapiro, 1993, *Discrete Event Systems* (Wiley, New York).
- Rust, J., 1997, Using Randomization to Break the Curse of Dimensionality, *Econometrica* 65, 487–516.
- Schwartz, E.S. and W.N. Torous, 1989, Prepayment and the Valuation of Mortgage-Backed Securities, *Journal of Finance* 44, 375–392.
- Scott, L.O., 1987, Option Pricing when the Variance Changes Randomly: Theory, Estimation, and an Application, *Journal of Financial and Quantitative Analysis* 22, 419–438.
- Shaw, J., 1995, Beyond VAR and Stress Testing, in *Monte Carlo: Methodologies and Applications for Pricing and Risk Management*, 231–244 (Risk Publications, London).
- Sobol', I.M., 1967, On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals, *USSR Computational Mathematics and Mathematical Physics* 7, 86–112.
- Spanier, J. and E.H. Maize, 1994, Quasi-Random Methods for Estimating Integrals Using Relatively Small Samples, *SIAM Review* 36, 18–44.
- Stein, M., 1987, Large Sample Properties of Simulations Using Latin Hypercube Sampling, *Technometrics* 29, 143–151.
- Stulz, R.M., 1982, Options on the Minimum or the Maximum of Two Risky Assets, *Journal of Financial Economics* 10, 161–185.
- Tezuka, S., 1994, A Generalization of Faure Sequences and its Efficient Implementation, Research Report RTO105 (IBM Research, Tokyo Research Laboratory, Kanagawa, Japan).
- Tezuka, S., 1995, *Uniform Random Numbers: Theory and Practice* (Kluwer Academic Publishers, Boston).
- Tilley, J.A., 1993, Valuing American Options in a Path Simulation Model, *Transactions of the Society of Actuaries* 45, 83–104.
- Turnbull, S.M. and L.M. Wakeman, 1991, A Quick Algorithm for Pricing European Average Options, *Journal of Financial and Quantitative Analysis* 26, 377–389.
- Van Rensberg J. and G.M. Torrie, 1993, Estimation of Multidimensional Integrals: Is

- Monte Carlo the Best Method?, *Journal of Physics A: Mathematical and General* 26, 943–953.
- Wiggins, J.B., 1987, Option Values under Stochastic Volatility: Theory and Empirical Evidence, *Journal of Financial Economics* 19, 351–372.
- Willard, G.A., 1997, Calculating Prices and Sensitivities for Path-Dependent Derivative Securities in Multifactor Models, *Journal of Derivatives* 5 (Fall), 45–61.
- Worzel, K.J., C. Vassiadou-Zeniou, and S.A. Zenios, 1994, Integrated Simulation and Optimization Models for Tracking Indices of Fixed-Income Securities, *Operations Research* 42, 223–233.
- Zaremba, S.K., 1968, The Mathematical Basis of Monte Carlo and Quasi-Monte Carlo Methods, *SIAM Review* 10, 310–314.

Part two

Interest Rate Modeling

A Geometric View of Interest Rate Theory

Tomas Björk

1 Introduction

1.1 Setup

We consider a bond market model (see Björk (1997), Musiela and Rutkowski (1997)) living on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, Q)$ where $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$. The basis is assumed to carry a standard m -dimensional Wiener process W , and we also assume that the filtration $\mathbf{F}$ is the internal one generated by W .

By $p(t, x)$ we denote the price, at t , of a zero coupon bond maturing at $t + x$, and the forward rates $r(t, x)$ are defined by

$$r(t, x) = -\frac{\partial \log p(t, x)}{\partial x}.$$

Note that we use the Musiela parameterization, where x denotes the time **to** maturity. The short rate R is defined as $R(t) = r(t, 0)$, and the money account B is given by $B(t) = \exp \left\{ \int_0^t R(s) ds \right\}$. The model is assumed to be free of arbitrage in the sense that the measure Q above is a martingale measure for the model. In other words, for every fixed time of maturity $T \geq 0$, the process $Z(t, T) = p(t, T - t)/B(t)$ is a Q -martingale.

Let us now consider a given forward rate model of the form

$$\begin{cases} dr(t, x) &= \beta(t, x)dt + \sigma(t, x)dW, \\ r(0, x) &= r^o(0, x), \end{cases} \quad (1)$$

where, for each x , β and σ are given optional processes. The initial curve $\{r^o(0, x); x \geq 0\}$ is taken as given. It is interpreted as the *observed forward rate curve*.

The standard Heath–Jarrow–Morton drift condition (Heath, Jarrow and Morton (1992)) can easily be transferred to the Musiela parameterization. The result (see Brace and Musiela (1994), Musiela (1993)) is as follows.

Proposition 1.1 (The forward rate equation) *Under the martingale measure Q the r -dynamics are given by*

$$dr(t, x) = \left\{ \frac{\partial}{\partial x} r(t, x) + \sigma(t, x) \int_0^x \sigma(t, u)^* du \right\} dt + \sigma(t, x) dW(t), \quad (2)$$

$$r(0, x) = r^o(0, x). \quad (3)$$

where $*$ denotes transpose.

1.2 Main problems

Suppose now that we are given a concrete model $\mathcal{M}$ within the above framework, i.e. suppose that we are given a concrete specification of the volatility process σ . We now formulate a couple of natural problems:

1. Take, in addition to $\mathcal{M}$, also as given a parameterized family $\mathcal{G}$ of forward rate curves. Under which conditions is the family $\mathcal{G}$ **consistent** with the dynamics of $\mathcal{M}$? Here consistency is interpreted in the sense that, given an initial forward rate curve in $\mathcal{G}$, the interest rate model $\mathcal{M}$ will only produce forward rate curves belonging to the given family $\mathcal{G}$.
2. When can the given, inherently infinite dimensional, interest rate model $\mathcal{M}$ be written as a **finite dimensional state space model**? More precisely, we seek conditions under which the forward rate process $r(t, x)$, induced by the model $\mathcal{M}$, can be realized by a system of the form

$$dZ_t = a(Z_t)dt + b(Z_t)dW_t, \quad (4)$$

$$r(t, x) = G(Z_t, x), \quad (5)$$

where Z (interpreted as the state vector process) is a finite dimensional diffusion, $a(z)$, $b(z)$ and $G(z, x)$ are deterministic functions and W is the same Wiener process as in (2).

As will be seen below, these two problems are intimately connected, and the main purpose of this chapter is to give an overview of some recent work in this area. The text is mainly based on Björk and Christensen (1999), Björk and Gombani (1999) and Björk and Svensson (1999), but the presentation given below is more focused on geometric intuition than the original articles, where full proofs, technical details and further results can be found. In the analysis below we use ideas from systems and control theory (see Isidori (1989)) as well as from nonlinear filtering theory (see Brockett (1981)). References to the literature will sometimes be given in the text, but will mainly be summarized in the Notes at the end of each section.

The organization of the text is as follows. In Section 2 we study the existence of a finite dimensional factor realization in the comparatively simple case when

the forward rate volatilities are deterministic. In Section 3 we study the general consistency problem, and in Section 4 we use the consistency results from Section 3 in order to give a fairly complete picture of the nonlinear realization problem.

2 Linear realization theory

In the general case, the forward rate equation (2) is a highly nonlinear infinite dimensional SDE but, as can be expected, the special case of linear dynamics is much easier to handle. In this section we therefore concentrate on linear forward rate models, and look for finite dimensional linear realizations.

2.1 Deterministic forward rate volatilities

For the rest of the section we only consider the case when the volatility $\sigma(t, x) = [\sigma_1(t, x), \dots, \sigma_m(t, x)]$ is a deterministic time-independent function $\sigma(x)$ of x only.

Assumption 2.1 *The volatility σ is a deterministic C^∞ -mapping $\sigma : R_+ \rightarrow R^m$.*

Denoting the function $x \mapsto r(t, x)$ by $r(t)$ we have, from (2),

$$dr(t) = \{\mathbf{F}r(t) + D\} dt + \sigma dW(t), \quad (6)$$

$$r(0) = r^o(0). \quad (7)$$

Here the linear operator $\mathbf{F}$ is defined by

$$\mathbf{F} = \frac{\partial}{\partial x}, \quad (8)$$

whereas the function D is given by

$$D(x) = \sigma(x) \int_0^x \sigma(s)^* ds. \quad (9)$$

The point to note here is that, because of our choice of a deterministic volatility $\sigma(x)$, the forward rate equation (6) is a **linear** (or rather affine) SDE. Because of this linearity (albeit in infinite dimensions) we therefore expect to be able to provide an explicit solution of (6). We now recall that a scalar equation of the form

$$dy(t) = [ay(t) + b] dt + cdW(t)$$

has the solution

$$y(t) = e^{at} y(0) + \int_0^t e^{a(t-s)} b ds + \int_0^t e^{a(t-s)} c dW(s),$$

and we are led to conjecture that the solution to (6) is given by the formal expression

$$r(t) = e^{\mathbf{F}t} r^o + \int_0^t e^{\mathbf{F}(t-s)} D ds + \int_0^t e^{\mathbf{F}(t-s)} \sigma dW(s).$$

The formal exponential $e^{\mathbf{F}t}$ acts on real valued functions, and we have to figure out how it operates. From the standard series expansion of the exponential function one is led to write

$$[e^{\mathbf{F}t} f](x) = \sum_{n=0}^{\infty} \frac{t^n}{n!} [F^n f](x). \quad (10)$$

In our case $F^n = \frac{\partial^n}{\partial x^n}$, so (assuming f to be analytic) we have

$$[e^{\mathbf{F}t} f](x) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \frac{\partial^n f}{\partial x^n}(x). \quad (11)$$

This is, however, just a Taylor series expansion of f around the point x , so for analytic f we have $[e^{\mathbf{F}t} f](x) = f(x + t)$. We have in fact the following precise result (which can be proved rigorously).

Proposition 2.2 *The operator $\mathbf{F}$ is the infinitesimal generator of the semigroup of left translations, i.e. for any $f \in C[0, \infty)$ we have*

$$[e^{\mathbf{F}t} f](x) = f(t + x).$$

The solution of the forward rate equation (6) is given as

$$r(t, x) = e^{\mathbf{F}t} r^o(0, x) + \int_0^t e^{\mathbf{F}(t-s)} D(x) ds + \int_0^t e^{\mathbf{F}(t-s)} \sigma(x) dW(s) \quad (12)$$

or equivalently by

$$r(t, x) = r^o(0, x + t) + \int_0^t D(x + t - s) ds + \int_0^t \sigma(x + t - s) dW(s). \quad (13)$$

From (12) it is clear by inspection that we may write the forward rate equation (6) as

$$dr_0(t, x) = \mathbf{F}r_0(t, x)dt + \sigma(x)dW(t), \quad r_0(0, x) = 0 \quad (14)$$

$$r(t, x) = r_0(t, x) + \delta(t, x), \quad (15)$$

where δ is given by

$$\delta(t, x) = r^o(0, x + t) + \int_0^t D(x + t - s) ds. \quad (16)$$

Since $\delta(t, x)$ is not affected by the input W , we see that the problem of finding a realization for the term structure system (6) is equivalent to that of finding a realization for (14). We are thus led to the following definition.

Definition 2.3 A matrix triple $[A, B, C(x)]$ is called an n -dimensional realization of the systems (6) and (14) if r_0 has the representation

$$dZ(t) = AZ(t)dt + BdW(t), \quad Z(0) = 0, \quad (17)$$

$$r_0(t, x) = C(x)Z(t). \quad (18)$$

Our main problems are now as follows.

- Take as *a priori* given a volatility structure $\sigma(x)$.
- When does there exist a finite dimensional realization?
- If there exists a finite dimensional realization, what is the minimal dimension?
- How do we construct a minimal realization from knowledge of σ ?
- Is there an economic interpretation of the state process Z in the realization?

2.2 Existence of finite linear realizations

We will now go on to study the existence of a finite dimensional realization of the stochastic system (14), and in order to get some ideas, suppose that there actually exists a finite dimensional realization of (14) of the form (17)–(18). Solving (14), we have

$$r_0(t, x) = \int_0^t e^{F(t-s)} \sigma(x) dW(s) = \int_0^t \sigma(x + t - s) dW(s),$$

while, from the realization (17)–(18), we also have

$$r_0(t, x) = C(x)Z(t) = C(x) \int_0^t e^{A(t-s)} B dW(s).$$

Thus we have, with probability one, for each x and each t ,

$$\int_0^t \sigma_x(t - s) dW(s) = \int_0^t C(x) e^{A(t-s)} B dW(s), \quad (19)$$

where we use subindex x to denote left translation, i.e. $f_x(t) = f(x + t)$. This leads us immediately to conjecture that the equation

$$\sigma_x(t) = C(x) e^{At} B$$

must hold for all x and t , and we have our first main result.

Proposition 2.4

1. The forward rate process has a finite dimensional linear realization if and only if the volatility function σ can be written in the form

$$\sigma(x) = C_0 e^{Ax} B. \quad (20)$$

2. If σ has the form (20) then a concrete realization of r_0 is given by

$$dZ(t) = AZ(t)dt + BdW(t), \quad Z(0) = 0, \quad (21)$$

$$r_0(t, x) = C(x)Z(t), \quad (22)$$

with A, B as in (20), and with $C(x) = C_0 e^{Ax}$. The forward rates $r(t, x)$ are then given by (15)–(16).

Proof It is clear from the discussion above that if there exists a finite realization, then we must have the factorization $\sigma_x(t) = C(x)e^{At}B$. Setting $x = 0$, and denoting $C(0)$ by C_0 , in this case gives us the relation (20). If, on the other hand, σ factors as in (20), then we simply define Z as in (21). A direct calculation as above then shows that we have $r_0(t, x) = C_0 e^{Ax} z(t)$. $\square$

Remark 2.5 Let us call a function of the form $ce^{Ax}b$, where c is a row vector, A is a square matrix and b is a column vector, a **quasi-exponential** (or QE) function. The general form of a quasi-exponential function f is given by

$$f(x) = \sum_i e^{\lambda_i x} + \sum_j e^{\alpha_j x} [p_j(x) \cos(\omega_j x) + q_j(x) \sin(\omega_j x)], \quad (23)$$

where $\lambda_i, \alpha_j, \omega_j$ are real numbers, whereas p_j and q_j are real polynomials.

QE functions will turn up again, so we list some simple properties.

Lemma 2.6 *The following hold for the quasi-exponential functions:*

- A function is QE if and only if it is a component of the solution of a vector valued linear ODE with constant coefficients.
- A function is QE if and only if it can be written as $f(x) = ce^{Ax}b$.
- If f is QE, then f' is QE.
- If f is QE, then its primitive function is QE.
- If f and g are QE, then fg is QE.

2.3 Transfer functions

Using ideas from linear systems theory, an alternative view of the realization problem is obtained by studying **transfer functions**, i.e. by going to the frequency domain. To get some intuition, consider again the equation

$$dr_0(t, x) = \mathbf{F}r_0(t, x)dt + \sigma(x)dW(t), \quad r_0(0, x) = 0. \quad (24)$$

Let us now formally “divide by dt ”, which gives us

$$\frac{dr_0}{dt}(t, x) = \mathbf{F}r_0(t, x) + \sigma(x)\frac{dW}{dt}(t),$$

where the formal time derivative $\frac{dW}{dt}(t)$ is interpreted as white noise. We interpret this equation as an input–output system where the random input signal $t \mapsto \frac{dW}{dt}(t)$ is transformed into the infinite dimensional output signal $t \mapsto r_0(t, \cdot)$. We thus view the equation as a version of the following controlled ODE:

$$\begin{aligned} \frac{dr_0}{dt}(t, x) &= \mathbf{F}r_0(t, x) + \sigma(x)u(t), \\ r_0(0) &= 0, \end{aligned} \quad (25)$$

where u is a deterministic input signal. Generally speaking, tricks like this do not work directly, since we are ignoring the difference between standard differential calculus, which is used to analyze (25), and Itô calculus which we use when dealing with SDEs. In this case, however, because of the linear structure, the second order Itô term will not come into play, so we are safe. (See the discussion in Section 3.4 around the Stratonovich integral for how to treat the nonlinear situation.)

It is now natural to study the transfer function for the system (25), which relates the Laplace transform of the input signal to the Laplace transform of the output signal.

Definition 2.7 *The transfer function, $K(s, x)$, for (25) is determined by the relation*

$$\tilde{r}_0(s, x) = K(s, x)\tilde{u}(s),$$

where $\tilde{\cdot}$ denotes the Laplace transform in the t -variable.

From the uniqueness of the Laplace transform we then have the following result.

Lemma 2.8 *The system*

$$dZ(t) = AZ(t)dt + BdW(t), \quad Z(0) = 0, \quad (26)$$

$$r_0(t, x) = C(x)Z(t) \quad (27)$$

is a realization of

$$dr_0(t, x) = \mathbf{F}r_0(t, x)dt + \sigma(x)dW(t), \quad r_0(0, x) = 0 \quad (28)$$

if and only if the deterministic control system

$$\frac{dr_0}{dt}(t, x) = \mathbf{F}r_0(t, x) + \sigma(x)u(t) \quad (29)$$

has the same transfer function as the system

$$\frac{dZ}{dt}(t) = AZ(t) + Bu(t), \quad (30)$$

$$r_0(t, x) = C(x)Z(t). \quad (31)$$

Furthermore we have

Lemma 2.9 *The transfer function $K(s, x)$ of (29) is given by*

$$K(s, x) = \mathcal{L}[\sigma_x](s),$$

where $\mathcal{L}$ denotes the Laplace transform, and σ_x denotes left translation.

Proof From (29) we have

$$r_0(t, x) = \int_0^t \sigma(x + t - s)u(s)ds = [\sigma_x \star u](t),$$

and thus

$$\tilde{r}_0(s, x) = \mathcal{L}[\sigma_x](s)\tilde{u}(s).$$

□

For concrete computation of a realization, the following result is useful.

Lemma 2.10

- *The transfer function of the system (30)–(31) is given by*

$$K(s, x) = C(x)[sI - A]^{-1}B.$$

- *The r_0 system has a finite realization if and only if there exists a factorization of the form*

$$\mathcal{L}[\sigma_x](s) = C(x)[sI - A]^{-1}B.$$

- *Denote the transfer function of r_0 by $K(s, x)$, and assume that there exists a finite dimensional realization. If we have found A , B and C such that*

$$K(s, 0) = C[sI - A]^{-1}B,$$

then a realization of r_0 is given by $[A, B, Ce^{Ax}]$.

Proof The first assertion is immediately obtained by taking the Laplace transform of (30)–(31). The second follows from Lemma 2.8, and the third from Proposition 2.4. $\square$

If we want to find a concrete realization for a given system, we thus have two possibilities. We can either look for a factorization of the volatility function as $\sigma(x) = Ce^{Ax}B$, or we can try to factor the transfer function as $K(s, 0) = C[sI - A]^{-1}B$. From a logical point of view the two approaches are equivalent, but from a practical point of view it is much easier to factor the transfer function than to factor the volatility. There are in fact a number of standard algorithms in the systems theoretic literature which construct a realization, given knowledge of the transfer functions. See Brockett (1970).

2.4 Minimal realizations

The purpose of this section is to determine the minimal dimension of a finite dimensional realization.

Definition 2.11 *The **dimension** of a realization $[A, B, C(x)]$ is defined as the dimension of the corresponding state space. A realization $[A, B, C(x)]$ is said to be **minimal** if there is no other realization with smaller dimension. The **McMillan degree**, $\mathcal{D}$, of the forward rate system is defined as the dimension of a minimal realization.*

In order to get a feeling for how to determine the McMillan degree, we note that r_0 has a finite dimensional realization if and only if r_0 evolves on a finite dimensional subspace in the infinite dimensional function space $\mathcal{H}$. Furthermore, it seems obvious that the McMillan degree equals the dimension of this subspace.

In order to determine the subspace above, let us again view the r_0 system as a special case of the following controlled equation, where we have suppressed x .

$$\begin{cases} \frac{dr_0}{dt} = \mathbf{F}r_0(t) + \sigma u(t), \\ r_0(0) = 0. \end{cases} \quad (32)$$

The solution of this equation is given by

$$r_0(t) = \int_0^t e^{\mathbf{F}(t-s)} \sigma u(s) ds = \int_0^t \sum_{n=0}^{\infty} \frac{(t-s)^n}{n!} \mathbf{F}^n \sigma u(s) ds.$$

This is a linear combination of vectors of the form $\mathbf{F}^n \sigma_i$, so we see that the smallest subspace $\mathcal{R}$ which contains $r_0(t)$ for all t and for all choices of the input signal u

is given by

$$\mathcal{R} = \text{span} [\sigma, \mathbf{F}\sigma, \mathbf{F}^2\sigma, \dots] = \text{span} [\mathbf{F}^k \sigma_i ; i = 1, \dots, m \ k = 0, 1, \dots]. \quad (33)$$

We thus have the following result.

Proposition 2.12 *Take the volatility function*

$$\sigma = [\sigma_1, \dots, \sigma_m]$$

as given. Then the McMillan degree, $\mathcal{D}$, is given by

$$\mathcal{D} = \dim(\mathcal{R}), \quad (34)$$

with $\mathcal{R}$ defined as in (33). The forward rate system thus admits a finite dimensional realization if and only if the space spanned by the components of σ and all their derivatives is finite dimensional.

2.5 Economic interpretation of the state space

In general, the state space of the minimal realization of a given system has no concrete (e.g. physical) interpretation. In our case, however, the states of the *minimal* realization turn out to have a simple economic interpretation in terms of a minimal set of “benchmark” forward rates.

Assume that $[A, B, C]$ is a minimal realization, of dimension n , of the forward rates as in (21)–(22). Let us choose a set of “benchmark” maturities $x_1, \dots, x_n$. We use the notation $\bar{x} = (x_1, \dots, x_n)$. Assume furthermore that the maturity vector $\bar{x}$ is chosen so that the matrix

$$T(\bar{x}) = \begin{bmatrix} Ce^{Ax_1} \\ \vdots \\ Ce^{Ax_n} \end{bmatrix}$$

is invertible. It can be shown (see Björk and Gombani (1999)) that, outside a set of measure zero, this can always be done as long as the maturities are distinct. We use the notation

$$r_0(t, \bar{x}) = \begin{bmatrix} r_0(t, x_1) \\ \vdots \\ r_0(t, x_n) \end{bmatrix}$$

and corresponding interpretations for column vectors like $r(t, \bar{x})$, $\delta(t, \bar{x})$ etc.

The following result shows how the entire term structure is determined by the benchmark forward rates.

Proposition 2.13 Assume that (21)–(22) is a minimal realization of the forward rates, and assume furthermore that a maturity vector $\bar{x} = (x_1, \dots, x_n)$ is chosen as above. Then the following hold.

- With notation as above, the vector $r(t, \bar{x})$ of benchmark forward rates has the dynamics

$$dr(t, \bar{x}) = [T(\bar{x})AT^{-1}(\bar{x})r(t, \bar{x}) + \Psi(t, \bar{x})]dt + T(\bar{x})BdW(t), \quad (35)$$

$$r(0, \bar{x}) = r^*(0, \bar{x}),$$

where the deterministic function Ψ is given by

$$\Psi(t, \bar{x}) = \frac{\partial r^*}{\partial x}(0, t\bar{e} + \bar{x}) + D(t\bar{e} + \bar{x}) - T(\bar{x})AT^{-1}(\bar{x})\delta(t, \bar{x}).$$

Here $\bar{e} \in R^n$ denotes the vector with unit components, i.e.

$$\bar{e} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- The system of benchmark forward rates determine the entire forward rate process according to the formula

$$r(t, x) = Ce^{Ax}T^{-1}(\bar{x})r(t, \bar{x}) - Ce^{Ax}T^{-1}(\bar{x})\delta(t, \bar{x}) + \delta(t, x). \quad (36)$$

- The correspondence between Z and r is given by

$$r_0(t, \bar{x}) = T(\bar{x})Z(t). \quad (37)$$

Proof See Björk and Gombani (1999). □

The conclusion is thus that the state variables of a *minimal* realization can be interpreted as an affine transformation of a vector of benchmark forward rates.

2.6 Examples

In this section we will give some simple illustrations of the theory. Note the handling of multiple roots of the matrix A , and the fact that the input noise can have dimension smaller than the dimension of A .

Example 2.14 $\sigma(x) = \sigma e^{-ax}$ We consider a model driven by a one-dimensional Wiener process, having the forward rate volatility structure

$$\sigma(x) = \sigma e^{-ax},$$

where σ in the right hand side denotes a constant. (The reader will probably recognize this example as the Hull–White model.) We start by determining the McMillan degree $\mathcal{D}$, and by Proposition 2.12 we have

$$\mathcal{D} = \dim(\mathcal{R}),$$

where the space $\mathcal{R}$ is given by

$$\mathcal{R} = \text{span} \left[\frac{d^k}{dx^k} \sigma e^{-ax} ; k \geq 0 \right].$$

It is obvious that $\mathcal{R}$ is one dimensional, and that it is spanned by the single function e^{-ax} . Thus the McMillan degree is given by $\mathcal{D} = 1$. We now want to apply Proposition 2.4 to find a realization, so we must factor the volatility function. In this case this is easy, since we have the trivial factorization $\sigma(x) = 1 \cdot e^{-ax} \cdot \sigma$. In the notation of Proposition 2.4 we thus have

$$\begin{aligned} C_0 &= 1, \\ A &= -a, \\ B &= \sigma. \end{aligned}$$

A realization of the forward rates is thus given by

$$\begin{aligned} dZ(t) &= -aZ(t)dt + \sigma dW(t), \\ r_0(t, x) &= e^{-ax} Z(t), \\ r(t, x) &= r_0(t, x) + \delta(t, x), \end{aligned}$$

and since the state space in this realization is of dimension one, the realization is minimal. We see that if $a > 0$ then the system is asymptotically stable.

We now go on to the interpretation of the state space, and since $\mathcal{D} = 1$ we can choose a single benchmark maturity. The canonical choice is of course $x_1 = 0$, i.e. we choose the instantaneous short rate $R(t)$ as the state variable. In the notation of Proposition 2.13 we then have

$$\begin{aligned} T(\bar{x}) &= 1, \\ r(t, \bar{x}) &= R(t), \end{aligned}$$

and we get rate dynamics

$$dR(t) = \{\Psi(t, 0) - aR(t)\} dt + \sigma dW(t).$$

Thus we see that we have indeed the Hull–White extension of the Vasiček model (1977). Note however that we do not have to choose the benchmark maturity as

$x_1 = 0$. We can in fact choose *any* fixed maturity, x_1 , and then use the corresponding forward rate as benchmark. This will give us the dynamics

$$dr(t, x_1) = \{\Psi(t, x_1) - ar(t, x_1)\} dt + e^{-ax_1} dW(t),$$

and now the entire forward rate curve will be determined by the x_1 -rate according to formula (36).

Example 2.15 $\sigma(x) = xe^{-ax}$

In this example we still have a single driving Wiener process, but the volatility function is now “hump-shaped”.

By taking derivatives of $\sigma(x)$ we immediately see, from Proposition 2.12, that $\mathcal{R}$ is given by

$$\mathcal{R} = \text{span} [xe^{-ax}, e^{-ax}],$$

so in this case $\mathcal{D} = 2$, and we have a two-dimensional minimal state space. In order to obtain a realization we compute the transfer function $K(s, x)$, which is given by Lemma 2.9 as

$$K(s, x) = \mathcal{L}[(x + \cdot)e^{-a(x+\cdot)}](s).$$

An easy calculation gives us

$$K(s, x) = \frac{e^{-ax}}{(a+s)^2} + \frac{xe^{-ax}}{(a+s)} = \frac{sx e^{-ax} + (1+ax)e^{-ax}}{(a+s)^2},$$

and we now look for a realization of this transfer function (for a fixed x). The obvious thing to do is to use the standard controllable realization (see Brockett (1970)), and we obtain

$$\begin{aligned} C(x) &= [xe^{-ax}, (1+ax)e^{-ax}], \\ A &= \begin{bmatrix} -2a & -a^2 \\ 1 & 0 \end{bmatrix}, \\ B &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \end{aligned}$$

Since $\mathcal{D} = 2$ and this realization is two-dimensional we have a minimal realization, given by

$$\begin{aligned} dZ_1(t) &= -2aZ_1(t)dt - a^2Z_2(t)dt + dW(t), \\ dZ_2(t) &= Z_1(t)dt, \\ r_0(t, x) &= xe^{-ax}Z_1(t) + (1+ax)e^{-ax}Z_2(t), \\ r(t, x) &= r_0(t, x) + \delta(t, x). \end{aligned}$$

We have a double eigenvalue of the system matrix A at $\lambda_1 = -a$, so if $a > 0$ the system is asymptotically stable.

2.7 Notes

This section is mainly based on Björk and Gombani (1999). The first paper to appear in this area was to our knowledge the preprint (Musielà (1993)), where the Musielà parameterization and the space $\mathcal{R}$ are discussed in some detail. See also the closely related and interesting preprints El Karoui and Lacoste (1993), El Karoui, Geman and Lacoste (1997) and Zabczyk (1992). Because of the linear structure, the theory above is closely connected to (and in a sense inverse to) the theory of affine term structures developed in Duffie and Kan (1996). The standard reference on infinite dimensional SDEs is Da Prato and Zabczyk (1992), where one also can find a presentation of the connections between control theory and infinite dimensional linear stochastic equations.

3 Invariant manifolds

In this section we study when a given submanifold of forward rate curves is invariant under the action of a given interest rate model. This problem is of interest from an applied as well as from a theoretical point of view. In particular we will use the results from this section to analyze problems about existence of finite dimensional factor realizations for interest rate models on forward rate form. Invariant manifolds are, however, also of interest in their own right, so we begin by discussing a concrete problem which naturally leads to the invariance concept.

3.1 Parameter recalibration

A standard procedure when dealing with concrete interest rate models on a high frequency (say, daily) basis can be described as follows:

1. At time $t = 0$, use market data to fit (calibrate) the model to the observed bond prices.
2. Use the calibrated model to compute prices of various interest rate derivatives.
3. The following day ($t = 1$), repeat the procedure in 1 above in order to recalibrate the model, etc.

To carry out the calibration in step 1 above, the analyst typically has to produce a forward rate curve $\{r^o(0, x); x \geq 0\}$ from the observed data. However, since only a finite number of bonds actually trade in the market, the data consist of a discrete set of points, and a need to fit a curve to these points arises. This curve-fitting

may be done in a variety of ways. One way is to use splines, but also a number of parameterized families of smooth forward rate curves have become popular in applications – the most well-known probably being the Nelson-Siegel (see Nelson and Siegel (1987)) family. Once the curve $\{r^o(0, x); x \geq 0\}$ has been obtained, the parameters of the interest rate model may be calibrated to this.

Now, from a purely logical point of view, the recalibration procedure in step 3 above is of course slightly nonsensical: if the interest rate model at hand is an exact picture of reality, then there should be no need to recalibrate. The reason that everyone insists on recalibrating is of course that any model in fact is only an approximate picture of the financial market under consideration, and recalibration allows the incorporation of newly arrived information in the approximation. Even so, the calibration procedure itself ought to take into account that it will be repeated. It appears that the optimal way to do so would involve a combination of time series and cross-section data, as opposed to the purely cross-sectional curve-fitting, where the information contained in previous curves is discarded in each recalibration. .

The cross-sectional fitting of a forward curve and the repeated recalibration is thus, in a sense, a pragmatic and somewhat non-theoretical endeavor. Nonetheless, there are some nontrivial theoretical problems to be dealt with in this context, and the problem to be studied in this section concerns the **consistency** between, on the one hand, the dynamics of a given interest rate model, and, on the other hand, the forward curve family employed.

What, then, is meant by consistency in this context? Assume that a given interest rate model $\mathcal{M}$ (e.g. the Hull–White model (1990)) in fact *is* an exact picture of the financial market. Now consider a particular family $\mathcal{G}$ of forward rate curves (e.g. the Nelson–Siegel family) and assume that the interest rate model is calibrated using this family. We then say that the pair $(\mathcal{M}, \mathcal{G})$ is **consistent** (or, that $\mathcal{M}$ and $\mathcal{G}$ are consistent) if all forward curves which may be produced by the interest rate model $\mathcal{M}$ are contained within the family $\mathcal{G}$. Otherwise, the pair $(\mathcal{M}, \mathcal{G})$ is inconsistent.

Thus, if $\mathcal{M}$ and $\mathcal{G}$ are consistent, then the interest rate model actually produces forward curves which belong to the relevant family. In contrast, if $\mathcal{M}$ and $\mathcal{G}$ are inconsistent, then the interest rate model will produce forward curves outside the family used in the calibration step, and this will force the analyst to change the model parameters all the time – not because the model is an approximation to reality, but simply because the family does not go well with the model.

Put into more operational terms this can be rephrased as follows.

- Suppose that you are using a fixed interest rate model $\mathcal{M}$. If you want to do recalibration, then your family $\mathcal{G}$ of forward rate curves should be chosen in

such a way as to be consistent with the model $\mathcal{M}$.

Note however that the argument can also be run backwards, yielding the following conclusion for empirical work.

- Suppose that a particular forward curve family $\mathcal{G}$ has been observed to provide a good fit, on a day-to-day basis, in a particular bond market. Then this gives you modeling information about the choice of an interest rate model in the sense that you should try to use/construct an interest rate model which is consistent with the family $\mathcal{G}$.

We now have a number of natural problems to study.

- I** Given an interest rate model $\mathcal{M}$ and a family of forward curves $\mathcal{G}$, what are necessary and sufficient conditions for consistency?
- II** Take as given a specific family $\mathcal{G}$ of forward curves (e.g. the Nelson–Siegel family). Does there exist any interest rate model $\mathcal{M}$ which is consistent with $\mathcal{G}$?
- III** Take as given a specific interest rate model $\mathcal{M}$ (e.g. the Hull–White model). Does there exist any finitely parameterized family of forward curves $\mathcal{G}$ which is consistent with $\mathcal{M}$?

In this section we will mainly address problem I above. Problem II has been studied, for special cases, in Filipović (1998a,b), whereas Problem III can be shown (see Proposition 4.6) to be equivalent to the problem of finding a finite dimensional factor realization of the model $\mathcal{M}$ and we provide a fairly complete solution in Section 4.

3.2 Invariant manifolds

We now move on to give precise mathematical definition of the consistency property discussed above, and this leads us to the concept of an **invariant manifold**.

Definition 3.1 (Invariant manifold) *Take as given the forward rate process dynamics (2). Consider also a fixed family (manifold) of forward rate curves $\mathcal{G}$. We say that $\mathcal{G}$ is locally **invariant** under the action of r if, for each point $(s, r) \in R_+ \times \mathcal{G}$, the condition $r_s \in \mathcal{G}$ implies that $r_t \in \mathcal{G}$, on a time interval with positive length. If r stays forever on $\mathcal{G}$, we say that $\mathcal{G}$ is globally invariant.*

The purpose of this section is to characterize invariance in terms of local characteristics of $\mathcal{G}$ and $\mathcal{M}$, and in this context local invariance is the best one can hope for. In order to save space, local invariance will therefore be referred to as invariance.

To get some intuitive feeling for the invariance concepts one can consider the following two-dimensional deterministic system

$$\begin{aligned}\frac{dy_1}{dt} &= y_2, \\ \frac{dy_2}{dt} &= -y_1.\end{aligned}$$

For this system it is obvious that the unit circle $\mathcal{C} = \{(y_1, y_2) : y_1^2 + y_2^2 = 1\}$ is globally invariant, i.e. if we start the system on $\mathcal{C}$ it will stay forever on $\mathcal{C}$. The ‘upper half’ of the circle, $\mathcal{C}_u = \{(y_1, y_2) : y_1^2 + y_2^2 = 1, y_2 > 0\}$, is on the other hand only locally invariant, since the system will leave $\mathcal{C}_u$ at the point $(1, 0)$. This geometric situation is in fact the generic one also for our infinite dimensional stochastic case. The forward rate trajectory will never leave a locally invariant manifold at a point in the relative interior of the manifold. Exit from the manifold can only take place at the relative boundary points. We have no general method for determining whether a locally invariant manifold is also globally invariant or not. Problems of this kind have to be solved separately for each particular case.

3.3 The formalized problem

3.3.1 The Space

As our basic space of forward rate curves we will use a weighted Sobolev space, where a generic point will be denoted by r .

Definition 3.2 Consider a fixed real number $\gamma > 0$. The space $\mathcal{H}_\gamma$ is defined as the space of all differentiable (in the distributional sense) functions

$$r : R_+ \rightarrow R$$

satisfying the norm condition $\|r\|_\gamma < \infty$. Here the norm is defined as

$$\|r\|_\gamma^2 = \int_0^\infty r^2(x) e^{-\gamma x} dx + \int_0^\infty \left(\frac{dr}{dx}(x) \right)^2 e^{-\gamma x} dx.$$

Remark 3.3 The variable x is as before interpreted as time to maturity. With the inner product

$$(r, q) = \int_0^\infty r(x)q(x)e^{-ax}dx + \int_0^\infty \left(\frac{dr}{dx}(x) \right) \left(\frac{dq}{dx}(x) \right) e^{-\gamma x} dx,$$

the space $\mathcal{H}_\gamma$ becomes a Hilbert space. Because of the exponential weighting function all constant forward rate curves will belong to the space. In the sequel we will suppress the subindex γ , writing $\mathcal{H}$ instead of $\mathcal{H}_\gamma$.

3.3.2 The Forward Curve Manifold

We consider as given a mapping

$$G : \mathcal{Z} \rightarrow \mathcal{H}, \quad (38)$$

where the parameter space $\mathcal{Z}$ is an open connected subset of R^d , i.e. for each parameter value $z \in \mathcal{Z} \subseteq R^d$ we have a curve $G(z) \in \mathcal{H}$. The value of this curve at the point $x \in R_+$ will be written as $G(z, x)$, so we see that G can also be viewed as a mapping

$$G : \mathcal{Z} \times R_+ \rightarrow R. \quad (39)$$

The mapping G is thus a formalization of the idea of a finitely parameterized family of forward rate curves, and we now define the forward curve manifold as the set of all forward rate curves produced by this family.

Definition 3.4 *The forward curve manifold $\mathcal{G} \subseteq \mathcal{H}$ is defined as*

$$\mathcal{G} = \text{Im } P_A G.$$

3.3.3 The Interest Rate Model

We take as given a volatility function σ of the form

$$\sigma : \mathcal{H} \times R_+ \rightarrow R^m,$$

i.e. $\sigma(r, x)$ is a functional of the infinite dimensional r -variable, and a function of the real variable x . Denoting the forward rate curve at time t by r_t we then have the following forward rate equation.

$$dr_t(x) = \left\{ \frac{\partial}{\partial x} r_t(x) + \sigma(r_t, x) \int_0^x \sigma(r_t, u)^* du \right\} dt + \sigma(r_t, x) dW_t. \quad (40)$$

Remark 3.5 For notational simplicity we have assumed that the r -dynamics are time homogeneous. The case when σ is of the form $\sigma(t, r, x)$ can be treated in exactly the same way. See Björk and Christensen (1999).

We need some regularity assumptions, and the main ones are as follows. See Björk (1997) for technical details.

Assumption 3.6 *We assume the following.*

- *The volatility mapping $r \mapsto \sigma(r)$ is smooth.*
- *The mapping $z \mapsto G(z)$ is a smooth embedding, so in particular the Fréchet derivative $G'_z(z)$ is injective for all $z \in \mathcal{Z}$.*
- *For every initial point $r_0 \in \mathcal{G}$, there exists a unique strong solution in $\mathcal{H}$ of Equation (40).*

3.3.4 The Problem

Our main problem is the following.

- Suppose that we are given
 - A volatility σ , specifying an interest rate model $\mathcal{M}$ as in (40)
 - A mapping G , specifying a forward curve manifold $\mathcal{G}$.
- Is $\mathcal{G}$ then invariant under the action of r ?

3.4 The invariance conditions

In order to study the invariance problem we need to introduce some compact notation.

Definition 3.7 We define $\mathbf{H}\sigma$ by

$$\mathbf{H}\sigma(r, x) = \int_0^x \sigma(r, s) ds.$$

Suppressing the x -variable, the Itô dynamics for the forward rates are thus given by

$$dr_t = \left\{ \frac{\partial}{\partial x} r_t + \sigma(r_t) \mathbf{H}\sigma(r_t)^* \right\} dt + \sigma(r_t) dW_t \quad (41)$$

and we write this more compactly as

$$dr_t = \mu_0(r_t) dt + \sigma(r_t) dW_t, \quad (42)$$

where the drift μ_0 is given by the bracket term in (41). To get some intuition we now formally “divide by dt ” and obtain

$$\frac{dr}{dt} = \mu_0(r_t) + \sigma(r_t) \dot{W}_t, \quad (43)$$

where the formal time derivative $\dot{W}_t$ is interpreted as an “input signal” chosen by chance. As in Section 2.3 we are thus led to study the associated deterministic control system

$$\frac{dr}{dt} = \mu_0(r_t) + \sigma(r_t) u_t. \quad (44)$$

The intuitive idea is now that $\mathcal{G}$ is invariant under (42) if and only if $\mathcal{G}$ is invariant under (44) for all choices of the input signal u . It is furthermore geometrically obvious that this happens if and only if the velocity vector $\mu(r) + \sigma(r)u$ is tangential to $\mathcal{G}$ for all points $r \in \mathcal{G}$ and all choices of $u \in R^m$. Since the tangent space of

$\mathcal{G}$ at a point $G(z)$ is given by $\text{Im}[G'_z(z)]$, where G'_z denotes the Fréchet derivative (Jacobian), we are led to conjecture that $\mathcal{G}$ is invariant if and only if the condition

$$\mu_0(r) + \sigma(r)u \in \text{Im}[G'_z(z)]$$

is satisfied for all $u \in R^m$. This can also be written

$$\begin{aligned}\mu_0(r) &\in \text{Im}[G'_z(z)], \\ \sigma(r) &\in \text{Im}[G'_z(z)],\end{aligned}$$

where the last inclusion is interpreted componentwise for σ .

This “result” is, however, not correct due to the fact that the argument above neglects the difference between ordinary calculus, which is used for (44), and Itô calculus, which governs (42). In order to bridge this gap we have to rewrite the analysis in terms of Stratonovich integrals instead of Itô integrals.

Definition 3.8 For given semimartingales X and Y , the **Stratonovich integral** of X with respect to Y , $\int_0^t X(s) \circ dY(s)$, is defined as

$$\int_0^t X_s \circ dY_s = \int_0^t X_s dY_s + \frac{1}{2} \langle X, Y \rangle_t. \quad (45)$$

The first term on the rhs is the Itô integral. In the present case, with only Wiener processes as driving noise, we can define the “quadratic variation process” $\langle X, Y \rangle$ in (45) by

$$d\langle X, Y \rangle_t = dX_t dY_t, \quad (46)$$

with the usual “multiplication rules” $dW \cdot dt = dt \cdot dt = 0$, $dW \cdot dW = dt$. We now recall the main result and *raison d’être* for the Stratonovich integral.

Proposition 3.9 (Chain rule) Assume that the function $F(t, y)$ is smooth. Then we have

$$dF(t, Y_t) = \frac{\partial F}{\partial t}(t, Y_t)dt + \frac{\partial F}{\partial y} \circ dY_t. \quad (47)$$

Thus, in the Stratonovich calculus, the Itô formula takes the form of the standard chain rule of ordinary calculus.

Returning to (42), the Stratonovich dynamics are given by

$$\begin{aligned}dr_t &= \left\{ \frac{\partial}{\partial x} r_t + \sigma(r_t) \mathbf{H} \sigma(r_t)^* \right\} dt - \frac{1}{2} d\langle \sigma(r_t), W_t \rangle \\ &+ \sigma(r_t) \circ dW_t.\end{aligned} \quad (48)$$

In order to compute the Stratonovich correction term above we use the infinite dimensional Itô formula (see Da Prato and Zabczyk (1992)) to obtain

$$d\sigma(r_t) = \{\cdot\cdot\cdot\} dt + \sigma'_r(r_t)\sigma(r_t)dW_t, \quad (49)$$

where σ'_r denotes the Fréchet derivative of σ w.r.t. the infinite dimensional r -variable. From this we immediately obtain

$$d\langle\sigma(r_t), W_t\rangle = \sigma'_r(r_t)\sigma(r_t)dt. \quad (50)$$

Remark 3.10 If the Wiener process W is multidimensional, then σ is a vector $\sigma = [\sigma_1, \dots, \sigma_m]$, and the rhs of (50) should be interpreted as

$$\sigma'_r(r_t)\sigma(r_t, x) = \sum_{i=1}^m \sigma'_{ir}(r_t)\sigma_i(r_t).$$

Thus (48) becomes

$$\begin{aligned} dr_t &= \left\{ \frac{\partial}{\partial x} r_t + \sigma(r_t) \mathbf{H} \sigma(r_t)^* - \frac{1}{2} \sigma'_r(r_t) \sigma(r_t) \right\} dt \\ &+ \sigma(r_t) \circ dW_t \end{aligned} \quad (51)$$

We now write (51) as

$$dr_t = \mu(r_t)dt + \sigma(r_t) \circ dW_t, \quad (52)$$

where

$$\mu(r, x) = \frac{\partial}{\partial x} r(x) + \sigma(r_t, x) \int_0^x \sigma(r_t, u)^* du - \frac{1}{2} [\sigma'_r(r_t) \sigma(r_t)](x). \quad (53)$$

Given the heuristics above, our main result is not surprising. The formal proof, which is somewhat technical, is left out. See Björk and Christensen (1999).

Theorem 3.11 (Main theorem) *The forward curve manifold $\mathcal{G}$ is locally invariant for the forward rate process $r(t, x)$ in $\mathcal{M}$ if and only if,*

$$G'_x(z) + \sigma(r) \mathbf{H} \sigma(r)^* - \frac{1}{2} \sigma'_r(r) \sigma(r) \in \text{Im}[G'_z(z)], \quad (54)$$

$$\sigma(r) \in \text{Im}[G'_z(z)], \quad (55)$$

hold for all $z \in \mathcal{Z}$ with $r = G(z)$.

Here, G'_z and G'_x denote the Fréchet derivative of G with respect to z and x , respectively. The condition (55) is interpreted componentwise for σ . Condition (54) is called *the consistent drift condition*, and (55) is called *the consistent volatility condition*.

Remark 3.12 It is easily seen that if the family G is invariant under shifts in the x -variable, then we will automatically have the relation

$$G'_x(z) \in \text{Im}[G'_z(z)],$$

so in this case the relation (54) can be replaced by

$$\sigma(r) \mathbf{H} \sigma(r)^* - \frac{1}{2} \sigma'_r(r) \sigma(r) \in \text{Im}[G'_z(z)],$$

with $r = G(z)$ as usual.

3.5 Examples

The results above are extremely easy to apply in concrete situations. As a test case we consider the Nelson–Siegel (see Nelson and Siegel (1987)) family of forward rate curves. We analyze the consistency of this family with the Ho–Lee and Hull–White interest rate models. It should be emphasized that these examples are chosen only in order to illustrate the general methodology. For more examples and details, see Björk and Christensen (1999).

3.5.1 The Nelson–Siegel family

The Nelson–Siegel (henceforth NS) forward curve manifold $\mathcal{G}$ is parameterized by $z \in R^4$, the curve $x \mapsto G(z, x)$ as

$$G(z, x) = z_1 + z_2 e^{-z_4 x} + z_3 x e^{-z_4 x}. \quad (56)$$

For $z_4 \neq 0$, the Fréchet derivatives are easily obtained as

$$G'_z(z, x) = [1, e^{-z_4 x}, x e^{-z_4 x}, -(z_2 + z_3 x) x e^{-z_4 x}], \quad (57)$$

$$G'_x(z, x) = (z_3 - z_2 z_4 - z_3 z_4 x) e^{-z_4 x}. \quad (58)$$

In order for the image of this map to be included in $\mathcal{H}_\gamma$, we need to impose the condition $z_4 > -\gamma/2$. In this case, the natural parameter space is thus $\mathcal{Z} = \{z \in R^4 : z_4 \neq 0, z_4 > -\gamma/2\}$. However, as we shall see below, the results are uniform w.r.t. γ . Note that the mapping G indeed is smooth, and for $z_4 \neq 0$, G and G'_z are also injective.

In the degenerate case $z_4 = 0$, we have

$$G(z, x) = z_1 + z_2 + z_3 x, \quad (59)$$

We return to this case below.

3.5.2 The Hull–White and Ho–Lee models

As our test case, we analyze the Hull and White (1990) (henceforth HW) extension of the Vasiček model. On short rate form the model is given by

$$dR(t) = \{\Phi(t) - aR(t)\} dt + \sigma dW(t), \quad (60)$$

where $a, \sigma > 0$. As is well known, the corresponding forward rate formulation is

$$dr(t, x) = \beta(t, x)dt + \sigma e^{-ax} dW_t. \quad (61)$$

Thus, the volatility function is given by $\sigma(x) = \sigma e^{-ax}$, and the conditions of Theorem 3.11 become

$$G'_x(z, x) + \frac{\sigma^2}{a} [e^{-ax} - e^{-2ax}] \in \text{Im}[G'_z(z, x)], \quad (62)$$

$$\sigma e^{-ax} \in \text{Im}[G'_z(z, x)]. \quad (63)$$

To investigate whether the NS manifold is invariant under HW dynamics, we start with (63) and fix a z -vector. We then look for constants (possibly depending on z) A , B , C , and D , such that for all $x \geq 0$ we have

$$\sigma e^{-ax} = A + B e^{-z_4 x} + C x e^{-z_4 x} - D(z_2 + z_3 x) x e^{-z_4 x}. \quad (64)$$

This is possible if and only if $z_4 = a$, and since (63) must hold for all choices of $z \in \mathcal{Z}$ we immediately see that HW is inconsistent with the full NS manifold (see also the Notes below).

Proposition 3.13 (Nelson–Siegel and Hull–White) *The Hull–White model is inconsistent with the NS family.*

We have thus obtained a negative result for the HW model. The NS manifold is “too small” for HW, in the sense that if the initial forward rate curve is on the manifold, then the HW dynamics will force the term structure off the manifold within an arbitrarily short period of time. For more positive results see Björk and Christensen (1999).

Remark 3.14 It is an easy exercise to see that the minimal manifold which is consistent with HW is given by

$$G(z, x) = z_1 e^{-ax} + z_2 e^{-2ax}.$$

In the same way, one may easily test the consistency between NS and the model obtained by setting $a = 0$ in (60). This is the continuous time limit of the Ho and Lee model (Ho and Lee (1986)), and is henceforth referred to as HL. Since we have a pedagogical point to make, we give the results on consistency, which are as follows.

Proposition 3.15 (Nelson–Siegel and Ho–Lee)

- (a) *The full NS family is inconsistent with the Ho–Lee model.*
 (b) *The degenerate family $G(z, x) = z_1 + z_3x$ is in fact consistent with Ho–Lee.*

Remark 3.16 We see that the minimal invariant manifold provides information about the model. From the result above, the HL model is closely tied to the class of **affine** forward rate curves. Such curves are unrealistic from an economic point of view, implying that the HL model is overly simplistic.

3.6 Notes

The section is based on Björk and Christensen (1999). As we very easily detected above, neither the HW nor the HL model is consistent with the Nelson–Siegel family of forward rate curves. A much more difficult problem is to determine whether **any** interest rate model is. This is Problem II in Section 3.1 for the NS family, and it has been solved recently (using different techniques) in Filipović (1998a), where it is shown that no nontrivial Wiener driven model is consistent with NS. Thus, for a model to be consistent with Nelson–Siegel, it must be deterministic. In Filipović (1998b) (which is a technical *tour de force*) this result is extended to a much larger exponential polynomial family than the NS family. In our presentation we have used strong solutions of the infinite dimensional forward rate SDE. This is of course restrictive. The invariance problem for weak solutions has recently been studied in Filipović (1999). An alternative way of studying invariance is by using some version of the Stroock–Varadhan support theorem, and this line of thought is carried out in depth in Zabczyk (1992).

4 Existence of nonlinear realizations

We now turn to Problem 2 in Section 1.2, i.e. the problem of when a given forward rate model has a finite dimensional factor realization. For ease of exposition we mostly confine ourselves to a discussion of the case of a single driving Wiener process and to time invariant forward rate dynamics. Multidimensional Wiener processes and time varying systems can be treated similarly, and for completeness we state the results for the multidimensional case. We will use some ideas and concepts from differential geometry, and a general reference here is Warner (1979). The section is based on Björk and Svensson (1999).

4.1 Setup

In order to study the realization problem we need (see Remark 4.4) a very regular space to work in.

Definition 4.1 Consider a fixed real number $\gamma > 0$. The space $\mathcal{B}_\gamma$ is defined as the space of all infinitely differentiable functions

$$r : \mathbb{R}_+ \rightarrow \mathbb{R}$$

satisfying the norm condition $\|r\|_\gamma < \infty$. Here the norm is defined as

$$\|r\|_\gamma^2 = \sum_{n=0}^{\infty} 2^{-n} \int_0^\infty \left(\frac{d^n r}{dx^n}(x) \right)^2 e^{-\gamma x} dx.$$

Note that $\mathcal{B}$ is not a space of distributions, but a space of functions. As with $\mathcal{H}$ we will often suppress the subindex γ . With the obvious inner product $\mathcal{B}$ is a pre-Hilbert space, and in Björk and Svensson (1999) the following result is proved.

Proposition 4.2 The space $\mathcal{B}$ is a Hilbert space, i.e. it is complete. Furthermore, every function in the space is fact real analytic, and can thus be uniquely extended to a holomorphic function in the entire complex plane.

We now take as given a volatility $\sigma : \mathcal{B} \rightarrow B$ and consider the induced forward rate model (on Stratonovich form)

$$dr_t = \mu(r_t)dt + \sigma(r_t) \circ dW_t, \quad (65)$$

where as before (see Section 3.4).

$$\mu(r) = \frac{\partial}{\partial x} r + \sigma(r) \mathbf{H} \sigma(r)^* - \frac{1}{2} \sigma'_r(r) \sigma(r). \quad (66)$$

We need some regularity assumptions.

Assumption 4.3 We assume that σ is chosen such that the following hold.

- The mapping σ is smooth.
- The mapping

$$r \mapsto \sigma(r) \mathbf{H} \sigma(r)^* - \frac{1}{2} \sigma'_r(r) \sigma(r)$$

is a smooth map from $\mathcal{B}$ to B .

Remark 4.4 The reason for our choice of $\mathcal{B}$ as the underlying space is that the linear operator $\mathbf{F} = d/dx$ is bounded in this space. Together with the assumptions above, this implies that both μ and σ are smooth vector fields on $\mathcal{B}$, thus ensuring

the existence of a strong local solution to the forward rate equation for every initial point $r^o \in \mathcal{B}$.

4.2 The geometric problem

Given a specification of the volatility mapping σ , and an initial forward rate curve r^o we now investigate when (and how) the corresponding forward rate process possesses a finite dimensional realization. We are thus looking for smooth d -dimensional vector fields a and b , an initial point $z_0 \in R^d$, and a mapping $G : R^d \rightarrow \mathcal{B}$ such that r , locally in time, has the representation

$$dZ_t = a(Z_t)dt + b(Z_t)dW_t, \quad Z_0 = z_0 \quad (67)$$

$$r(t, x) = G(Z_t, x). \quad (68)$$

Remark 4.5 Let us clarify some points. Firstly, note that in principle it may well happen that, given a specification of σ , the r -model has a finite dimensional realization given a particular initial forward rate curve r^o , while being infinite dimensional for all other initial forward rate curves in a neighborhood of r^o . We say that such a model is a **non-generic** or **accidental** finite dimensional model. If, on the other hand, r has a finite dimensional realization for all initial points in a neighborhood of r^o , then we say that the model is a **generically** finite dimensional model. In this text we are solely concerned with the generic problem. Secondly, let us emphasize that we are looking for **local** (in time) realizations.

We can now connect the realization problem to our studies of invariant manifolds.

Proposition 4.6 *The forward rate process possesses a finite dimensional realization if and only if there exists an invariant finite dimensional submanifold $\mathcal{G}$ with $r^o \in \mathcal{G}$.*

Proof See Björk and Christensen (1999) for the full proof. The intuitive argument runs as follows. Suppose that there exists a finite dimensional invariant manifold $\mathcal{G}$ with $r^o \in \mathcal{G}$. Then $\mathcal{G}$ has a local coordinate system, and we may define the Z process as the local coordinate process for the r -process. On the other hand it is clear that if r has a finite dimensional realization as in (67)–(68), then every forward rate curve that will be produced by the model is of the form $x \mapsto G(z, x)$ for some choice of z . Thus there exists a finite dimensional invariant submanifold $\mathcal{G}$ containing the initial forward rate curve r^o , namely $\mathcal{G} = \text{Im } G$. $\square$

Using Theorem 3.11 we immediately obtain the following geometric characterization of the existence of a finite realization.

Corollary 4.7 *The forward rate process possesses a finite dimensional realization if and only if there exists a finite dimensional manifold $\mathcal{G}$ containing r^o , such that, for each $r \in \mathcal{G}$, the following conditions hold:*

$$\mu(r) \in T_{\mathcal{G}}(r),$$

$$\sigma(r) \in T_{\mathcal{G}}(r).$$

Here $T_{\mathcal{G}}(r)$ denotes the tangent space to $\mathcal{G}$ at the point r , and the vector fields μ and σ are as above.

4.3 The main result

Given the volatility vector field σ , and hence also the field μ , we now are faced with the problem of determining whether there exists a finite dimensional manifold $\mathcal{G}$ with the property that μ and σ are tangential to $\mathcal{G}$ at each point of $\mathcal{G}$. In the case when the underlying space is finite dimensional, this is a standard problem in differential geometry, and we will now give the heuristics.

To get some intuition we start with a simpler problem and therefore consider the space $\mathcal{B}$ (or any other Hilbert space), and a smooth vector field f on the space. For each fixed point $r^o \in \mathcal{B}$ we now ask whether there exists a finite dimensional manifold $\mathcal{G}$ with $r^o \in \mathcal{G}$ such that f is tangential to $\mathcal{G}$ at every point. The answer to this question is yes, and the manifold can in fact be chosen to be one-dimensional. To see this, consider the infinite dimensional ODE

$$\frac{dr_t}{dt} = f(r_t), \quad (69)$$

$$r_0 = r^o. \quad (70)$$

If r_t is the solution, at time t , of this ODE, we use the notation

$$r_t = e^{f t} r^o.$$

We have thus defined a group of operators $\{e^{f t} : t \in R\}$, and we note that the set $\{e^{f t} r^o : t \in R\} \subseteq \mathcal{B}$ is nothing else than the integral curve of the vector field f , passing through r^o . If we define $\mathcal{G}$ as this integral curve, then our problem is solved, since f will be tangential to $\mathcal{G}$ by construction.

Let us now take two vector fields f_1 and f_2 as given, where the reader informally can think of f_1 as σ and f_2 as μ . We also fix an initial point $r^o \in \mathcal{B}$ and the question is if there exists a finite dimensional manifold $\mathcal{G}$, containing r^o , with the property that f_1 and f_2 are both tangential to $\mathcal{G}$ at each point of $\mathcal{G}$. We call such a manifold a **tangential manifold** for the vector fields. At a first glance it would seem that there always exists a tangential manifold, and that it can even be chosen to be two-dimensional. The geometric idea is that we start at r^o and let f_1 generate the

integral curve $\{e^{f_1 s} r^o : s \geq 0\}$. For each point $e^{f_1 s} r^o$ on this curve we now let f_2 generate the integral curve starting at that point. This gives us the object $e^{f_2 t} e^{f_1 s} r^o$ and thus it seems that we sweep out a two-dimensional surface $\mathcal{G}$ in $\mathcal{B}$. This is our obvious candidate for a tangential manifold.

In the general case this idea will, however, not work, and the basic problem is as follows. In the construction above we started with the integral curve generated by f_1 and then applied f_2 , and there is of course no guarantee that we will obtain the same surface if we start with f_2 and then apply f_1 . We thus have some sort of commutativity problem, and the key concept is the **Lie bracket**.

Definition 4.8 *Given smooth vector fields f and g on $\mathcal{B}$, the Lie bracket $[f, g]$ is a new vector field defined by*

$$[f, g](r) = f'(r)g(r) - g'(r)f(r). \quad (71)$$

The Lie bracket measures the lack of commutativity on the infinitesimal scale in our geometric program above, and for the procedure to work we need a condition which says that the lack of commutativity is “small”. It turns out that the relevant condition is that the Lie bracket should be in the linear hull of the vector fields.

Definition 4.9 *Let $f_1, \dots, f_n$ be smooth independent vector fields on some space X . Such a system is called a **distribution**, and the distribution is said to be **involutive** if*

$$[f_i, f_j](x) \in \text{span}\{f_1(x), \dots, f_n(x)\}, \quad \forall i, j,$$

where the span is the linear hull over the real numbers.

We now have the following basic result, which extends a classic result from finite dimensional differential geometry (see Warner (1979)).

Theorem 4.10 (Frobenius) *Let $f_1, \dots, f_k$ be independent smooth vector fields in $\mathcal{B}$ and consider a fixed point $r^o \in \mathcal{B}$. Then the following statements are equivalent.*

- *For each point r in a neighborhood of r^o , there exists a k -dimensional tangential manifold passing through r .*
- *The system $f_1, \dots, f_k$ of vector fields is (locally) involutive.*

Proof See Björk and Svensson (1999), which provides a self contained proof of the Frobenius theorem in Banach space. □

Let us now go back to our interest rate model. We are thus given the vector fields μ, σ , and an initial point r^o , and the problem is whether there exists a finite dimensional tangential manifold containing r^o . Using the infinite dimensional

Frobenius theorem, this situation is now easily analyzed. If $\{\mu, \sigma\}$ is involutive then there exists a two-dimensional tangential manifold. If $\{\mu, \sigma\}$ is not involutive, this means that the Lie bracket $[\mu, \sigma]$ is not in the linear span of μ and σ , so then we consider the system $\{\mu, \sigma, [\mu, \sigma]\}$. If this system is involutive there exists a three-dimensional tangential manifold. If it is not involutive at least one of the brackets $[\mu, [\mu, \sigma]]$, $[\sigma, [\mu, \sigma]]$ is not in the span of $\{\mu, \sigma, [\mu, \sigma]\}$, and we then adjoin this (these) bracket(s). We continue in this way, forming brackets of brackets, and adjoining these to the linear hull of the previously obtained vector fields, until the point when the system of vector fields thus obtained actually is closed under the Lie bracket operation.

Definition 4.11 *Take the vector fields $f_1, \dots, f_k$ as given. The **Lie algebra** generated by $f_1, \dots, f_k$ is the smallest linear space (over R) of vector fields which contains $f_1, \dots, f_k$ and is closed under the Lie bracket. This Lie algebra is denoted by*

$$\mathcal{L} = \{f_1, \dots, f_k\}_{\text{LA}}$$

The **dimension** of $\mathcal{L}$ is defined, for each point $r \in \mathcal{B}$, as

$$\dim [\mathcal{L}(r)] = \dim \text{span} \{f_1(r), \dots, f_k(r)\}.$$

Putting all these results together, we have the following main result on finite dimensional realizations.

Theorem 4.12 (Main result) *Take the volatility mapping $\sigma = (\sigma_1, \dots, \sigma_m)$ as given. Then the forward rate model generated by σ generically admits a finite dimensional realization if and only if*

$$\dim \{\mu, \sigma_1, \dots, \sigma_m\}_{\text{LA}} < \infty$$

in a neighborhood of r^o .

The result above thus provides a general solution to Problem II from Section 1.2. For any given specification of forward rate volatilities, the Lie algebra can in principle be computed, and the dimension can be checked. Note, however, that the theorem is a pure existence result. If, for example, the Lie algebra has dimension five, then we know that there exists a five-dimensional realization, but the theorem does not directly tell us how to construct a concrete realization. This is the subject of ongoing research. Note also that realizations are not unique, since any diffeomorphic mapping of the factor space R^d onto itself will give a new equivalent realization.

When computing the Lie algebra generated by μ and σ , the following observations are often useful.

Lemma 4.13 *Take the vector fields $f_1, \dots, f_k$ as given. The Lie algebra $\mathcal{L} = \{f_1, \dots, f_k\}_{\text{LA}}$ remains unchanged under the following operations.*

- *The vector field $f_i(r)$ may be replaced by $\alpha(r)f_i(r)$, where α is any smooth nonzero scalar field.*
- *The vector field $f_i(r)$ may be replaced by*

$$f_i(r) + \sum_{j \neq i} \alpha_j(r) f_j(r),$$

where α_j is any smooth scalar field.

Proof The first point is geometrically obvious, since multiplication by a scalar field will only change the length of the vector field f_i , and not its direction, and thus not the tangential manifold. Formally it follows from the “Leibnitz rule” $[f, \alpha g] = \alpha [f, g] - (\alpha' f)g$. The second point follows from the bilinear property of the Lie bracket together with the fact that $[f, f] = 0$. $\square$

4.4 Applications

In this section we give some simple applications of the theory developed above. For more examples and results, see Björk and Svensson (1999).

4.4.1 Constant Volatility

We start with the simplest case, which is when the volatility $\sigma(r, x)$ is a constant vector in $\mathcal{B}$. We are thus back in the framework of Section 2, and we assume for simplicity that we have only one driving Wiener process. Then we have no Stratonovich correction term and the vector fields are given by

$$\begin{aligned} \mu(r, x) &= \mathbf{F}r(x) + \sigma(x) \int_0^x \sigma(s) ds, \\ \sigma(r, x) &= \sigma(x). \end{aligned}$$

where as before $\mathbf{F} = \frac{\partial}{\partial x}$.

The Fréchet derivatives are trivial in this case. Since $\mathbf{F}$ is linear (and bounded in our space), and σ is constant as a function of r , we obtain

$$\begin{aligned} \mu'_r &= \mathbf{F}, \\ \sigma'_r &= 0. \end{aligned}$$

Thus the Lie bracket $[\mu, \sigma]$ is given by

$$[\mu, \sigma] = \mathbf{F}\sigma,$$

and in the same way we have

$$[\mu, [\mu, \sigma]] = \mathbf{F}^2 \sigma.$$

Continuing in the same manner it is easily seen that the relevant Lie algebra $\mathcal{L}$ is given by

$$\mathcal{L} = \{\mu, \sigma\}_{\text{LA}} = \text{span} \{\mu, \sigma, \mathbf{F}\sigma, \mathbf{F}^2\sigma, \dots\} = \text{span} \{\mu, \mathbf{F}^n \sigma ; n = 0, 1, 2, \dots\}.$$

It is thus clear that $\mathcal{L}$ is finite dimensional (at each point r) if and only if the function space

$$\text{span} \{\mathbf{F}^n \sigma ; n = 0, 1, 2, \dots\}$$

is finite dimensional. We have thus obtained our old condition from Proposition 2.12 and we have the following result which extends Proposition 2.4 by in principle allowing the realization to be nonlinear.

Proposition 4.14 *Under the above assumptions, there exists a finite dimensional realization if and only if σ is a quasi-exponential function.*

4.4.2 Constant Direction Volatility

We go on to study the most natural extension of the deterministic volatility case (still in the case of a scalar Wiener process), namely the case when the volatility is of the form

$$\sigma(r, x) = \varphi(r)\lambda(x). \quad (72)$$

In this case the individual vector field σ has the constant direction $\lambda \in \mathcal{H}$, but is of varying length, determined by φ , where φ is allowed to be any smooth functional of the entire forward rate curve. In order to avoid trivialities we make the following assumption.

Assumption 4.15 *We assume that $\varphi(r) \neq 0$ for all $r \in \mathcal{H}$.*

After a simple calculation the drift vector μ turns out to be

$$\mu(r) = \mathbf{F}r + \varphi^2(r)D - \frac{1}{2}\varphi'(r)[\lambda]\varphi(r)\lambda, \quad (73)$$

where $\varphi'(r)[\lambda]$ denotes the Fréchet derivative $\varphi'(r)$ acting on the vector λ , and where the constant vector $D \in \mathcal{H}$ is given by

$$D(x) = \lambda(x) \int_0^x \lambda(s)ds.$$

We now want to know under what conditions on φ and λ we have a finite dimensional realization, i.e. when the Lie algebra generated by

$$\begin{aligned}\mu(r) &= \mathbf{F}r + \varphi^2(r)D - \frac{1}{2}\varphi'(r)[\lambda]\varphi(r)\lambda, \\ \sigma(r) &= \varphi(r)\lambda,\end{aligned}$$

is finite dimensional. Under Assumption 4.15 we can use Lemma 4.13, to see that the Lie algebra is in fact generated by the simpler system of vector fields

$$\begin{aligned}f_0(r) &= \mathbf{F}r + \Phi(r)D, \\ f_1(r) &= \lambda,\end{aligned}$$

where we have used the notation

$$\Phi(r) = \varphi^2(r).$$

Since the field f_1 is constant, it has zero Fréchet derivative. Thus the first Lie bracket is easily computed as

$$[f_0, f_1](r) = \mathbf{F}\lambda + \Phi'(r)[\lambda]D.$$

The next bracket to compute is $[[f_0, f_1], f_1]$ which is given by

$$[[f_0, f_1], f_1] = \Phi''(r)[\lambda; \lambda]D.$$

Note that $\Phi''(r)[\lambda; \lambda]$ is the second order Fréchet derivative of Φ operating on the vector pair $[\lambda; \lambda]$. This pair is to be distinguished (notice the semicolon) from the Lie bracket $[\lambda, \lambda]$ (with a comma), which of course would be equal to zero. We now make a further assumption.

Assumption 4.16 *We assume that $\Phi''(r)[\lambda; \lambda] \neq 0$ for all $r \in \mathcal{H}$.*

Given this assumption we may again use Lemma 4.13 to see that the Lie algebra is generated by the following vector fields

$$\begin{aligned}f_0(r) &= \mathbf{F}r, \\ f_1(r) &= \lambda, \\ f_3(r) &= \mathbf{F}\lambda, \\ f_4(r) &= D.\end{aligned}$$

Of these vector fields, all but f_0 are constant, so all brackets are easy. After elementary calculations we see that in fact

$$\{\mu, \sigma\}_{\text{LA}} = \text{span} \{ \mathbf{F}r, \mathbf{F}^n \lambda, \mathbf{F}^n D; n = 0, 1, \dots \}.$$

From this expression it follows immediately that a necessary condition for the Lie algebra to be finite dimensional is that the vector space spanned by $\{\mathbf{F}^n \lambda; n \geq 0\}$ is finite dimensional. This occurs if and only if λ is quasi-exponential (see Remark 2.5). If, on the other hand, λ is quasi-exponential, then we know from Lemma 2.6, that D is also quasi-exponential, since it is the integral of the QE function λ multiplied by the QE function λ . Thus the space $\{\mathbf{F}^n D; n = 0, 1, \dots\}$ is also finite dimensional, and we have proved the following result.

Proposition 4.17 *Under Assumptions 4.15 and 4.16, the interest rate model with volatility given by $\sigma(r, x) = \varphi(r)\lambda(x)$ has a finite dimensional realization if and only if λ is a quasi-exponential function. The scalar field φ is allowed to be any smooth field.*

4.4.3 When is the Short Rate a Markov Process?

One of the classical problems concerning the HJM approach to interest rate modeling is that of determining when a given forward rate model is realized by a short rate model, i.e. when the short rate is Markovian. We now briefly indicate how the theory developed above can be used in order to analyze this question. For the full theory see Björk and Svensson (1999).

Using the results above, we immediately have the following general necessary condition.

Proposition 4.18 *The forward rate model generated by σ is a generic short rate model, i.e. the short rate is generically a Markov process, only if*

$$\dim \{\mu, \sigma\}_{\text{LA}} \leq 2. \quad (74)$$

Proof If the model is really a short rate model, then bond prices are given as $p(t, x) = F(t, R_t, x)$ where F solves the term structure PDE. Thus bond prices, and forward rates are generated by a two-dimensional factor model with time t and the short rate R as the state variables. $\square$

Remark 4.19 The most natural case is $\dim \{\mu, \sigma\}_{\text{LA}} = 2$. It is an open problem whether there exists a non-deterministic generic short rate model with $\dim \{\mu, \sigma\}_{\text{LA}} = 1$.

Note that condition (74) is only a necessary condition for the existence of a short rate realization. It guarantees that there exists a two-dimensional realization, but the question remains whether the realization can be chosen in such a way that the short rate and running time are the state variables. This question is completely resolved by the following central result.

Theorem 4.20 *Assume that the model is not deterministic, and take as given a time invariant volatility $\sigma(r, x)$. Then there exists a short rate realization if and only if the vector fields $[\mu, \sigma]$ and σ are parallel, i.e. if and only if there exists a scalar field $\alpha(r)$ such that the following relation holds (locally) for all r .*

$$[\mu, \sigma](r) = \alpha(r)\sigma(r). \quad (75)$$

Proof See Björk and Svensson (1999). □

It turns out that the class of generic short rate models is very small indeed. We have, in fact, the following result, which was first proved in Jeffrey (1995) (using techniques different from those above). See Björk and Svensson (1999) for a proof based on Theorem 4.20.

Theorem 4.21 *Consider an HJM model with one driving Wiener process and a volatility structure of the form*

$$\sigma(r, x) = g(R, x).$$

where $R = r(0)$ is the short rate. Then the model is a generic short rate model if and only if g has one of the following forms.

- *There exists a constant c such that*

$$g(R, x) \equiv c.$$

- *There exist constants a and c such that.*

$$g(R, x) = ce^{-ax}.$$

- *There exist constants a and b , and a function $\alpha(x)$, where α satisfies a certain Riccati equation, such that*

$$g(R, x) = \alpha(x)\sqrt{aR + b}.$$

We immediately recognize these cases as the Ho–Lee model, the Hull–White extended Vasicek model, and the Hull–White extended Cox–Ingersoll–Ross model (Cox, Ingersoll and Ross (1985)). Thus, in this sense the only generic short rate models are the affine ones, and the moral of this, perhaps somewhat surprising, result is that most short rate models considered in the literature are not generic but “accidental”. To understand the geometric picture one can think of the following program.

1. Choose an arbitrary short rate model, say of the form

$$dR_t = a(R_t)dt + b(R_t)dW_t$$

with a fixed initial point R_0 .

2. Solve the associated PDE in order to compute bond prices. This will also produce:
 - An initial forward rate curve $\hat{r}^o(x)$.
 - Forward rate volatilities of the form $g(R, x)$.
3. Forget about the underlying short rate model, and take the forward rate volatility structure $g(R, x)$ as given in the forward rate equation.
4. Initiate the forward rate equation with an arbitrary initial forward rate curve $r^o(x)$.

The question is now whether the thus constructed forward rate model will produce a Markovian short rate process. Obviously, if you choose the initial forward rate curve r^o as $r^o = \hat{r}^o$, then you are back where you started, and everything is OK. If, however, you choose another initial forward rate curve rather than $\hat{r}^o$, say the observed forward rate curve of today, then it is no longer clear that the short rate will be Markovian. What the theorem above says is that only the models listed above will produce a Markovian short rate model for all initial points in a neighborhood of $\hat{r}^o$. If you take another model (like, say, the Dothan model) then a generic choice of the initial forward rate curve will produce a short rate process which is not Markovian.

4.5 Notes

The section is based on Björk and Svensson (1999) where full proofs and further results can be found, and where also the time varying case is considered. In our study of the constant direction model above, φ was allowed to be any smooth functional of the entire forward rate curve. The simpler special case when φ is a point evaluation of the short rate, i.e. of the form $\varphi(r) = h(r(0))$ has been studied in Bhar and Chiarella (1997), Inui and Kijima (1998) and Ritchken and Sankarasubramanian (1995). All these cases falls within our present framework and the results are included as special cases of the general theory above. A different case, treated in Chiarella and Kwon (1998), occurs when σ is a finite point evaluation, i.e. when $\sigma(t, r) = h(t, r(x_1), \dots, r(x_k))$ for fixed benchmark maturities $x_1, \dots, x_k$. In Chiarella and Kwon (1998) it is studied when the corresponding finite set of benchmark forward rates is Markovian.

A classic paper on Markovian short rates is Carverhill (1994), where a deterministic volatility of the form $\sigma(t, x)$ is considered. Theorem 4.21 was first stated and proved in Jeffrey (1995). See Eberlein and Raible (1999) for an example with a driving Lévy process.

The geometric ideas presented above and in Björk and Svensson (1999) are intimately connected to controllability problems in systems theory, where they

have been used extensively (see Isidori (1989)). They have also been used in filtering theory, where the problem is to find a finite dimensional realization of the unnormalized conditional density process, the evolution of which is given by the Zakai equation. See Brockett (1981) for an overview of these areas.

References

- Bhar, R. and Chiarella, C. (1997), Transformation of Heath–Jarrow–Morton models to markovian systems. *European Journal of Finance* **3**, 1, 1–26.
- Björk, T. (1997), Interest Rate Theory. In W. Runggaldier (ed.), *Financial Mathematics*. Springer Lecture Notes in Mathematics, Vol. 1656. Springer-Verlag, Berlin.
- Björk, T. and Christensen, B.J. (1999), Interest rate dynamics and consistent forward rate curves. *Mathematical Finance* **9**, 4, 323–48.
- Björk, T. and Gombani, A. (1999), Minimal realization of interest rate models. *Finance and Stochastics* **3**, 4, 413–32.
- Björk, T. and Svensson, L. (1999), On the existence of finite dimensional nonlinear realizations of interest rate models. Forthcoming in *Mathematical Finance*.
- Brace, A. and Musiela, M. (1994), A multi factor Gauss Markov implementation of Heath Jarrow and Morton. *Mathematical Finance* **4**, 3, 563–76.
- Brockett, R.W. (1970), *Finite Dimensional Linear Systems*. Wiley, New York.
- Brockett, R.W. (1981), Nonlinear systems and nonlinear estimation theory. In *Stochastic systems: The Mathematics of Filtering and Identification and Applications* (eds. Hazewinkel, M and Willems, J.C.) Reidel, Dordrecht.
- Carverhill, A. (1994), When is the spot rate Markovian? *Mathematical Finance*, **4**, 305–12.
- Chiarella, C and Kwon, K. (1998), Forward rate dependent Markovian transformations of the Heath–Jarrow–Morton term structure model. Working paper. School of Finance and Economics, University of Technology, Sydney.
- Cox, J., Ingersoll, J. and Ross, S. (1985), A theory of the term structure of interest rates. *Econometrica* **53**, 385–408.
- Da Prato, G. and Zabczyk, J. (1992), *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, Cambridge.
- Duffie, D. and Kan, R. (1996), A yield factor model of interest rates. *Mathematical Finance*, **6**, 379–406.
- Eberlein, E. and Raible, S. (1999), Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31–53.
- El Karoui, N. and Lacoste, V (1993), Multifactor models of the term structure of interest rates. Preprint.
- El Karoui, N., Geman, H. and Lacoste, V (1997), On the role of state variables in interest rate models. Preprint
- Filipović, D. (1998a): A note on the Nelson–Siegel family. *Mathematical Finance* **9**, 4, 349–59.
- Filipović, D. (1998b): Exponential–polynomial families and the term structure of interest rates. To appear in *Bernoulli*.
- Filipović, D. (1999), Invariant manifolds for weak solutions of stochastic equations. To appear in *Probability Theory and Related Fields*.
- Heath, D., Jarrow, R. and Morton, A. (1992), Bond pricing and the term structure of interest rates. *Econometrica* **60** 1, 77–106.

- Ho, T. and Lee, S. (1986), Term structure movements and pricing interest rate contingent claims. *Journal of Finance* **41**, 1011–29.
- Hull, J. and White, A. (1990), Pricing interest-rate-derivative securities. *The Review of Financial Studies* **3**, 573–92.
- Inui, K. and Kijima, M. (1998), A markovian framework in multi-factor Heath–Jarrow–Morton models. *JFQA* **333** 3, 423–40.
- Isidori, A. (1989), *Nonlinear Control Systems*. Springer-Verlag, Berlin.
- Jeffrey, A. (1995), Single factor Heath–Jarrow–Morton term structure models based on Markovian spot interest rates. *JFQA* **30** 4, 619–42.
- Musiela, M. (1993), Stochastic PDEs and term structure models. Preprint.
- Musiela, M. and Rutkowski, M. (1997), *Martingale Methods in Financial Modeling*. Springer-Verlag, Berlin, Heidelberg, New York.
- Nelson, C. and Siegel, A. (1987), Parsimonious modelling of yield curves. *Journal of Business*, **60**, 473–89.
- Ritchken, P. and Sankarasubramanian, L. (1995), Volatility structures of forward rates and the dynamics of the term structure. *mathematical Finance*, **5**, 1, 55–72.
- Vasiček, O. (1977), An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–88.
- Warner, F.W. (1979), *Foundations of Differentiable Manifolds and Lie Groups*. Scott, Foresman, Hill.
- Zabczyk, J. (1992), Stochastic invariance and consistency of financial models. Preprint. Scuola Normale Superiore, Pisa.

Towards a Central Interest Rate Model

Alan Brace, Tim Dun and Geoff Barton

1 Introduction

In recent years, the appearance of a new class of term structure of interest rate models has attracted the interest of practitioners. These so-called Market Models provide both an arbitrage-free pricing framework and pricing formulae that conform to the current (and accepted) market practice.

This class of model can effectively be split into two types: those that model forward Libor rates, and those that model forward swap rates. The Libor rate models, such as those introduced in Miltersen *et al.* (1997), Brace *et al.* (1997) and Musiela and Rutkowski (1997a,b), allow caps to be priced in a manner consistent with market practice, while the swap rate models, such as the one proposed by Jamshidian (1997), do the same for swaptions. However, these two approaches are fundamentally incompatible because Libor rates and swap rates cannot both be lognormal in an arbitrage-free framework.

The formulae currently in use in the market are based on extensions of the well-known Black–Scholes option formula, and are, in fact, known as the Black cap and swaption formulae. In the case of swaptions, the swap rate replaces the stock price as being the market observable parameter assumed to follow lognormal dynamics. Other concepts that are related to (and easily calculated using) the Black–Scholes option formula can also be extended to the case of swaptions, such as the option sensitivities or *Greeks*. These give an indication as to the likely magnitude and direction of the change in option price under changes in the swap rate value and/or volatility.

The Black formulae, however, are incapable of producing arbitrage-free prices for exotics, nor are they of much use as a ‘central’ interest rate model to do bank-wide risk management. These shortfalls constitute the original motivation for the development of term structure models. So how do the two types of Market Model mentioned above perform in these areas?

When pricing exotics, the natural tendency is to choose the most appropriate model for the task, hence Libor models for Libor based exotics, such as barrier caps triggered by Libor, and swap rate models for swap rate based exotics, such as barrier swaptions triggered by the swap rate. The case of cross-market exotics, however, is not so simple – how does one treat barrier swaptions triggered by Libor, and how does one calibrate simultaneously to both cap and swaption markets?

In the authors' opinion, the Libor model is the unifying model – the Central Interest Rate Model – capable of encompassing the global properties of the swap rate model and tackling the problems related above. This is primarily because it is the most tractable mathematically, with Libor rates being lognormal under their own measures, without the restriction of only certain families of swap rates being lognormal. The model also prices swaptions and swap rate exotics, and, as we intend to argue in this paper, in practice it prices swaptions in a manner close to that of the market – and by extension – to the forward swap rate model. This indicates a closeness between the two types of Market Model.¹ We propose in this study, therefore, to examine the Libor model and its ability to price and hedge pure swap market products in comparison to the Black swaption formula, under arbitrary yield and volatility specifications, with the aim of revealing the closeness of the two approaches.

Our methodology is as follows. First, in Section 2, the notation and equations involved in swaption pricing within the Libor model are introduced. The Black swaption formula is also presented, along with the equations necessary to calculate the swaption Greeks and hedge swaptions. In Section 3, the actual distributional properties of the swap rate within the Libor model are examined analytically, to see if it cannot be approximately modelled by a lognormal process. An expression is then derived for the volatility of this swap rate allowing the approximate pricing of swaptions inside the Libor model using a Black type formula. In Section 4, approximation techniques are applied to derive equations inside the Libor model for swaption Greeks with respect to the swap rate. Here, only approximate relations at best may be expected, since in the Libor model, the swap rate is a weighted sum of Libor rates, and not a single quantity as implied by the Black formula. These Greeks will, however, provide us with another mechanism for comparing the swaption modelling capabilities of the Libor model. Simulation techniques are then used to test the approximations from Sections 3 and 4 on a range of swaptions for two quite different volatility structures, with the results presented in Section 5. Tests are carried out to determine if the swaption Greeks derived are meaningful by undertaking a delta-hedging simulation and seeing if Libor model swaptions can be

¹ This closeness was first alluded to in the observation in Brace *et al.* (1997) that the Libor model swaption formula essentially reduces to the Black formula when yield and volatility are flat. Other authors to examine this behaviour include Jamshidian (1997) and Rebonato (1999).

successfully hedged within the Libor model framework using Black-style hedging techniques. The results from these tests are also presented in Section 5. Finally, Section 6 states our conclusions on the work done, while the appendices contain additional results, both numerical and mathematical, for the interested reader.

2 Model preliminaries

In this section, we introduce the fundamental equations behind the lognormal Libor model, together with swap and swaption pricing within this model. The equivalent market pricing equations are then presented, and option sensitivities (or Greeks) defined. The section ends with a description of a method for translating the Greeks into actual hedges. Note that all the definitions, results and formulae in this section hold for both single and multi-factor models.

2.1 Lognormal Libor model

We consider the discrete tenor version of the lognormal forward Libor model, as described in Musiela and Rutkowski (1997a,b), and Jamshidian (1997), as opposed to the continuous tenor model in Brace *et al.* (1997).

We start with an equi-spaced tenor structure defined by

$$T_j = T_0 + j\delta \text{ for } j = 1, \dots, n$$

where δ is a constant typically of value three or six months. Time t values of zero coupon bonds expiring on the tenor dates are expressed as $P(t, T_j)$, while the forward time T price for a zero coupon bond maturing at $T_j \geq T$ is

$$F_T(t, T_j) = \frac{P(t, T_j)}{P(t, T)}.$$

The forward Libor rate $K(t, T_j)$, expressing the simple forward interest rate between tenor dates T_j and T_{j+1} , is related to the zero coupon bonds by

$$K(t, T_j) = \frac{1}{\delta} \left(\frac{P(t, T_j)}{P(t, T_{j+1})} - 1 \right).$$

We assume that we are equipped with a complete filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying the ‘usual conditions’ (see Chapter 14 in Musiela and Rutkowski (1997a)). The dynamics of the forward Libor processes are then described by the stochastic differential equation

$$dK(t, T_{j-1}) = K(t, T_{j-1})\gamma(t, T_{j-1}) \cdot dW_{T_j}(t) \quad (1)$$

where $\gamma(t, T_{j-1})$ is the forward Libor volatility function, and W_{T_j} represents Brownian motion under the $\mathbb{P}$ -equivalent forward measure $\mathbb{P}_{T_j}$. Adjacent forward measures are related by

$$dW_{T_j}(t) = dW_{T_{j-1}}(t) + \frac{\delta K(t, T_{j-1})}{1 + \delta K(t, T_{j-1})} \gamma(t, T_{j-1}) dt. \quad (2)$$

Consider now a forward payer swap, paid in arrears, with n equal rolls starting at time T_0 . In terms of zero coupon bonds, Libor rates and a strike value κ , the time t value of the swap $Pswap(t)$ can be written as

$$Pswap(t) = Pswap(t, T_0, n) = \delta \sum_{j=1}^n P(t, T_j) (K(t, T_{j-1}) - \kappa). \quad (3)$$

The swap rate $\omega(t)$ is that unique value of the strike which gives the swap contract zero value, and is given by

$$\omega(t) = \omega(t, T_0, n) = \frac{\sum_{j=1}^n P(t, T_j) K(t, T_{j-1})}{\sum_{j=1}^n P(t, T_j)} = \frac{\sum_{j=1}^n F_{T_0}(t, T_j) K(t, T_{j-1})}{\sum_{j=1}^n F_{T_0}(t, T_j)}. \quad (4)$$

A swaption is formally defined as an option maturing at time T_0 , on an underlying swap with strike κ . If the swap rate is greater than the strike at option maturity, then the swaption pays the difference between the two rates. The swaption price can, therefore, be expressed as

$$Pswpn(t) = \delta \sum_{j=1}^n P(t, T_j) \mathbb{E}_{T_j} \left\{ (K(T, T_{j-1}) - \kappa) \mathbb{I}(A) \mid \mathcal{F}_t \right\} \quad (5)$$

where $A = \{Swap(T) \geq 0\}$ is the event that the swap ends up in-the-money. This expression does not allow an analytic solution, however a good approximation can be found following the approach in Brace *et al.* (1997) or Brace (1996). This approximation was originally derived for the continuous tenor version of the model, however it is equally valid in the discrete tenor model as no dates outside of the discrete tenor structure appear in the formulae.

Define the n -dimensional random vector

$$X = (X_j) \stackrel{def}{=} \left(\int_t^{T_0} \gamma(s, T_{j-1}) \cdot dW_{T_j}(s) \right)$$

and approximate it by a Gaussian random vector by using a deterministic approximation (here a Wiener chaos expansion of order 0) to the stochastic drift term in (2). The mean vector μ and covariance matrix λ of our approximation under the $\mathbb{P}_{T_0}$ -measure are then given by

$$X \sim \mathbf{N}(\mu, \lambda),$$

$$\begin{aligned}\mu &= (\mu_j) = \left(\sum_{i=1}^j \frac{\delta K(t, T_{i-1})}{1 + \delta K(t, T_{i-1})} \lambda_{ij} \right), \\ \lambda &= (\lambda_{ij}) = \left(\int_t^{T_0} \gamma(s, T_{i-1}) \cdot \gamma(s, T_{j-1}) ds \right),\end{aligned}\quad (6)$$

where $\mathbf{N}(\cdot)$ represents the multi-dimensional Gaussian cumulative distribution function.

We find in practice that the symmetric matrix λ (which we will term the *swaption covariance matrix*) is often of rank one, meaning that it can be expressed as the cross product of a vector with itself, as in $\lambda = \Gamma \times \Gamma^T$. Such a decomposition can be easily found through an eigenvector/eigenvalue analysis of the matrix.

Using this rank one approximation Γ , we find the value of s satisfying the relation

$$\sum_{j=1}^n \frac{K(t, T_{j-1}) \exp(\Gamma_j(s + d_j) - \frac{1}{2}\Gamma_j^2) - \kappa}{\prod_{i=1}^j \left(1 + \delta K(t, T_{i-1}) \exp(\Gamma_i(s + d_i) - \frac{1}{2}\Gamma_i^2) \right)} = 0 \quad (7)$$

with

$$d_j = \sum_{i=1}^j \frac{\delta K(t, T_{i-1})}{1 + \delta K(t, T_{i-1})} \Gamma_i,$$

and the approximate swaption price is then given by

$$P_{swpn}(t) \approx \delta \sum_{j=1}^n P(t, T_j) \left(K(t, T_{j-1}) \mathbf{N}(h_j) - \kappa \mathbf{N}(h_j - \Gamma_j) \right) \quad (8)$$

where

$$h_j = -(s + d_j - \Gamma_j). \quad (9)$$

Equation (8) provides an accurate approximation as long as the assumption holds that the covariance matrix λ is of rank one. This assumption and its implications are discussed in more detail in Sections 4.1, 5.3 and 5.5.

2.2 Market swaption formula

In the Market (or Black) swaption pricing formula, swap rates are implicitly assumed lognormal under a single measure $\mathbb{P}_m$. For a swap of n rolls, maturing at time T_0 , this implies the following relation between the forward swap rate $\omega(t) = \omega(t, T_0, n)$ and its associated volatility $\sigma(t) = \sigma(t, T_0, n)$:

$$d\omega(t) = \omega(t)\sigma(t) \cdot dW(t),$$

where $W(t)$ is Brownian motion under $\mathbb{P}_m$. In terms of $\omega(t)$, the present values of a payer swap and corresponding payer swaption are

$$\begin{aligned}
 P_{\text{swap}}(t) &= P_{\text{swap}}(t, T_0, n) = \delta \sum_{j=1}^n P(t, T_j) (\omega(t) - \kappa), \\
 P_{\text{swpn}}(t) &= P_{\text{swpn}}(t, T_0, n) = \delta \sum_{j=1}^n P(t, T_j) \mathbb{E} \{ (\omega(T_0) - \kappa)^+ | \mathcal{F}_t \} \\
 &= \delta \sum_{j=1}^n P(t, T_j) \mathbf{B}(t),
 \end{aligned} \tag{10}$$

where $\mathbf{B}(t)$ is Black's call formula

$$\mathbf{B}(t) = \omega(t) \mathbf{N}(h) - \kappa \mathbf{N}(h - \sqrt{\zeta}), \tag{11}$$

in this case with

$$\begin{aligned}
 h &= \frac{\ln \frac{\omega(t)}{\kappa} + \frac{1}{2}\zeta}{\sqrt{\zeta}}, \\
 \zeta &= \int_t^{T_0} |\sigma(s, T_0, n)|^2 ds.
 \end{aligned} \tag{12}$$

We denote the term ζ as the *swaption zeta*, representing a volatility term which also contains information on the time to maturity of the option. We will use it below to define a version of the option vega. For the sake of convenience, we denote the sum $\sum_{j=1}^n \delta P(t, T_j)$ as the *present value of a basis point*, or *PVBP*. In other references this sum has been given various other names, including the *coupon process*, the *level*, or even the *annuity price*.

The definition of sensitivities (or Greeks) for swaptions differs slightly from standard Black–Scholes type options due to the presence of the *PVBP* term and the fact that the swap rate is a forward rather than a spot value. We define, therefore, our Greeks in terms of forward values into the swaption discounted by the *PVBP* – this being a sensible definition in terms of hedging – as will be discussed in Section 2.3. This reduces the expressions for the Greeks to partial derivatives of the Black term $\mathbf{B}(t)$, as in

$$\text{Swaption delta} \quad \Delta = \frac{\partial}{\partial \omega} \left(\frac{P_{\text{swpn}}(t)}{\delta \sum_{j=1}^n P(t, T_j)} \right) = \frac{\partial \mathbf{B}}{\partial \omega} = \mathbf{N}(h), \tag{13}$$

$$\text{Swaption gamma} \quad \Gamma = \frac{\partial^2}{\partial \omega^2} \left(\frac{P_{\text{swpn}}(t)}{\delta \sum_{j=1}^n P(t, T_j)} \right) = \frac{\partial^2 \mathbf{B}}{\partial \omega^2} = \frac{1}{\omega \sqrt{\zeta}} \mathbf{N}'(h), \tag{14}$$

and

$$\text{Swaption vega} \quad \Lambda = \frac{\partial}{\partial \zeta} \left(\frac{P_{\text{swpn}}(t)}{\delta \sum_{j=1}^n P(t, T_j)} \right) = \frac{\partial \mathbf{B}}{\partial \zeta} = \frac{\omega}{2\sqrt{\zeta}} \mathbf{N}'(h), \quad (15)$$

where, as indicated above, we define our vega term slightly differently from the traditional way in that it is the derivative with respect to the swaption zeta, rather than an annualised volatility value as in Black–Scholes. This is done simply to ease computation later. Note that $\mathbf{N}'(\cdot)$ represents the Gaussian density function.

Note also that our gamma and vega are connected by the relation

$$\Lambda = \frac{1}{2} \omega^2 \Gamma, \quad (16)$$

and we would expect our approximate formulae for Γ and Λ in the lognormal Libor model (derived in Section 4) to satisfy this same constraint.

2.3 Swaption hedging

For Black–Scholes type options, the option Δ not only describes the first-order sensitivity of the option value to the underlying, but it also represents the probability of exercise of the option and hence can be used for hedging – giving the required hedge ratio into the underlying. The extension of this to the case of swaptions is complicated by the presence of the *PVBP* discount term in the pricing formula (10), and the fact that the swap rate is not a traded asset. One method² is to hedge using the underlying forward swap and the *PVBP* as the hedging instruments. The hedge then consists of two elements

- a delta hedge of amount $\Delta = \mathbf{N}(h)$ (from Section 2.2) into the underlying forward swap $P_{\text{swap}}(t)$, and
- a bucket hedge of $(\mathbf{B}(t) - \Delta(\omega(t) - \kappa))$ into the *PVBP*.

This produces a portfolio which matches the swaption in value, and – with continual rebalancing – should match the swaption payoff at maturity. Often in practice the swaption is delta-hedged with the underlying swap while the *PVBP* terms are absorbed into the underlying book as cash flows, where they are hedged as part of the general exposure in different time buckets.

3 Swap rate dynamics in the Libor model

The Libor model is deliberately constructed in such a way that the forward Libor rates will be lognormal under certain probability measures – called *forward measures* – induced by using zero coupon bond prices as the numeraire. Similarly

² For other methods see Dudenhausen *et al.* (1998) or Dun *et al.* (1999).

the lognormal swap rate model chooses a specific numeraire so that under the measure it induces the forward swap rates will be lognormal. While this numeraire is quite valid within the Libor model framework, analytic tractability can only be obtained if we know the swap rate dynamics under one of the forward measures. Hence the aim of this section is to investigate the possibility of the swap rate being approximately lognormal under a certain forward measure – in this case the one corresponding to the maturity of the swaption $\mathbb{P}_{T_0}$ – and to find an expression for its corresponding volatility.

3.1 Swap rate measure in the Libor model

The swap rate measure is the one induced by taking the $PVBP = \sum_{j=1}^n \delta P(t, T_j)$ as the numeraire. Under this measure the swap rate $\omega(t, T_0, n)$ will be a martingale. Denoting this measure, and the Brownian motion under it, as $\tilde{\mathbb{P}}_{T_0}$ and $\tilde{W}_{T_0}(t)$ respectively, we can demonstrate the relationship between $\tilde{\mathbb{P}}_{T_0}$ and the Libor model maturity forward measure $\mathbb{P}_{T_0}$ as follows. Taking an arbitrary zero coupon bond $P(t, T_k)$ and applying Itô's lemma to the quotient of it and the $PVBP$, we obtain

$$\begin{aligned} d\left(\frac{P(t, T_k)}{\delta \sum_{j=1}^n P(t, T_j)}\right) &= d\left(\frac{F_{T_0}(t, T_k)}{\delta \sum_{j=1}^n F_{T_0}(t, T_j)}\right) \\ &= \frac{F_{T_0}(t, T_k)}{\delta \sum_{j=1}^n F_{T_0}(t, T_j)} \left(\frac{\sum_{j=1}^n F_{T_0}(t, T_j) \sigma(t, j)}{\sum_{j=1}^n F_{T_0}(t, T_j)} - \sigma(t, k) \right) \\ &\quad \times \left(dW_{T_0}(t) + \frac{\sum_{j=1}^n F_{T_0}(t, T_j) \sigma(t, j)}{\sum_{j=1}^n F_{T_0}(t, T_j)} dt \right), \end{aligned} \quad (17)$$

where we define $\sigma(t, n)$ as the stochastic function

$$\sigma(t, n) = \sum_{i=1}^n \frac{\delta K(t, T_{i-1})}{1 + \delta K(t, T_{i-1})} \gamma(t, T_{i-1}).$$

The expression (17) is a martingale under $\tilde{\mathbb{P}}_{T_0}$, which implies

$$d\tilde{W}_{T_0}(t) = dW_{T_0}(t) + \frac{\sum_{j=1}^n F_{T_0}(t, T_j) \sigma(t, j)}{\sum_{j=1}^n F_{T_0}(t, T_j)} dt, \quad (18)$$

giving us an explicit relation between Brownian motion under the swap rate measure $\tilde{\mathbb{P}}_{T_0}$ and the swaption maturity forward measure $\mathbb{P}_{T_0}$. Further, by applying (2) recursively we arrive at

$$d\tilde{W}_{T_0}(t) = \frac{\sum_{j=1}^n F_{T_0}(t, T_j) dW_{T_j}(t)}{\sum_{j=1}^n F_{T_0}(t, T_j)}, \quad (19)$$

implying not only that $\tilde{\mathbb{P}}_{T_0}$ is an equivalent measure to the forward measures $\mathbb{P}_{T_j}$, but the Brownian motion $\tilde{W}_{T_0}$ under this measure is in fact a weighted average of the W_{T_j} . Given this relationship, and recalling that the swap rate will be a martingale under $\tilde{\mathbb{P}}_{T_0}$, we feel justified in looking for a lognormal approximation to the swap rate $\omega(t, T_0, n)$ under any other of the $\mathbb{P}_{T_j}$, and in particular $\mathbb{P}_{T_0}$. Effectively we are choosing to neglect the drift term in (18), an assertion that we will verify by simulation in Section 5.1. Our next step is, assuming an approximate lognormal swap rate distribution under $\mathbb{P}_{T_0}$, to derive an expression for its volatility.

3.2 Approximate swap rate volatility

As the swap rate definition (4) is effectively a weighted (by forward prices $F_{T_0}(t, T_j) / \sum_{i=1}^n F_{T_0}(t, T_i)$) average of Libor rates $K(t, T_j)$, it seems evident that the contribution to the swap rate volatility by the $K(t, T_j)$ will be significantly greater than that of the $F_{T_0}(t, T_j)$. In fact, in this analysis and much of that which follows, we will assume that the contribution in terms of volatility of the $F_{T_0}(t, T_j)$ is negligible and regard them (and hence also the $P(t, T_j)$) as essentially constant at their initial values. This assumption is tested and justified by simulation means in Section 5.2.

Examining the individual terms which make up the swap rate (4), we see that they are martingales under the T_0 -forward measure $\mathbb{P}_{T_0}$, as demonstrated by Equations (20) and (21) below.

$$\frac{dF_{T_0}(t, T_j)}{F_{T_0}(t, T_j)} = -\sigma(t, j) \cdot dW_{T_0}(t) \quad (20)$$

$$\frac{d(F_{T_0}(t, T_j) K(t, T_{j-1}))}{F_{T_0}(t, T_j) K(t, T_{j-1})} = (\gamma(t, T_{j-1}) - \sigma(t, j)) \cdot dW_{T_0}(t). \quad (21)$$

These terms will become lognormal if the stochastic term $\sigma(t, j)$ is approximated deterministically. In this case, both the numerator and denominator of (4) will be sums of lognormal processes, and these sums will also be approximately lognormal, as in the standard approximations used to price average rate options. Hence, the swap rate $\omega(t, T, n)$, being the ratio of approximate lognormal processes under $\mathbb{P}_{T_0}$, ought to be approximately lognormal itself (with a drift) under the same measure. Following this reasoning, we model the swap rate dynamics under $\mathbb{P}_{T_0}$ as

$$d\omega(t, T, n) = \omega(t, T, n) (\mu(t, T_0, n)dt + \gamma(t, T_0, n) \cdot dW_{T_0}(t)) \quad (22)$$

and, neglecting the volatility contribution of the $F_{T_0}(t, T_j)$ as suggested above, we obtain the following approximate expression for the swap rate volatility $\gamma(t, T_0, n)$

in terms of the Libor rate volatilities $\gamma(t, T_j)$,

$$\begin{aligned}\gamma(t, T_0, n) &= \frac{\sum_{j=1}^n P(0, T_j) K(0, T_{j-1}) \gamma(t, T_{j-1})}{\sum_{j=1}^n P(0, T_j) K(0, T_{j-1})} \\ &= \frac{\sum_{j=1}^n F_{T_0}(0, T_j) K(0, T_{j-1}) \gamma(t, T_{j-1})}{\sum_{j=1}^n F_{T_0}(0, T_j) K(0, T_{j-1})}.\end{aligned}\quad (23)$$

The ability of this equation to predict Libor model swaption volatilities and prices for a given yield curve and Libor volatility function $\gamma(t, T)$ will be tested in Section 5.³

4 Greeks in the Libor model

Another mechanism for assessing the closeness of swaption pricing within the Libor model to the Black swaption formula is through the calculation of the swaption Greeks. In this section we use approximation techniques to derive equations for the swaption delta, gamma and vega under arbitrary volatility specifications.

As seen in Section 2.2, the definition and computation of the swaption delta, gamma and vega are straightforward in the framework implied by the Black swaption formula. Here, the swap rate is a real variable with respect to which we can differentiate, and its corresponding volatility can be expressed likewise – even if the model is multi-factor.

For the Libor model, however, the swap rate is not a single quantity but a forward price-weighted sum of Libor rates – all of which can, to a certain extent, behave independently. This means that we do not have a real central variable with respect to which we can differentiate in order to define and compute swaption Greeks. The Libor rates are, however, related together by the swaption covariance matrix (defined in Section 2.1) and this matrix is often of rank one for both single and multi-factor volatility structures. This effectively implies that the Libor rates can, in fact, be described by a single variable. Taking this idea further, it implies – given the assumption of a rank one covariance matrix – the existence of a variable with which we can differentiate and define Greeks in the Libor model. This notion will be central to our approximation calculations below.

Note that all the equations derived in this section will be examined numerically in Section 5.

³ Note that an equivalent expression to (23) is independently derived by Rebonato (1999) who also employs simulation techniques to verify his results.

4.1 Approximations

Here we give a formal list and explanation of the approximations and assumptions required to derive the equations for the swaption Greeks within the Libor model. Labelling them A1 to A4, we have:

- A1. The discount terms $(F_{T_0}(t, T_j), P(t, T_j))$ are constant at their initial time zero values;
- A2. The swaption covariance matrix is of rank one;
- A3. The volatility function is one-factor separable; and
- A4. The forward probability measures can be merged into one single measure.

Approximation A1 was previously introduced in Section 3.2 where it was observed that the contribution of the volatility of the forward prices (and hence the zero coupon bonds) is essentially negligible. Assumption A2 is required in order to interrelate the Libor rates, and is, in fact, equivalent to A3, which is only included as a separate assumption for reasons of clarity. A3 assumes that we can approximate our (in general multi-factor) volatility function $\gamma(t, T)$ by a single-factor separable model, as in

$$\gamma_{approx}(t, T) = \psi(t) \phi(T). \quad (24)$$

While this assumption seems quite restrictive, we note (see Appendix B) that it is entirely equivalent to Assumption A2, in that the volatility structure is separable if and only if the swaption covariance matrix is of rank one. Numerical results suggest that for most (non-extreme) volatility structures, the swaption covariance matrix is very close to rank one, validating both assumptions A2 and A3. This is considered in more detail in Section 5.3. The approximation (24) is constructed in such a way that it returns the rank one swaption covariance matrix

$$\begin{aligned} (\lambda_{i,j}) &= \left(\int_t^{T_0} \gamma(s, T_{i-1}) \cdot \gamma(s, T_{j-1}) ds \right) \\ &= \left(\phi(T_{i-1}) \phi(T_{j-1}) \int_t^{T_0} \psi^2(s) ds \right) = \Gamma \times \Gamma^T, \end{aligned}$$

implying

$$\Gamma_j = \phi(T_{j-1}) \sqrt{\int_t^{T_0} \psi^2(s) ds}. \quad (25)$$

Approximation A4 is used in simplifying the relationship between the Libor rates and in the computation of the swaption gamma and vega. Essentially it is analogous to the implicit assumption in the Black swaption formula (mentioned in Section 2.2) that the swap rates are assumed lognormal under a single measure $\mathbb{P}_m$.

We assume that calendar time $t = 0$ and introduce the abbreviated notation $K_j \sim K(0, T_j)$, $P_j \sim P(0, T_j)$, and $\phi_j \sim \phi(T_j)$, and the variable U satisfying

$$dU = \psi(t) dW(t),$$

where $W(t)$ is Brownian motion under the single measure into which all the forward measures have been merged. Applying assumptions A1, A3 and A4 to Equations (1) and (4), we have the following simplified equations for the Libor and swap rate processes

$$\begin{aligned} dK(t, T_{j-1}) &= K(t, T_{j-1}) \psi(t) \phi_{j-1} dW_{T_j}(t) \\ &= K(t, T_{j-1}) \phi_{j-1} dU, \end{aligned} \quad (26)$$

and

$$d\omega = \frac{\sum_j P_j K_{j-1} \phi_{j-1}}{\sum_j P_j} dU. \quad (27)$$

With these assumptions/approximations, we can now proceed to derive equations for the swaption Greeks in the Libor model.

4.2 Libor model delta

In the case of single-factor volatility functions, a swaption delta can be derived with minimal approximation by eliminating stochastic terms in the stochastic differential equations for the swap and swaption. Here we consider a different method involving differentiation inside the expectation term, a method which will be further utilised in Section 4.3 to derive an expression for the swaption gamma. Note however that both methods would produce an equivalent expression for the swaption delta.

Define Δ_{i-1} to be the partial derivative of the swaption price with respect to the Libor rate $K(0, T_{i-1})$. Denoting the swaption price $P_{swpn}(0)$ as S , we have, using (5),

$$\begin{aligned} \Delta_{i-1} &= \frac{\partial S}{\partial K_{i-1}} = \frac{\partial}{\partial K_{i-1}} \left(\delta \sum_j P_j \mathbb{E}_{T_j} (K(T, T_{j-1}) - \kappa) \mathbb{I}(A) \right) \\ &= \delta \sum_j P_j \mathbb{E}_{T_j} \left\{ \frac{\partial K(T, T_{j-1})}{\partial K_{i-1}} \mathbb{I}(A) + (K(T, T_{j-1}) - \kappa) \frac{\partial \mathbb{I}(A)}{\partial K_{i-1}} \right\}. \end{aligned}$$

By measure transformation, the second term inside the expectation can be shown to equate to

$$P(0, T) \mathbf{E}_T \left\{ \text{Swap}(T) \frac{\partial \mathbb{I}(A)}{\partial K_{i-1}} \right\} = 0$$

since

$$\frac{\partial \mathbb{I}(A)}{\partial K(0, T_{i-1})} = 0 \quad \text{if} \quad \text{Swap}(T) \neq 0.$$

Using the integrated version of Equation (1), we can then show that the remaining expression reduces to

$$\Delta_{i-1} = \delta P_i \mathbf{N}(h_i) \quad (28)$$

where the h_i are given by (9).

Treating U as a real variable, we now obtain an expression for the swaption delta in the Libor model using the definition (13) from Section 2.2,

$$\Delta = \frac{\partial}{\partial \omega} \left(\frac{S}{\delta \sum_j P_j} \right) = \frac{1}{\delta \sum_j P_j} \frac{\partial S}{\partial \omega} \quad (29)$$

$$\begin{aligned} &= \frac{1}{\delta \sum_j P_j} \sum_j \frac{\partial S}{\partial K_{j-1}} \frac{\partial K_{j-1}}{\partial U} \frac{\partial U}{\partial \omega} \\ &= \frac{1}{\delta \sum_j P_j} \sum_j \frac{\Delta_{j-1} K_{j-1} \phi_{j-1} \sum_i P_i}{\sum_i P_i K_{i-1} \phi_{i-1}} \\ &= \frac{\sum_j P_j \mathbf{N}(h_j) K_{j-1} \phi_{j-1}}{\sum_j P_j K_{j-1} \phi_{j-1}}. \end{aligned} \quad (30)$$

Equation (30) is tested against the Black swaption Δ in Section 5.6, and in terms of swaption hedging in Section 5.8.

4.3 Libor model gamma

Building on the approach of Section 4.2, we can now derive an expression for the swaption gamma in the Libor model. The first step is to calculate second derivatives of the Libor model swaption with respect to the $K(\cdot)$ – which we will denote as $\Gamma_{i,k}$ – and then, using the assumptions of Section 4.1, obtain a single number that can be compared to the gamma given by the Black formula. We have⁴

$$\begin{aligned} \Gamma_{i-1,k-1} &= \frac{\partial^2 P_{\text{swpn}}(0)}{\partial K_{i-1} \partial K_{k-1}} \\ &= \delta P_i \mathbb{E}_{T_i} \left\{ \frac{\partial K(T, T_{i-1})}{\partial K_{i-1}} \frac{\partial \mathbb{I}(\text{Swap}(T))}{\partial K_{k-1}} \right\} \end{aligned}$$

⁴ Use the formulae $\frac{d(x)^+}{dx} = \mathbb{I}(x)$, $\frac{d\mathbb{I}(x)}{dx} = \delta\{x\}$, where $\mathbb{I}(\cdot)$ is the Heaviside function and $\delta\{\cdot\}$ is the Dirac delta function.

$$\begin{aligned}
&= \delta^2 P_i \mathbb{E}_{T_i} \left\{ \frac{\partial K(T, T_{i-1})}{\partial K_{i-1}} P(T, T_k) \frac{\partial K(T, T_{k-1})}{\partial K_{k-1}} \right. \\
&\quad \left. \times \delta \left\{ \delta \sum_{j=1}^n P(T, T_j) \{K(T, T_{j-1}) - \kappa\} \right\} \right\}.
\end{aligned}$$

With assumption A4, and setting $Z \sim \mathbf{N}(0, 1)$, it follows that,⁵

$$\begin{aligned}
\Gamma_{i-1, k-1} &\simeq \delta^2 P_i P_k \mathbb{E} \left\{ \mathbf{e}(\Gamma_i Z) \mathbf{e}(\Gamma_k Z) \delta \left\{ \delta \sum_j P(T, T_j) \{K_{j-1} \mathbf{e}(\Gamma_j Z) - \kappa\} \right\} \right\} \\
&= \delta^2 P_i P_k \exp(\Gamma_i \Gamma_k) \\
&\quad \times \mathbb{E} \left\{ \delta \left\{ \delta \sum_j P(T, T_j) (K_{j-1} \mathbf{e}(\Gamma_j [Z + \Gamma_i + \Gamma_k]) - \kappa) \right\} \right\}.
\end{aligned}$$

Assuming that the ‘s’ satisfying (7) also approximately satisfies

$$\sum_j P(T, T_j) (K_{j-1} \exp(\Gamma_j s - \frac{1}{2} \Gamma_j^2) - \kappa) = 0, \quad (31)$$

then we have

$$\begin{aligned}
\Gamma_{i-1, k-1} &\simeq \frac{\delta P_i P_k \exp(\Gamma_i \Gamma_k) \mathbf{N}'(s - \Gamma_i - \Gamma_k)}{\sum_j P_j K_{j-1} \Gamma_j \exp(\Gamma_j s - \frac{1}{2} \Gamma_j^2)} \\
&= \frac{\delta P_i P_k \mathbf{N}'(s - \Gamma_i) \mathbf{N}'(s - \Gamma_k)}{\sum_j P_j K_{j-1} \Gamma_j \mathbf{N}'(s - \Gamma_j)}. \quad (32)
\end{aligned}$$

Using our definition for the swaption gamma (14), we can derive an expression in terms of the partial derivatives derived above, giving

$$\begin{aligned}
\Gamma &= \frac{\partial^2}{\partial \omega^2} \left(\frac{S}{\delta \sum_j P_j} \right) = \frac{1}{\delta \sum_j P_j} \frac{\partial}{\partial \omega} \left(\frac{\partial S}{\partial \omega} \right) \\
&= \frac{1}{\delta \sum_j P_j} \sum_j \frac{\partial}{\partial K_{j-1}} \left(\frac{\partial S}{\partial \omega} \right) \frac{\partial K_{j-1}}{\partial U} \frac{\partial U}{\partial \omega}. \quad (33)
\end{aligned}$$

Recall from Section 4.2 that we have

$$\begin{aligned}
\frac{\partial K_{j-1}}{\partial U} \frac{\partial U}{\partial \omega} &= \frac{\sum_i P_i}{\sum_i P_i K_{i-1} \Gamma_i} K_{j-1} \Gamma_j \\
\frac{\partial S}{\partial \omega} &= \sum_j \frac{\partial S}{\partial K_{j-1}} \frac{\partial K_{j-1}}{\partial U} \frac{\partial U}{\partial \omega} \\
&= \frac{\sum_i P_i}{\sum_i P_i K_{i-1} \Gamma_i} \sum_j \Delta_{j-1} K_{j-1} \Gamma_j,
\end{aligned}$$

⁵ If X is a random variable under some given measure, then $\mathbf{e}(X) = \exp\left(X - \frac{1}{2} \text{Var } X\right)$.

and substituting these into (33) and taking the partial derivative gives us

$$\begin{aligned} \Gamma = & \frac{\sum_j P_j}{\delta \left(\sum_j P_j K_{j-1} \Gamma_j \right)^2} \sum_i \sum_j \Gamma_i \Gamma_j K_{i-1} K_{j-1} \Gamma_{i-1, j-1} \\ & + \frac{\sum_j P_j}{\delta \left(\sum_j P_j K_{j-1} \Gamma_j \right)^2} \left[\left(\sum_j P_j K_{j-1} \Gamma_j \right) \left(\sum_j \Delta_{j-1} K_{j-1}^2 \Gamma_j^2 \right) \right. \\ & \left. - \left(\sum_j \Delta_{j-1} K_{j-1} \Gamma_j \right) \left(\sum_j P_j K_{j-1} \Gamma_j^2 \right) \right] \end{aligned}$$

in which the second term can be shown to be the difference of two quantities of similar order of magnitude and is hence taken to be zero. Substitution of (32) and collecting terms gives us our final expression for the Libor model swaption gamma

$$\Gamma = \sum_j P_j \frac{\sum_j P_j K_{j-1} \Gamma_j \mathbf{N}'(s - \Gamma_j)}{\left(\sum_j P_j K_{j-1} \Gamma_j \right)^2}. \quad (34)$$

4.4 Libor model vega

Finally, we wish to derive an equation for the swaption vega in the Libor model. Combining the approximate swap rate volatility equation (23) with Assumption A3 of an instantaneous one-factor separable volatility (24), we obtain

$$\gamma(t, T_0, n) = \psi(t) \frac{\sum_j P_j K_{j-1} \phi_{j-1}}{\sum_j P_j K_{j-1}}.$$

The swaption *zeta* in the Libor model corresponding to (12) is

$$\begin{aligned} \zeta &= \int_0^{T_0} |\gamma(s, T_0, n)|^2 ds \\ &= \left(\int_0^{T_0} \psi^2(s) ds \right) \left(\frac{\sum_j P_j K_{j-1} \phi_{j-1}}{\sum_j P_j K_{j-1}} \right)^2, \end{aligned}$$

and following the methodology presented in Section 2.2 we want to partially differentiate with respect to this variable to obtain the vega. To do this, we will denote by V the integral $\int_0^{T_0} \psi^2(s) ds$ and assume that this constitutes the variable part of ζ , implying

$$\frac{\partial \zeta}{\partial V} = \left(\frac{\sum_j P_j K_{j-1} \phi_{j-1}}{\sum_j P_j K_{j-1}} \right)^2. \quad (35)$$

From the definition of the vega (15), we have

$$\begin{aligned}\Lambda &= \frac{\partial}{\partial \zeta} \left(\frac{P_{\text{swpn}}(0)}{\delta \sum_j P_j} \right) = \frac{1}{\delta \sum_j P_j} \frac{\partial S}{\partial \zeta} \\ &= \frac{1}{\delta \sum_j P_j} \frac{\partial S}{\partial V} \frac{\partial V}{\partial \zeta},\end{aligned}$$

where, in this case, we can obtain the partial derivative $\partial S/\partial V$ by direct differentiation of the swaption formula (8). Using the additional assumption (implicit in the use of (31)) that $d_j \approx 0$, gives us

$$\begin{aligned}\frac{\partial S}{\partial V} &= \delta \sum_j P_j \left(K_{j-1} \frac{\partial h_j}{\partial V} \mathbf{N}'(h_j) - \kappa \frac{\partial(h_j - \Gamma_j)}{\partial V} \mathbf{N}'(h_j - \Gamma_j) \right) \\ &= \delta \sum_j P_j \left(K_{j-1} \left(-\frac{\partial s}{\partial V} + \frac{\partial \Gamma_j}{\partial V} \right) \mathbf{N}'(-s + \Gamma_j) + \kappa \frac{\partial s}{\partial V} \mathbf{N}'(s) \right) \\ &= \delta \left(-\frac{\partial s}{\partial V} \right) \mathbf{N}'(s) \sum_j P_j (K_{j-1} \exp(s\Gamma_j - \frac{1}{2}\Gamma_j^2) - \kappa) \\ &\quad + \delta \sum_j P_j K_{j-1} \frac{\partial \Gamma_j}{\partial V} \mathbf{N}'(s - \Gamma_j),\end{aligned}$$

where the first term can be seen to satisfy (31) and so can be taken as zero. Partial differentiation of (25) yields

$$\frac{\partial \Gamma_j}{\partial V} = \frac{\phi_{j-1}}{2\sqrt{V}} = \frac{\Gamma_j}{2V}$$

and hence

$$\frac{\partial S}{\partial V} = \frac{\delta}{2V} \sum_j P_j K_{j-1} \Gamma_j \mathbf{N}'(s - \Gamma_j).$$

Substituting from above as necessary, the vega is therefore

$$\begin{aligned}\Lambda &= \frac{1}{\delta \sum_j P_j} \frac{\partial S}{\partial V} \frac{\partial V}{\partial \zeta} \\ &= \frac{1}{2V \sum_j P_j} \left(\frac{\sum_j P_j K_{j-1}}{\sum_j P_j K_{j-1} \phi_{j-1}} \right)^2 \sum_j P_j K_{j-1} \Gamma_j \mathbf{N}'(s - \Gamma_j) \\ &= \frac{1}{2 \sum_j P_j} \left(\frac{\sum_j P_j K_{j-1}}{\sum_j P_j K_{j-1} \Gamma_j} \right)^2 \sum_j P_j K_{j-1} \Gamma_j \mathbf{N}'(s - \Gamma_j).\end{aligned}\tag{36}$$

Noting from (4) that $\omega = \sum_j P_j K_{j-1} / \sum_j P_j$, we see that the gamma and vega equations (34) and (36) satisfy the constraint (16) imposed on them in Section 2.2,

$$\Lambda = \frac{1}{2} \left(\frac{\sum_j P_j K_{j-1}}{\sum_j P_j} \right)^2 \Gamma = \frac{1}{2} \omega^2 \Gamma.$$

5 Numerical testing and results

Ultimately, the closeness of swaption pricing within the Libor model to the Black swaption formula must be tested numerically. In this section, the assumptions fundamental to the analysis are verified, the regime used to test the equations is explained, and the results of the numerical testing presented.

In order to test the approximate equations for volatility, pricing and Greeks thoroughly, a range of swaptions, strike values, yield curves and volatility specifications is required. In this light, it was decided to test a matrix consisting of 15 swaptions with maturity values ranging from 0.5 to 4 years, lengths of 1 to 8 years, and at strike values in-, at- and out-of-the-money. The tests were conducted for two separate volatility specifications – the first a single-factor homogeneous parameterisation to actual historic data, chosen to reflect typical market conditions – and the second, an artificial two-factor volatility function chosen to mimic a pathological market situation and *stress test* the results. Further details on the volatility specifications and their associated yield curves are given in Appendix C.

With the Black pricing formula, the price and Greeks can all be computed upon specification of the Black volatility σ . This is not the case in the Libor model, where an equivalent Black volatility can be obtained only by first computing the price and then ‘backing out’ the volatility by solving Equation (10) for a constant valued volatility function σ . Given that any comparison between prices and Greeks would be meaningless if not computed at a Black volatility equivalent to both frameworks, we define the Libor model *true price* as that value obtained from simulation, and the *true volatility* as the value obtained by backing out the *true price* at-the-money. The necessity of this distinction becomes apparent when one notes that Libor swaption pricing formula (8) only gives an approximate price, and one that can deviate from the true value under certain circumstances. The simulated price, however, is a reflection of the exact price, and, exploiting variance reduction means, can be made as accurate as required. This provides us with a number, free of approximation, which can be used objectively for comparison purposes.

We start, however, by verifying the assumptions used in deriving the various approximations.

Normal Probability Plot for the 1/8 Swap Rate

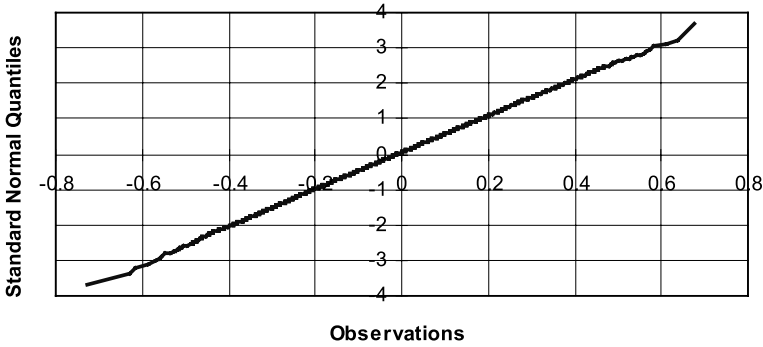


Fig. 1. Normal probability plot of the log of the swap rates simulated under the Libor model for a 1/8 swap using the second volatility structure.

5.1 Lognormality of the swap rates

In Section 3.1 it was postulated that the swap rate ω could be modelled as being approximately lognormal under the $\mathbb{P}_{T_0}$ forward measure. This was tested numerically by simulating swap rates under the appropriate measure within the Libor model framework. The simulation was performed by discretising the stochastic differential equations for the Libor rates (1) to produce sets of future yield curves from which the swap rates could be extracted.⁶ Statistical tests were then applied to the swap rates to determine the nature of the resulting distributions.

Figure 1 is an example of one of those statistical tests; a normal probability plot of the log of the simulated swap rates, in this case for an eight year swap, maturing in one year, simulated using the pathological volatility structure. A normal probability plot allows one to determine if random observations come from a normally distributed population; a straight line indicating the affirmative. Slight deviations at either end of the line are common, as a finite number of samples will never be able to fit the infinite tails of the normal distribution exactly. The test can be formalised through the use of quantitative statistical tests (such as the Shapiro–Wilk test), or a goodness-of-fit test between the expected and observed sample frequencies. The latter was used in this case.

All the swaptions for *both* volatility structures gave similar results to those in Figure 1, and at a 95% confidence level, were shown to follow a lognormal probability distribution.

⁶ See Brace (1998) for details of the simulation routine used, and Glasserman *et al.* (2000) for detailed analysis of a range of simulation methods in the forward Libor model.

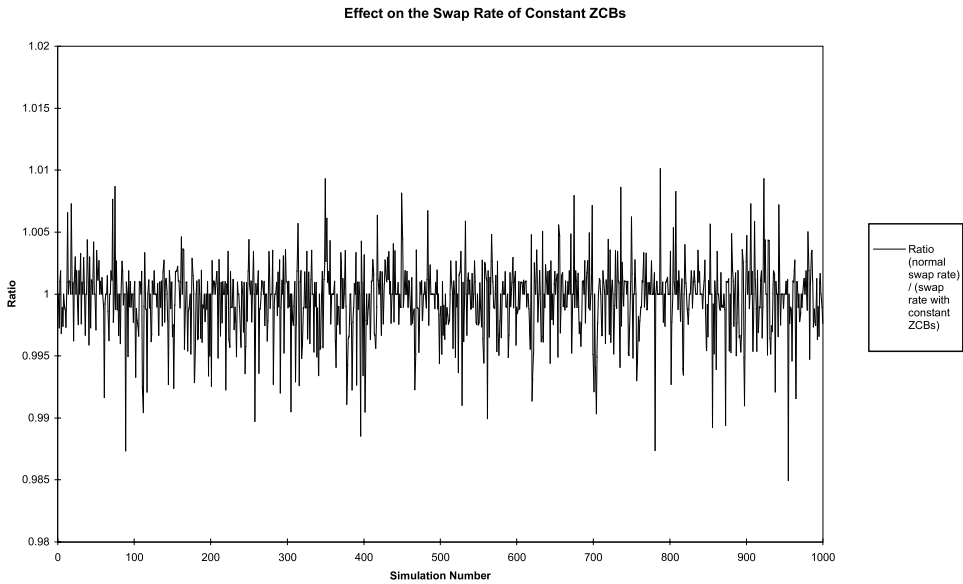


Fig. 2. The ratio between simulated swap rates with and without the effect of the zero coupon bonds.

5.2 Swap rate approximation

The approximations in Sections 3–4 rely on the assumption that the contribution of the volatility of the discount terms (forward prices and zero coupon bonds) towards the overall volatility of the swap rate is negligible, and that the discount terms can be considered constant at their initial values.

Figure 2 confirms the validity of this assumption on the swap rate for a 1/5 swap, simulated using the second volatility structure. It shows the ratio of the simulated swap rate calculated using all the discount terms, to the value obtained by taking these terms as constant. A value of 1 indicates that the calculation methods are equivalent. This figure demonstrates that the assumption is quite reasonable, leading to errors in the swap rate that are generally below one per cent.

5.3 Rank one covariance matrix

The Libor model swaption formula (8) and all the analysis in Section 4 are fundamentally dependent on the assumption that the swaption covariance matrix λ is of rank one. A symmetric matrix is of rank one when it has only one non-zero eigenvalue. A rank one approximation to an arbitrary symmetric matrix will only be accurate if the ratio of the second largest to the largest eigenvalue is small.

Table 1. *Ratio of the first and second eigenvalues for the swaption covariance matrices (both volatility structures).*

Volatility structure	Swaption length	Swaption maturity			
		0.25	1	2	4
1	1	0.0%	7.5%	1.5%	2.1%
	2	0.0%	1.5%	3.2%	3.5%
	4	0.0%	2.1%	3.5%	5.9%
	8	0.0%	1.6%	2.7%	
2	1	0.5%	1.0%	1.6%	1.6%
	2	4.4%	6.7%	8.2%	4.8%
	4	30.8%	27.9%	17.3%	6.4%
	8	20.5%	13.0%	7.9%	

In the case of the Libor model, the rank of the swaption covariance matrix will depend on the form of the volatility function $\gamma(t, T)$, and the maturity and length of the individual swaption. A swaption is said to be exhibiting *rank two behaviour* when the rank one price (8) begins to deviate from the *true* price. This seems to occur for an eigenvalue ratio of 5% or above, with 20–30% representing extreme values.

Table 1 shows this ratio for all the swaptions and volatility structures considered in this paper. A value of 0 represents a swaption covariance matrix of rank one. The second volatility structure was chosen for its *pathological* nature, and this is reflected in the more extreme values for the eigenvalue ratio seen here. It would not be surprising, therefore, if the approximations of Section 4 were to break down for some of the swaptions under the second volatility structure.

5.4 Swap rate volatility

In Section 3.2, we derived the approximate equation (23) for the equivalent Black volatility of a Libor model swaption. In Table 2 we compare values given by this equation to the *true* volatility, defined in Section 5 as the volatility implied by the at-the-money simulation price of the corresponding swaption within the Libor model framework.

The results indicate that the volatility approximation is quite accurate, with all the values for rank one swaptions within about 12 basis points, with this figure rising to 80 basis points for the more extreme rank two swaptions occurring under the second volatility structure. In general, however, the approximate volatility equation (23) provides a good indication for the Libor model *true* volatility.

Table 2. *Black volatility verification results for both volatility structures.*

Volatility structure	Swaption length	Volatility description	Swaption maturity			
			0.25	1	2	4
1	1	true	4.64%	5.73%	10.14%	17.59%
		approximation	4.65%	5.74%	10.15%	17.61%
	2	true	6.97%	9.37%	14.23%	18.58%
		approximation	6.98%	9.38%	14.24%	18.58%
	4	true	14.02%	15.53%	17.56%	18.51%
		approximation	14.07%	15.57%	17.59%	18.56%
	8	true	15.32%	15.80%	16.57%	
		approximation	15.44%	15.90%	16.65%	
2	1	true	23.16%	19.81%	17.46%	17.76%
		approximation	23.20%	19.85%	17.50%	17.75%
	2	true	18.60%	16.64%	16.26%	18.06%
		approximation	18.72%	16.74%	16.17%	18.04%
	4	true	15.79%	15.81%	16.67%	20.24%
		approximation	15.85%	15.68%	16.41%	20.13%
	8	true	18.37%	19.05%	20.34%	
		approximation	17.88%	18.35%	19.54%	

5.5 *Swaption prices*

Table 3 compares swaption prices for the first volatility structure. Three different prices are given – the *true* value obtained by simulation, an approximate value obtained by using the Black swaption formula (10) with the swap rate volatility approximation (23), and the Libor model rank one price (8). The prices are expressed in basis points (bp), where 1 bp = \$100 per \$1M face value. As with the previous swaption volatilities, for the rank one swaptions, the volatility approximation provides a reasonable estimate of the swaption price. As to be expected, the Libor model price performs better in most situations. The deviation between the *true* and rank one prices is evident in the rank two swaptions under the second volatility structure (shown in Appendix A), and it is not surprising to note that under these circumstances the volatility approximation mirrors the rank one price more than the *true* price.

In general, however, these results show that a Libor model swaption behaves very much like a Black swaption with the volatility given by Equation (23).

Table 3. *Swaption price comparisons for the first volatility structure (all values expressed in basis points).*

Swaption length	Strike	Price description	Swaption maturity			
			0.25	1	2	4
1	IN	true	12.52	30.34	68.87	126.86
		vol approx	12.53	30.35	68.96	126.93
		rank 1	12.52	30.32	68.85	126.91
	AT	true	6.18	15.37	37.22	78.97
		vol approx	6.18	15.37	37.25	79.04
		rank 1	6.18	15.35	37.20	79.02
	OUT	true	2.29	5.59	13.06	25.16
		vol approx	2.29	5.59	13.00	25.18
		rank 1	2.29	5.58	13.05	25.17
2	IN	true	37.29	94.79	178.16	254.77
		vol approx	37.34	95.08	178.61	254.79
		rank 1	37.29	94.77	178.07	254.72
	AT	true	18.56	49.42	100.45	160.62
		vol approx	18.59	49.51	100.54	160.65
		rank 1	18.56	49.40	100.36	160.56
	OUT	true	6.83	17.86	34.55	50.74
		vol approx	6.83	17.68	34.17	50.77
		rank 1	6.83	17.85	34.54	50.69
4	IN	true	140.40	282.57	397.87	475.04
		vol approx	140.93	283.66	398.71	475.78
		rank 1	140.38	282.38	397.51	475.20
	AT	true	71.82	154.19	231.58	299.12
		vol approx	72.09	154.57	231.90	299.90
		rank 1	71.81	154.02	231.16	299.26
	OUT	true	26.16	54.22	77.45	94.53
		vol approx	26.04	53.63	77.18	94.79
		rank 1	26.18	54.24	77.44	94.35
8	IN	true	272.66	507.00	666.23	
		vol approx	273.80	509.09	668.80	
		rank 1	272.47	506.52	665.36	
	AT	true	139.66	276.30	383.50	
		vol approx	140.79	278.07	385.49	
		rank 1	139.57	276.10	382.81	
	OUT	true	50.09	95.81	128.49	
		vol approx	50.68	96.33	129.04	
		rank 1	50.12	96.03	128.74	

Table 4. *Delta comparisons for Libor and Black swaptions for the first volatility structure.*

Swaption length	Strike	Model	Swaption maturity			
			0.25	1	2	4
1	IN	Black	0.750	0.750	0.750	0.750
		Libor	0.751	0.751	0.752	0.750
	AT	Black	0.505	0.511	0.529	0.570
		Libor	0.506	0.512	0.531	0.570
	OUT	Black	0.250	0.250	0.250	0.250
		Libor	0.251	0.250	0.252	0.250
2	IN	Black	0.750	0.750	0.750	0.750
		Libor	0.752	0.755	0.755	0.750
	AT	Black	0.507	0.519	0.540	0.574
		Libor	0.508	0.523	0.545	0.574
	OUT	Black	0.250	0.250	0.250	0.250
		Libor	0.251	0.255	0.255	0.250
4	IN	Black	0.751	0.750	0.750	0.750
		Libor	0.756	0.757	0.755	0.751
	AT	Black	0.514	0.531	0.549	0.573
		Libor	0.519	0.538	0.554	0.574
	OUT	Black	0.249	0.249	0.250	0.249
		Libor	0.255	0.257	0.254	0.250
8	IN	Black	0.752	0.751	0.751	
		Libor	0.755	0.756	0.754	
	AT	Black	0.515	0.531	0.547	
		Libor	0.518	0.536	0.550	
	OUT	Black	0.248	0.248	0.248	
		Libor	0.251	0.253	0.252	

5.6 Swaption delta

The validity of the approximate swaption delta equation is illustrated in Table 4 which compares Δ values for a range of equivalently priced Black and Libor model swaptions at-, in- and out-of-the-money for the first volatility structure. The Black swaption delta is calculated using the *true* swap rate volatility (see Section 5.4), with the strike values chosen so that the Δ values in- and out-of-the-money are approximately 0.75 and 0.25, respectively.

The results show that the approximate method gives good agreement to the Black swaption Δ – showing slight, yet consistent, over-estimation of the true values.

Even for the more extreme swaptions under the second volatility structure (see Appendix A), the agreement is quite acceptable, with the values deviating by 4.5% at most, with the average deviation being 0.1%. Note, however, that this deviation, for both volatility structures, tends to increase slightly as the swaptions move out-of-the-money.

5.7 Swaption gamma and vega

Libor model gamma and vega equations (34) and (36) were tested against their Black counterparts (14) and (15), respectively, with the results shown in Table 5. As in Section 5.6, the Black swaption Greeks are calculated using the *true* volatility, and the same in- and out-of-the-money strike prices are used. Note that the Λ results will be entirely analogous to the Γ results, as Λ is directly proportional to Γ , as given by (16).

We see in general for both Γ and Λ that the agreement between the swaption behaviours is not as good as for the Δ , yet is still quite acceptable, with most of the Libor model results within 5% of the Black values. Note that the Libor model equations tend to underestimate the values in-the-money, while overestimating out-of-the-money. Note also that the agreement between the values deteriorates with longer swaption maturity and length. This is also true for the second volatility structure, shown in Appendix A.

5.8 Swaption delta-hedging

The Libor model Δ equation (30) gives an approximation to the partial derivatives of the swaption price with respect to the swap rate. However, as explained in Section 2.3, in the Black–Scholes framework (or here, in the framework implied by the Black swaption formula) the Δ is more than just a partial derivative – it represents a probability of exercise of the option – and is fundamental to the concept of hedging. It would be interesting to know if this concept can also be extended to the case of the approximate Libor model delta.

To test this, yield curve movements were simulated in the Libor model framework and swaptions hedged using the methodology from Section 2.3 and the approximate Δ formula (30). Rebalancing was effected at a frequency of five times per quarter, and, due to the lack of *true* (or simulation) prices and volatilities, the hedging was based on values given by the rank one Libor model price formula (8). For comparison purposes, the delta-hedge was run in conjunction with a Libor model hedge encompassing all the relevant Libor rates treated individually – as predicted from the partial derivatives with respect to the Libor rates given by (28).

Table 5. *Gamma and vega comparisons for Libor model and Black swaptions (for the first volatility structure).*

Greek type	Swaption length	Strike	Model	Swaption maturity			
				0.25	1	2	4
Gamma	1	IN	Black	193.5	73.7	28.1	11.3
			Libor	192.8	73.4	27.9	11.1
		AT	Black	243.0	92.5	35.3	14.0
			Libor	242.8	92.5	35.2	13.9
		OUT	Black	193.5	73.7	28.1	11.3
			Libor	194.1	73.9	28.4	11.4
	2	IN	Black	124.3	44.1	20.1	10.7
			Libor	123.4	43.3	19.6	10.4
		AT	Black	156.1	55.3	25.1	13.2
			Libor	155.8	55.1	24.9	13.1
		OUT	Black	124.2	44.1	20.1	10.7
			Libor	124.7	44.7	20.5	10.9
	4	IN	Black	59.6	26.2	16.1	10.6
			Libor	58.2	25.3	15.5	10.1
		AT	Black	74.9	32.8	20.1	13.1
			Libor	74.5	32.6	19.9	12.9
		OUT	Black	59.6	26.2	16.1	10.6
			Libor	60.4	27.0	16.7	11.0
	8	IN	Black	52.9	25.2	16.8	
			Libor	51.3	24.0	15.7	
		AT	Black	66.6	31.6	21.0	
			Libor	65.9	31.2	20.6	
		OUT	Black	52.9	25.2	16.8	
			Libor	53.6	26.1	17.5	
Vega	1	IN	Black	0.484	0.208	0.087	0.036
			Libor	0.482	0.208	0.086	0.036
		AT	Black	0.607	0.262	0.109	0.045
			Libor	0.607	0.262	0.109	0.044
		OUT	Black	0.484	0.208	0.087	0.036
			Libor	0.485	0.209	0.088	0.037
	2	IN	Black	0.334	0.130	0.062	0.034
			Libor	0.332	0.128	0.061	0.033
		AT	Black	0.420	0.164	0.078	0.042
			Libor	0.419	0.163	0.077	0.042
		OUT	Black	0.334	0.130	0.062	0.034
			Libor	0.336	0.132	0.063	0.035

Table 5. (*cont.*)

Greek type	Swaption length	Strike	Model	Swaption maturity			
				0.25	1	2	4
Vega	4	IN	Black	0.172	0.080	0.051	0.035
			Libor	0.168	0.077	0.049	0.033
		AT	Black	0.216	0.100	0.063	0.043
			Libor	0.215	0.099	0.063	0.042
		OUT	Black	0.172	0.080	0.051	0.035
			Libor	0.174	0.082	0.052	0.036
	8	IN	Black	0.162	0.080	0.055	
			Libor	0.156	0.076	0.051	
		AT	Black	0.203	0.100	0.068	
			Libor	0.201	0.099	0.067	
		OUT	Black	0.161	0.080	0.055	
			Libor	0.164	0.082	0.057	

A more detailed explanation of the mathematics and methodology of the hedging simulation is beyond the scope of this chapter and can be found in Dun *et al.* (1999).

Table 6 presents the results of these hedging tests in the form of means and standard deviations of the hedging profit and loss (P/L) for both volatility structures. A zero mean P/L with a small standard deviation is clearly the preferred outcome in any hedging exercise.

The results show that the approximate Libor Δ performs equally as well as individual hedges into the Libor rates – both in terms of P/L mean and standard deviation. All the rank one swaptions have been successfully hedged, with average P/Ls close to zero, while the rank two swaptions show some bias. This bias seems to be approximately equal to the difference between the *true* and rank one prices, and could probably be reduced by using the *true* volatility as the basis for the hedges rather than a rank one volatility as mentioned above. In general, however, the results imply that the approximate Libor model Δ is useful for hedging, and that the intuition attached to the delta value in Black swaptions is also valid in the Libor model framework.

6 Conclusions

In conclusion, we have derived approximate equations within the lognormal forward Libor model which indicate that swaption pricing in this framework is quite close to market practice. A simple equation can be used to estimate the Black volatility of Libor model swaptions, which can then be priced using the Black

Table 6. *Simulated delta hedging means (and standard deviations) for both volatility structures. Values expressed in basis points.*

Volatility structure	Swaption length	Hedging method	Swaption maturity							
			0.25		1		2		4	
1	1	Approx delta	0.0	(2.3)	0.0	(3.0)	0.0	(6.1)	0.1	(8.4)
		Libor rates	0.0	(2.3)	0.0	(3.0)	0.0	(6.1)	0.1	(8.4)
	2	Approx delta	0.0	(6.7)	0.1	(9.6)	0.0	(14.5)	-0.1	(16.2)
		Libor rates	0.0	(6.7)	0.1	(9.6)	0.0	(14.5)	-0.1	(16.2)
	4	Approx delta	0.0	(26.7)	0.3	(28.8)	-0.3	(30.8)	0.0	(28.4)
		Libor rates	0.0	(26.7)	0.3	(28.8)	-0.3	(30.8)	0.0	(28.3)
	8	Approx delta	0.4	(50.2)	-0.6	(52.6)	-0.8	(51.3)		
		Libor rates	0.4	(50.2)	-0.6	(52.5)	-0.8	(51.2)		
2	1	Approx delta	0.0	(13.7)	0.0	(14.4)	0.0	(14.4)	0.0	(8.0)
		Libor rates	0.0	(13.7)	0.0	(14.4)	0.0	(14.4)	0.0	(8.0)
	2	Approx delta	0.0	(23.5)	0.0	(23.9)	0.2	(21.1)	-0.2	(14.4)
		Libor rates	0.0	(23.4)	0.0	(23.9)	0.2	(21.0)	-0.2	(14.4)
	4	Approx delta	0.1	(36.4)	-1.4	(36.2)	-4.6	(33.5)	-0.7	(26.8)
		Libor rates	0.1	(36.4)	-1.4	(36.1)	-4.6	(33.4)	-0.7	(26.8)
	8	Approx delta	-9.8	(64.7)	-15.9	(65.1)	-14.0	(60.6)		
		Libor rates	-9.9	(64.7)	-15.9	(65.3)	-14.0	(60.6)		

swaption formula. Equations for swaption Greeks in the Libor model were derived and shown to retain their Black swaption significance, while Libor model swaptions could be successfully hedged with the swaption delta derived. Estimates are accurate while the assumption of a rank one swaption covariance matrix holds, although even when violated, the estimates are still surprisingly close to the true values. Swaption maturity, length and strike value do exhibit a slight influence on the estimates.

Overall, the results support the idea that the Libor model could be used for all swaption pricing – as well as caps and exotics pricing – since it can be calibrated to both caps and swaptions markets simultaneously. Conversely, the results could be used to support the idea in Jamshidian (1997) that models which are robust and adapted to the products being priced should be used – even if this means using mutually exclusive models – since we have shown that the Libor and Black (and hence by extension the swap rate) approaches are, numerically, not so different.

This study still leaves some questions unanswered, providing scope for further work. This includes, for example, the derivation of analytic bounds for the approximations presented here, an analysis of the closeness of the models when pricing exotics, and an investigation into the impact of using the assumptions of Section 4.1 to simplify exotics pricing.

Appendix A. Results for the second volatility structure

Comparisons of prices, deltas, gammas and vegas for the second volatility structure not tabulated in the body of the paper appear in Tables 7–9.

Appendix B. Rank one and separable volatility

If the volatility function is separable, all swaption quadratic variation matrices are of rank one. On the other hand, if a swaption quadratic variation matrix is of rank one, for arbitrary T and $T_i = T + i\delta$, we must have

$$\int_0^t \gamma^2(s, T) ds \int_0^t \gamma^2(s, T_i) ds = \left(\int_0^t \gamma(s, T) \gamma(s, T_i) ds \right)^2.$$

The following lemma shows that if this condition is strengthened, separability follows.

Lemma 1 *Let the LFM volatility function $\gamma(\cdot)$ be well behaved, and satisfy*

$$\int_0^t \gamma^2(s, u) ds \int_0^t \gamma^2(s, v) ds = \left(\int_0^t \gamma(s, u) \gamma(s, v) ds \right)^2 \quad (37)$$

for all relevant t, u, v . Then $\gamma(\cdot)$ is separable.

Table 7. *Swaption price comparisons for the second volatility structure.*

Swaption length	Strike	Price description	Swaption maturity			
			0.25	1	2	4
1	IN	true	69.84	123.44	159.82	121.77
		vol approx	69.91	123.62	160.09	121.74
		rank 1	69.83	123.44	159.87	121.76
	AT	true	36.95	69.31	92.79	75.99
		vol approx	37.01	69.44	93.04	75.95
		rank 1	36.94	69.31	92.86	75.98
	OUT	true	13.07	23.62	30.89	24.24
		vol approx	13.08	23.63	30.98	24.16
		rank 1	13.06	23.62	30.95	24.20
2	IN	true	121.42	220.52	249.97	229.35
		vol approx	121.84	221.27	249.57	229.20
		rank 1	121.26	220.12	250.11	229.37
	AT	true	63.03	120.88	143.94	143.69
		vol approx	63.43	121.58	143.18	143.52
		rank 1	62.87	120.52	144.02	143.72
	OUT	true	22.48	41.67	49.02	45.80
		vol approx	22.66	41.96	48.07	45.55
		rank 1	22.37	41.50	48.93	45.68
4	IN	true	194.98	343.86	416.41	433.46
		vol approx	195.34	342.57	413.61	432.54
		rank 1	194.66	342.88	413.32	433.03
	AT	true	100.24	188.39	241.57	279.52
		vol approx	100.60	186.81	237.83	278.05
		rank 1	99.93	187.18	237.41	278.76
	OUT	true	35.99	66.04	83.07	88.45
		vol approx	36.18	64.78	79.73	86.78
		rank 1	35.82	65.07	79.32	87.50
8	IN	true	337.74	599.84	728.68	
		vol approx	333.90	590.10	715.81	
		rank 1	329.26	587.22	719.30	
	AT	true	178.09	340.50	441.36	
		vol approx	173.28	328.00	424.19	
		rank 1	167.57	324.92	429.17	
	OUT	true	65.70	122.64	153.97	
		vol approx	62.01	112.37	139.49	
		rank 1	57.64	110.47	144.33	

Table 8. *Delta comparisons for Libor model and Black swaptions for the second volatility structure.*

Swaption			Swaption maturity			
length	Strike	Model	0.25	1	2	4
1	IN	Black	0.750	0.750	0.750	0.750
		Libor	0.751	0.751	0.751	0.750
	AT	Black	0.523	0.539	0.549	0.570
		Libor	0.524	0.540	0.550	0.570
	OUT	Black	0.250	0.249	0.249	0.250
		Libor	0.250	0.251	0.250	0.250
2	IN	Black	0.751	0.751	0.749	0.750
		Libor	0.753	0.753	0.750	0.750
	AT	Black	0.519	0.533	0.546	0.572
		Libor	0.520	0.535	0.546	0.572
	OUT	Black	0.248	0.248	0.252	0.250
		Libor	0.250	0.250	0.252	0.250
4	IN	Black	0.751	0.749	0.748	0.750
		Libor	0.752	0.750	0.751	0.752
	AT	Black	0.516	0.532	0.547	0.580
		Libor	0.516	0.531	0.547	0.582
	OUT	Black	0.249	0.252	0.255	0.252
		Libor	0.249	0.251	0.250	0.253
8	IN	Black	0.745	0.744	0.745	
		Libor	0.759	0.757	0.755	
	AT	Black	0.518	0.538	0.557	
		Libor	0.521	0.543	0.563	
	OUT	Black	0.257	0.260	0.262	
		Libor	0.245	0.254	0.262	

Proof Set

$$a(t, u) = \sqrt{\int_0^t \gamma^2(s, u) ds},$$

$$\dot{a}(t, u) = \frac{\partial a(t, u)}{\partial t},$$

rewrite (37) as

$$\int_0^t \gamma(s, u) \gamma(s, v) ds = a(t, u) a(t, v),$$

Table 9. *Gamma and vega comparisons for Libor model and Black swaptions for the second volatility structure.*

Greek type	Swaption			Swaption maturity			
	length	Strike	Model	0.25	1	2	4
Gamma	1	IN	Black	32.0	15.9	10.6	10.7
			Libor	31.8	15.7	10.4	10.6
		AT	Black	40.1	19.8	13.2	13.2
			Libor	40.0	19.8	13.1	13.2
		OUT	Black	32.0	15.8	10.6	10.7
			Libor	32.1	16.0	10.7	10.8
	2	IN	Black	35.7	17.2	13.0	10.9
			Libor	35.1	16.8	12.8	10.7
		AT	Black	44.9	21.6	16.2	13.4
			Libor	44.6	21.4	16.2	13.4
		OUT	Black	35.7	17.2	13.0	10.9
			Libor	35.8	17.4	13.3	11.1
	4	IN	Black	40.7	20.2	14.3	10.4
			Libor	40.0	20.0	14.2	10.0
		AT	Black	51.1	25.2	17.7	12.8
			Libor	50.9	25.5	18.2	12.7
		OUT	Black	40.6	20.2	14.4	10.4
			Libor	41.0	20.8	15.1	11.0
	8	IN	Black	39.4	19.3	13.7	
			Libor	41.0	19.6	13.5	
		AT	Black	48.9	23.9	16.8	
			Libor	53.4	25.7	17.6	
		OUT	Black	39.6	19.5	13.9	
			Libor	43.1	21.8	15.6	
Vega	1	IN	Black	0.118	0.081	0.078	0.037
			Libor	0.117	0.080	0.077	0.037
		AT	Black	0.147	0.101	0.098	0.046
			Libor	0.147	0.101	0.097	0.046
		OUT	Black	0.118	0.081	0.078	0.037
			Libor	0.118	0.082	0.079	0.038
	2	IN	Black	0.163	0.106	0.074	0.036
			Libor	0.160	0.103	0.073	0.035
		AT	Black	0.205	0.132	0.092	0.044
			Libor	0.204	0.131	0.092	0.044
		OUT	Black	0.163	0.105	0.074	0.036
			Libor	0.163	0.107	0.076	0.036

Table 9. (cont.)

Greek type	Swaption length	Strike	Model	Swaption maturity			
				0.25	1	2	4
Vega	4	IN	Black	0.199	0.101	0.064	0.030
			Libor	0.196	0.100	0.064	0.029
		AT	Black	0.250	0.126	0.080	0.036
			Libor	0.249	0.127	0.082	0.036
		OUT	Black	0.199	0.101	0.065	0.030
			Libor	0.200	0.104	0.068	0.031
	8	IN	Black	0.155	0.074	0.046	
			Libor	0.161	0.075	0.045	
		AT	Black	0.192	0.091	0.056	
			Libor	0.210	0.098	0.059	
		OUT	Black	0.155	0.074	0.046	
			Libor	0.169	0.083	0.052	

differentiate with respect to time t to get

$$\frac{\gamma(t, u)}{a(t, u)} \frac{\gamma(t, v)}{a(t, v)} = \frac{\dot{a}(t, u)}{a(t, u)} + \frac{\dot{a}(t, v)}{a(t, v)},$$

and then with respect to v to get

$$\frac{\partial}{\partial v} \left(\frac{\gamma(t, v)}{a(t, v)} \right) \bigg/ \frac{\partial}{\partial v} \left(\frac{\dot{a}(t, v)}{a(t, v)} \right) = \frac{a(t, u)}{\gamma(t, u)}. \quad (38)$$

Since the left hand side of (38) is a function of only t and v , while the right hand side is a function of only t and u , both must be functions of just t . For some function $b(t)$, we must therefore have

$$\int_0^t \gamma^2(s, u) ds = b(t) \gamma^2(t, u).$$

Differentiation with respect to t , rearrangement, and then integration with respect to t gives

$$\begin{aligned} \frac{\partial \gamma^2(t, u)}{\partial t} \bigg/ \gamma^2(t, u) &= [1 - \dot{b}(t)] / b(t), \\ \ln[\gamma(t, u)] &= \frac{1}{2} \int_0^t [1 - \dot{b}(s)] / b(s) ds + c(u), \end{aligned}$$

Market Fit Volatility Function

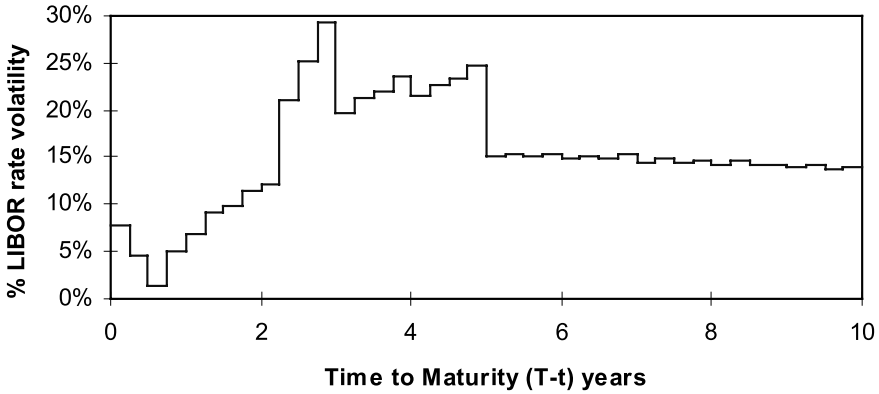


Fig. 3. Graphical representation of the first volatility structure.

where $c(\cdot)$ is an arbitrary function of u . Setting

$$\begin{aligned}\psi(t) &= \exp\left(\frac{1}{2} \int_0^t [1 - \dot{b}(s)] / b(s) ds\right), \\ \phi(u) &= \exp(c(u)),\end{aligned}$$

gives

$$\gamma(t, u) = \psi(t)\phi(u),$$

which is the result. □

Appendix C. Yield curve and volatility structures

C.1 Market fit volatility structure

The first volatility structure (Figure 3) is a simple one-factor homogeneous parametrisation to market data – the first six months of 1997 UK market data being used here. The yield curve used (Figure 4) is a typical one for that period of time.

C.2 Pathological volatility structure

The second volatility structure was chosen intentionally to be *pathological*, or representative of an extreme market situation. The functions were also optimised in order to ensure that some of the 15 swaptions to be tested had extreme rank two swaption covariance matrices.

Market Yield Curve - Forward LIBOR values

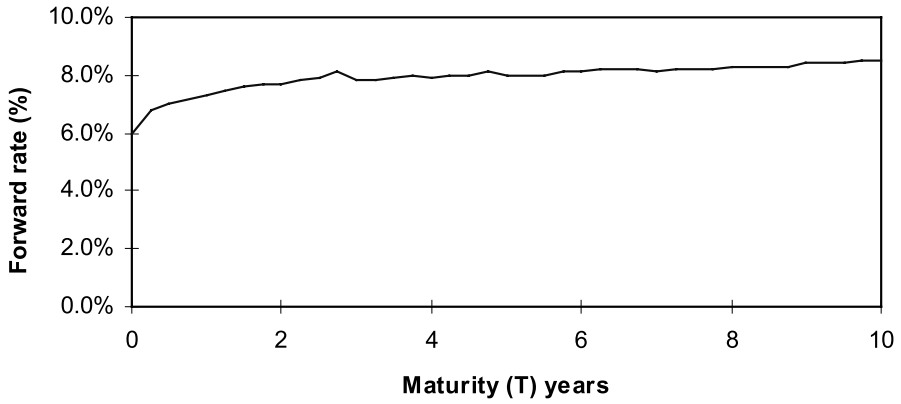


Fig. 4. Forward Libor rates used in conjunction with the first volatility structure.

Yield Curve



Fig. 5. Yield curve associated with the second volatility structure.

The functional form chosen for the yield curve was

$$\text{Yield}(T) = \begin{cases} 0.07 + 0.03T/3 & \text{for } T < 3 \\ 0.10 - 0.02(T - 3)/7 & \text{otherwise} \end{cases}$$

2-Factor Pathological Volatility Function

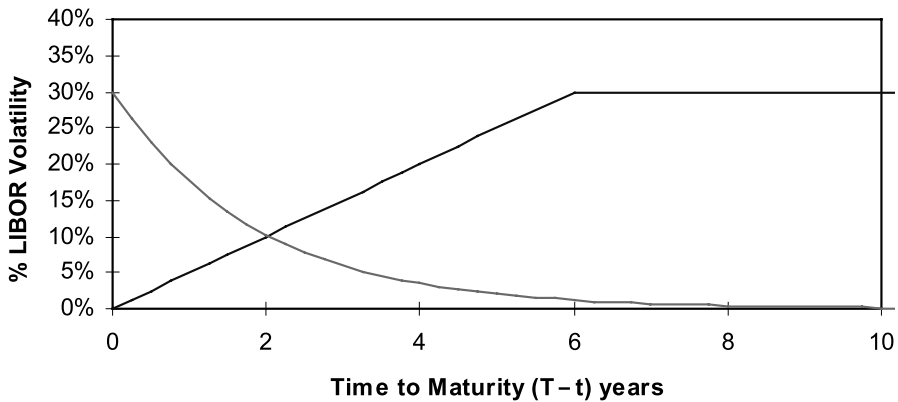


Fig. 6. Graphical representation of the two factors of the second volatility structure.

and is shown in Figure 5, while the equations for the volatility were

$$\begin{aligned} \gamma_1(t, T) &= \begin{cases} 0.05(T - t) & \text{for } (T - t) < 6 \\ 0.3 & \text{otherwise} \end{cases} \\ \gamma_2(t, T) &= 0.3 \exp(-0.54(T - t)) \end{aligned}$$

and these are graphed in Figure 6.

References

- Brace, A. (1996), Dual swap and swaption formulae in the normal and lognormal models. *University of New South Wales Preprint*.
- Brace, A. (1998), Simulation in the GHJM and LFM models. *FMMA notes*.
- Brace, A., Gatarek, D. and Musiela, M. (1997), The market model of interest rate dynamics. *Math. Finance* **7**, 127–54.
- Dudenhause, A., Schlögl, E. and Schlögl, L. (1998), Robustness of Gaussian hedges under parameter and model misspecification. *Working paper*, University of Bonn.
- Dun, T., E., Schlögl and Barton, G. (1999), Simulated swaption delta-hedging in the lognormal forward LIBOR model. Forthcoming in the *International Journal of Theoretical and Applied Finance* **4**(1) 2001.
- Glasserman, P. and Zhao, X. (2000), Arbitrage-free discretization of lognormal forward LIBOR and swap rate models. *Finance Stochast* **4**(1), 35–68.
- Hunt, P., Kennedy, J. and Pelsser, A. (1997), Markov functional interest rate models. *ABN Amro preprint*.
- Jamshidian, F. (1997), Libor and swap market models and measures. *Finance Stochast.* **1**, 293–330.
- Miltersen, K., Sandmann, K. and Sondermann, D. (1997), Closed form solutions for term structure derivatives with lognormal interest rates. *J. Finance* **52**, 407–30.

- Musiela, M. and Rutkowski, M. (1997a) *Martingale Methods in Financial Modelling*. Springer-Verlag, Berlin.
- Musiela, M., Rutkowski, M. (1997b) Continuous-time term structure models: a forward measure approach. *Finance Stochast.* **1**, 261–91.
- Plackett, R.L. (1954), A reduction formula for normal multivariate integrals *Biometrika* **41**, 351–60.
- Rebonato, R. (1999), On the pricing implications of the joint log-normal assumptions for the swaption and cap markets. *Journal of Computational Finance* **2**(3), 57–76.

Infinite Dimensional Diffusions, Kolmogorov Equations and Interest Rate Models

B. Goldys and M. Musiela

1 Introduction

The common feature of interest rate models is, that taking the Heath, Jarrow and Morton model Heath *et al.* (1992) as a starting point they naturally lead to infinite dimensional Markov processes which describe the arbitrage free dynamics of forward rates. By a forward rate $r(t, x)$ we mean the continuously compounded forward rate prevailing at time t over the time interval $[t + x, t + x + dx]$. Usually, the time evolution of forward curves $r(t, \cdot)$ is completely determined by the initial curve and the volatility structure. The question how to determine the volatility structure is a delicate one and different approaches can be chosen to address this problem; for possible answers see Musiela (1993), Brace and Musiela (1994), Goldys *et al.* (1995) or Brace *et al.* (1997). In this chapter, however, we assume that the volatility structure $\{\sigma(t, x) : t \geq 0, x \geq 0\}$ is a known vector-valued stochastic process. In that case the forward rate process $\{r(t, x) : t \geq 0, x \geq 0\}$ must satisfy the following stochastic partial differential equation

$$dr(t, x) = \frac{\partial}{\partial x} \left(\left(r(t, x) + \frac{1}{2} |\sigma(t, x)|^2 \right) dt + \sigma(t, x) dW(t) \right) \quad (1.1)$$

for all $t, x \geq 0$, where W is a d -dimensional Brownian motion. It has been shown in Musiela (1993) that (1.1) is sufficient for the nonarbitrage condition. We will concentrate on two models:

- Gaussian $r(t, x)$ model for its theoretical and computational simplicity, BGM model.

We start with the derivation of the stochastic PDE which is satisfied by the forward rate process $\{r(t, x) : t, x \geq 0\}$. We model the uncertainty of future interest rate movements using an infinite family of Wiener processes $\{W_k : k \geq 1\}$ defined on the common stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$. We assume that $(\mathcal{F}_t)$ is a $\mathbb{P}$ -augmentation of the natural filtration $\sigma(W_k(s) : s \leq t, k \geq 1)$. Let

$\{X(t, x) : t, x \geq 0\}$ be an arbitrary random field. We say that X is adapted to the filtration $(\mathcal{F}_t)$ if

$$\sigma(X(s, x) : s \leq t, x \geq 0) \subset \mathcal{F}_t$$

for every $t \geq 0$.

Let $P(t, T)$ denote the price at time $t \geq 0$ of a zero coupon bond with maturity $T \geq t$. We assume that

$$P(t, T) = \exp\left(-\int_0^{T-t} r(t, u) du\right) \quad (1.2)$$

for a certain measurable random field $\{r(t, x) : t, x \geq 0\}$ which is locally bounded: for every $T > 0$

$$\sup_{t, x \leq T} |r(t, x)| < \infty, \quad \mathbb{P}\text{-a.s.} \quad (1.3)$$

It follows that the process of saving account

$$\beta(t) = \exp\left(\int_0^t r(u, 0) du\right), \quad t \geq 0,$$

is well defined. The discounted price of the zero coupon is defined as

$$N(t, T) = \frac{P(t, T)}{\beta(t)}, \quad t \leq T. \quad (1.4)$$

Theorem 1.1 *Let (1.3) hold and let the random field r be adapted to $(\mathcal{F}_t)$. Assume that for every $T > 0$ the process $\{N(t, T) : t \leq T\}$ is a $(\mathbb{P}, (\mathcal{F}_t))$ -martingale and, moreover,*

$$\mathbb{E} \int_0^R \langle \log N(\cdot, T) \rangle_t dT < \infty, \quad R > 0. \quad (1.5)$$

Then there exists a family $\{\sigma_k : k \geq 1\}$ of adapted random fields such that for every $T > 0$ and $k \geq 1$

$$\sup_{t, x \leq T} |\sigma_k(t, x)| < \infty, \quad \mathbb{P}\text{-a.s.},$$

$$\sum_{k=1}^{\infty} \int_0^T \int_0^T \sigma_k^2(t, x) dx dt < \infty, \quad \mathbb{P}\text{-a.s.},$$

and

$$\begin{aligned} & \int_0^x r(t, u) du + \int_0^t r(s, 0) ds = \int_0^{x+t} r(0, u) du \\ & + \sum_{k=1}^{\infty} \int_0^t \sigma_k(s, x+t-s) dW_k(s) + \frac{1}{2} \sum_{k=1}^{\infty} \int_0^t \sigma_k^2(s, x+t-s) ds. \end{aligned}$$

Proof For every $T > 0$ the process $N(\cdot, T)$ is continuous and positive. Fix $R > 0$ and define the process N for all $t \geq 0$ and $T \in [0, R]$ putting $N(t, T) = N(T, T)$ for $t \geq T$. Then for every $T \leq R$ the process $\{N(t, T) : t \leq R\}$ is a continuous square integrable martingale. Therefore, for every $T > 0$ there exists a continuous local martingale $M(\cdot, T)$ with $M(0, T) = 0$ such that

$$N(t, T) = P(0, T) \exp \left(-M(t, T) - \frac{1}{2} \langle M(\cdot, T) \rangle_t \right), \quad T \leq R,$$

and $M(t, T) = M(T, T)$ for $t \geq T$. By (1.5) $M(t, \cdot)$ is a $L^2(0, R)$ -valued continuous martingale for every $R > 0$. It follows from Theorem 8.2 in Da Prato and Zabczyk (1992) that there exists a family $\{h_k : k \geq 1\}$ of predictable $L^2(0, R)$ -valued processes, such that for $t, T \leq R$

$$M(t, T) = \sum_{k=1}^{\infty} \int_0^t h_k(s, T) dW_k(s)$$

and

$$\mathbb{E} \sum_{k=1}^{\infty} \int_0^R \int_0^t h_k^2(s, T) dT ds < \infty.$$

It is easy to see that the processes $h_k, k \geq 1$, may be chosen independently of R . Hence, for $t, x \geq 0$ we may define $\sigma_k(t, x) = h_k(t, x + t)$ and then

$$\begin{aligned} N(t, x + t) = & \exp \left(- \int_0^{t+x} r(0, u) du \right. \\ & \left. - \sum_{k=1}^{\infty} \int_0^t \sigma_k(s, x + t - s) dW_k(s) - \frac{1}{2} \sum_{k=1}^{\infty} \int_0^t \sigma_k^2(s, x + t - s) ds \right) \end{aligned}$$

and the theorem follows. $\square$

In the sequel we assume that for each $x \geq 0$

$$dr(t, x) = g(t, x)dt + \sum_{k=1}^{\infty} \tau_k(t, x) dW_k(t). \quad (1.6)$$

The random fields $\{g(t, x) : t, x \geq 0\}$ and $\{\tau_k(t, x) : t, x \geq 0\}, k \geq 1$, satisfy the following conditions.

(C1) For every $T > 0$

$$\sup_{t, x \leq T} |g(t, x)| < \infty, \quad \mathbb{P}\text{-a.s.},$$

and for every $T > 0$ and $k \geq 1$

$$\sup_{t, x \leq T} |\tau_k(t, x)| < \infty \quad \mathbb{P}\text{-a.s.}$$

(C2) For every $T > 0$

$$\sum_{k=1}^{\infty} \int_0^T \int_0^T \tau_k^2(t, x) dx dt < \infty. \quad \mathbb{P}\text{-a.s.}$$

(C3) For every $t > 0$

$$\sigma(g(s, x) : s \leq t, x \geq 0) \cup \sigma(\tau_k(s, x) : s \leq t, x \geq 0, k \geq 1) \subset \mathcal{F}_t.$$

(C4) $\sigma\{r(0, x) : x \geq 0\} \in \mathcal{F}_0$ and for every $T > 0$

$$\sup_{x \leq T} |r(0, x)| < \infty.$$

Theorem 1.2 Assume that for all $t, x \geq 0$

$$\int_0^x g(t, u) du = r(t, x) - r(t, 0) + \frac{1}{2} \sum_{k=1}^{\infty} \left(\int_0^{T-t} \tau_k(t, u) du \right)^2. \quad (1.7)$$

Then for all $T > 0$ the process

$$M_T(t) = \frac{P(t, T)}{\beta(t)}, \quad t \in [0, T],$$

is a $\mathbb{P}$ -local martingale and a $\mathbb{P}$ -martingale, if in addition the process $\{r(t, x) : t, x \geq 0\}$ is bounded on $[0, T] \times \Omega$ for all $T > 0$.

Proof We have

$$\begin{aligned} d \log P(t, T) &= -d \left(\int_0^{T-t} r(t, u) du \right) \\ &= r(t, T-t) dt - \int_0^{T-t} \left(g(t, u) du + \sum_{k=1}^{\infty} \tau_k(t, u) dW_k(t) \right) du \\ &= r(t, T-t) dt - \left(\int_0^{T-t} g(t, u) du \right) dt \\ &\quad - \sum_{k=1}^{\infty} \left(\int_0^{T-t} \tau_k(t, u) du \right) dW_k(t). \end{aligned}$$

Hence, the quadratic variation of $\log P(\cdot, T)$ is given by

$$d \langle \log P(\cdot, T) \rangle (t) = \sum_{k=1}^{\infty} \left(\int_0^{T-t} \tau_k(t, u) du \right)^2 dt.$$

Therefore,

$$dP(t, T) = P(t, T) \left(r(t, T-t) - \int_0^{T-t} g(t, u) du \right.$$

$$+ \frac{1}{2} \sum_{k=1}^{\infty} \left(\int_0^{T-t} \tau_k(t, u) du \right)^2 \Big) dt - P(t, T) \sum_{k=1}^{\infty} \int_0^{T-t} \tau_k(t, u) dW_k(t).$$

The last equation yields

$$\begin{aligned} \frac{P(t, T)}{\beta(t)} &= P(0, T) \exp \left(- \sum_{k=1}^{\infty} \int_0^t \left(\int_0^{T-s} \tau_k(s, u) du \right) dW_k(s) \right. \\ &\quad \left. - \frac{1}{2} \sum_{k=1}^{\infty} \int_0^t \left(\int_0^{T-s} \tau_k(s, u) du \right)^2 ds \right) \end{aligned} \quad (1.8)$$

which concludes the proof. $\square$

Remark 1.3 The above theorem has been proved in Musiela (1993) for the finite dimensional Wiener process, that is for a certain $d \geq 1$, $\tau_k = 0$ for $k > d$. An extension to the case when the number of driving Wiener processes is infinite has been proposed in Santa-Clara and Sornette (1997).

We will reparametrize equation (1.8) putting $T = t + x$. Since

$$P(0, t + x) = \exp \left(- \int_0^{t+x} r(0, u) du \right),$$

we find that (1.8) takes the form

$$\begin{aligned} \frac{P(t, t + x)}{\beta(t)} &= \exp \left(- \int_0^{t+x} r(0, u) du \right) \\ &\quad \cdot \exp \left(- \sum_{k=1}^{\infty} \int_0^t \left(\int_0^{t+x-s} \tau_k(s, x) dx \right) dW_k(s) \right. \\ &\quad \left. - \frac{1}{2} \sum_{k=1}^{\infty} \int_0^t \left(\int_0^{t+x-s} \tau_k(s, x) dx \right)^2 ds \right). \end{aligned} \quad (1.9)$$

Under the appropriate regularity conditions on the coefficients τ_k we obtain formally from (1.9)

$$\begin{aligned} r(t, x) &= r(0, t + x) + \sum_{k=1}^{\infty} \int_0^t \tau_k(s, x + t - s) \left(\int_0^{x+t-s} \tau_k(s, u + t - s) du \right) ds \\ &\quad + \sum_{k=1}^{\infty} \int_0^t \tau_k(s, x + t - s) dW_k(s). \end{aligned} \quad (1.10)$$

If we assume that $\tau_k(s, x) = f_k(r(u, y) : u \leq s, y \geq 0)(x)$ for $k \geq 1$ then (1.10) defines a stochastic integral equation for the random field $\{r(t, x) : t, x \geq 0\}$. Such an approach has been studied in Kennedy (1994) and Hamza and Klebaner (1995).

In this chapter we take another approach, well known in the theory of stochastic partial differential equations. We will transform (1.10) into a stochastic evolution equation in an appropriate function space. To this end we define first a scale of weighted L^2 -spaces in the following way.

First, we assume that for every $t \geq 0$ the forward curve $r(t, x)$ is defined for all $x \geq 0$. Hence, the state of the forward rate process $r(t)$ at time t is the curve $\{r(t, x) : x \geq 0\}$. In order to allow bounded, for example constant forward rates, we assume that for a certain $\alpha > 0$

$$\int_0^\infty r^2(t, x)e^{-\alpha x} dx < \infty \quad \mathbb{P} - \text{a.s.}$$

It follows that a state space for the process $\{r(t) : t \geq 0\}$ is the space $L_\alpha^2(0, \infty)$ of functions with the finite norm

$$\|f\|_\alpha^2 = \int_0^\infty f^2(x)e^{-\alpha x} dx.$$

The space $L_\alpha^2(0, \infty)$ is a Hilbert space with the inner product

$$\langle f, g \rangle_\alpha = \int_0^\infty f(x)g(x)e^{-\alpha x} dx.$$

For $f \in L_\alpha^2(0, \infty)$ we define the semigroup of left shifts

$$S(t)f(\cdot) = f(t + \cdot), \quad t \geq 0.$$

Then (1.10) may be rewritten as

$$\begin{aligned} r(t) &= S(t)r_0 + \sum_{k=1}^\infty \int_0^t S(t-s)\tau_k(s) \left(\int_0^\cdot \tau_k(s, u)du \right) ds \\ &\quad + \sum_{k=1}^\infty \int_0^t S(t-s)\tau_k(s)dW_k(s). \end{aligned}$$

We will restrict our considerations to the class of forward rate processes defined by the Markovian dynamics on $L_\alpha^2(0, \infty)$, that is we assume that

$$\tau_k(s) = \tau_k(s, r(s))(\cdot) \in L_\alpha^2(0, \infty),$$

where the same notation τ_k is preserved. Then

$$\begin{aligned} r(t) &= S(t)r_0 + \sum_{k=1}^\infty \int_0^t S(t-s)\tau_k(s, r(s)) \left(\int_0^\cdot \tau_k(s, r(s))(u)du \right) ds \\ &\quad + \sum_{k=1}^\infty \int_0^t S(t-s)\tau_k(s, r(s))dW_k(s). \end{aligned} \tag{1.11}$$

Let $\tau : L_\alpha^2(0, \infty) \rightarrow \mathbb{R}$ be defined by the formula

$$G(t, f)(x) = \sum_{k=1}^{\infty} \tau_k(t, f)(x) \int_0^x \tau_k(t, f)(u) du.$$

where $G : L_\alpha^2(0, \infty) \rightarrow L_\alpha^2(0, \infty)$ and

$$\tau(t, f) = \sum_{k=1}^{\infty} \tau_k(t, f(t)) e_k$$

Let $\{e_k : k \geq 1\}$ be a complete orthonormal system in $L_\alpha^2(0, \infty)$. We denote by

$$W(t) = \sum_{k=1}^{\infty} W_k(t) e_k, \quad t \geq 0,$$

the standard cylindrical Wiener process on $L_\alpha^2(0, \infty)$. By this we mean that W is a process of continuous random functionals on $L_\alpha^2(0, \infty)$ with the properties:

$$\langle W(t), f \rangle \sim N(0, t \|f\|^2), \quad t \geq 0, f \in L_\alpha^2(0, \infty),$$

$$\mathbb{E} \langle W(t), f \rangle \langle W(s), g \rangle = \langle f, g \rangle \min(s, t).$$

Then, (1.11) takes the form of the following integral equation in $L_\alpha^2(0, \infty)$

$$r(t) = S(t)r_0 + \int_0^t S(t-s)G(s, r(s))ds + \int_0^t S(t-s)\tau(s, r(s))dW(s). \quad (1.12)$$

Definition 1.4 The $L_\alpha^2(0, \infty)$ -valued $(\mathcal{F}_t)$ -predictable process r is a solution to (1.12) with the initial condition $r_0 \in L_\alpha^2(0, \infty)$ if

(a) for all $t \geq 0$

$$\int_0^t \|G(s, r(s))\| ds + \sum_{k=1}^{\infty} \int_0^t \|\tau_k(s, r(s))\|_2^2 ds < \infty, \quad \mathbb{P}\text{-a.s.},$$

where

$$\|\tau(s, r(s))\|_2^2 = \sum_{k=1}^{\infty} \|\tau_k(s, r(s))\|^2.$$

(b) for every $t \geq 0$ equation (1.12) holds $\mathbb{P}$ -a.s.

In the theorem below we use the general theory of equations of type (1.12) developed in Da Prato and Zabczyk (1992) to provide conditions for existence and uniqueness of solutions to (1.12).

Theorem 1.5 Assume that piecewise continuous functions $\tau_k : \mathbb{R}_+ \times \mathbb{R} \rightarrow \mathbb{R}_+$, $k \leq d$ satisfy the following conditions: for every $T > 0$ there exists $C_T > 0$ such that

$$\sup_{x \geq 0, t \leq T} \tau_k(t, x) < \infty$$

$$|\tau_k(t, x) - \tau_k(t, y)| \leq C_T |x - y|, \quad t \leq T.$$

Then for every $\alpha > 0$ there exists a unique solution to (1.12) for every $r_0 \in L^2_\alpha(0, \infty)$.

Remark 1.6 The above theorem does not assure positivity of forward rates. If we assume that $r_0 \geq 0$ then under appropriate conditions on τ_k one may obtain existence and uniqueness of nonnegative solutions. We do not pursue this topic here. For an example of equation (1.12) with nonnegative solutions see Goldys *et al.* (1995).

It is well known that equation (1.10) is intimately related to a stochastic partial differential equation

$$\left\{ \begin{array}{l} dr(t, x)(t, x) = \left(\frac{\partial r}{\partial x}(t, x) + \sum_{k=1}^{\infty} \tau_k(t, r(t, x)) \int_0^x \tau_k(t, r(t, y)) dy \right) dt \\ \quad + \sum_{k=1}^{\infty} \tau_k(t, r(t, x)) dW_k(t), \\ r(0, x) = r_0(x). \end{array} \right. \quad (1.13)$$

We will discuss this relationship at the level of the evolution equation (1.12). In the space $L^2_\alpha(0, \infty)$ we introduce an operator $A = \frac{\partial}{\partial x}$ with the domain

$$\text{dom}(A) = H^1_\alpha(0, \infty) = \left\{ f \in L^2_\alpha(0, \infty) : \int_0^\infty \left| \frac{\partial f}{\partial x}(x) \right|^2 e^{-\alpha x} dx < \infty \right\},$$

where the derivative is meant in the generalized sense. Equation (1.13) considered in $L^2_\alpha(0, \infty)$ takes the form

$$\left\{ \begin{array}{l} dr(t) = (Ar(t) + G(t, r(t))) dt + \tau(t, r(t)) dW(t), \\ r(0) = r_0. \end{array} \right. \quad (1.14)$$

The latter equation, however, does not need to have classical solutions unless further regularity conditions are imposed on the data (see below). In general we define a solution to (1.14) in the mild sense as a solution to (1.12). The relationship between the two equations is clarified by the next theorem, which follows from the general theory developed in Da Prato and Zabczyk (1992).

Theorem 1.7 Assume that the functions $\tau_k, k \leq d$, satisfy assumptions of theorem 1.5 and let r be a solution to (1.12). Then the following holds.

(i) Equation (1.13) holds x -a.e. if and only if $\tau_k(t, \cdot) \in H_\alpha^1$ for all $t \geq 0$ and $r_0 \in H_\alpha^1$.

(ii) There exist sequences $(\tau_k^n(t, \cdot)), (r_0^n) \subset H_\alpha^1, k \leq d$ converging in the $L_\alpha^2(0, \infty)$ -norm to $\tau_k(t, \cdot)$ and r_0 respectively and such that the corresponding solutions of (1.13) satisfy the condition

$$\lim_{n \rightarrow \infty} \mathbb{E} \int_0^T \|r^n(t) - r(t)\|_\alpha^2 dt = 0.$$

Proof The standard proof of this theorem is omitted. □

2 The BGM Model

In this section our starting point is the model of Libor rate process proposed in Brace *et al.* (1997).

Let $L(t, x)$ denote the Libor rate process defined by the formula

$$1 + \delta L(t, x) = \frac{P(t, t + x)}{P(t, t + x + \delta)}, \quad t, x \geq 0,$$

where $\delta > 0$ (for example $\delta = 0.25$) is fixed. We assume that all zero coupons may be expressed in terms of a certain forward rate process r given in (1.2) but we shift our attention to the process $\log L(t, x)$ which is supposed to satisfy an equation

$$d(\log L(t, x)) = \alpha(t, x)dt + \gamma(t, x)dW(t), \quad x \geq 0, \quad (2.1)$$

W is a d -dimensional Wiener process. We need conditions on the drift term α which assure that there is no arbitrage.

We assume that the measurable function $\gamma : [0, \infty) \times [0, \infty) \rightarrow \mathbb{R}^d$ is deterministic,

$$M_\gamma = \sup_{t, x > 0} |\gamma(t, x)| + \sup_{t \geq 0, x \leq \delta} \sum_{k=0}^{\infty} |\gamma(t, x + k\delta)| < \infty. \quad (2.2)$$

Let l be a solution to the following stochastic evolution equation in $L_\alpha^2(0, \infty)$:

$$\begin{cases} dl(t) = (Al(t) + F(t, l(t)))dt + \gamma(t)dW(t), \\ l(0) = \phi \in L_\alpha^2(0, \infty), \end{cases} \quad (2.3)$$

where

$$F(t, \phi)(x) = \sum_{k=0}^{[x/\delta]} \frac{\delta \exp(\phi(x - k\delta))}{1 + \delta \exp(\phi(x - k\delta))} \langle \gamma(t, x - k\delta), \gamma(t, x) \rangle - \frac{1}{2} |\gamma(t, x)|^2.$$

If this equation has a solution then we may define the process L via the formula $l(t, x) = \log L(t, x)$. In turn (2) allows us to define the family of zero coupons and finally the forward rate process $r(t)$ can be defined provided the appropriate regularity conditions are satisfied. It was shown in Brace *et al.* (1997) that if l is a solution to (2.3) then the corresponding process of forward rates satisfies the nonarbitrage condition (1.5).

Theorem 2.1 Assume (2.1). Then the following holds.

(a) For every $\alpha > 0$ there exists a unique solution to (2.3) in the space $L_\alpha^2(0, \infty)$.

(b) Let $\alpha \leq 0$ and

$$N_\gamma^2 = \sup_{t \geq 0} \int_0^\infty e^{-\alpha x} |\gamma(t, x)|^2 dx < \infty. \quad (2.4)$$

Then there exists a unique solution to (2.3) in $L_\alpha^2(0, \infty)$.

Proof Note first that

$$|F(t, \phi)(x)| \leq |\gamma(t, x)| \sum_{k=0}^{[x/\delta]} |\gamma(t, x - k\delta)| + \frac{1}{2} |\gamma(t, x)|^2$$

and therefore

$$\begin{aligned} & \int_0^\infty e^{-\alpha x} |F(t, \phi)(x)|^2 dx \\ & \leq 2 \int_0^\infty e^{-\alpha x} |\gamma(t, x)|^2 \left(\sum_{k=0}^{[x/\delta]} |\gamma(t, x - k\delta)| \right)^2 dx + \frac{1}{2} \int_0^\infty e^{-\alpha x} |\gamma(t, x)|^4 dx \\ & \leq 2 \sum_{n=0}^\infty e^{-\alpha \delta n} \int_0^\delta |\gamma(t, x + n\delta)|^2 \left(\sum_{k=0}^n |\gamma(t, x + k\delta)| \right)^2 dx \\ & \quad + \frac{1}{2} M_\gamma^2 \int_0^\infty e^{-\alpha x} |\gamma(t, x)|^2 dx. \end{aligned} \quad (2.5)$$

Therefore, for $\alpha > 0$

$$\|F(t, \phi)\|^2 \leq 2\delta M_\gamma^4 \sum_{n=0}^\infty n^2 e^{-\alpha \delta n} + \frac{1}{2\alpha} M_\gamma^4 < \infty.$$

If $\alpha \leq 0$ then (2.3), (2.4) and (2.5) yield

$$\|F(t, \phi)\|^2 \leq \frac{3}{2} M_\gamma^2 \|\gamma(t)\|^2.$$

Hence, for every $\alpha \in \mathbb{R}$ the mapping $F : [0, \infty) \times L_\alpha^2(0, \infty) \rightarrow L_\alpha^2(0, \infty)$ is uniformly bounded. We will show now that

$$\|F(t, \phi) - F(t, \psi)\| \leq M_F \|\phi - \psi\|, \quad \phi, \psi \in L_\alpha^2(0, \infty). \quad (2.6)$$

Since

$$\left| \frac{e^x}{1+e^x} - \frac{e^y}{1+e^y} \right| \leq \frac{1}{2}|x-y|,$$

we obtain, proceeding similarly as in (2.5),

$$\begin{aligned} \|F(t, \phi) - F(t, \psi)\|^2 &\leq \frac{1}{4} \sum_{n=0}^{\infty} e^{-\alpha \delta n} \int_0^\delta |\gamma(t, x + n\delta)|^2 \\ &\quad \left(\sum_{k=0}^n |\gamma(t, x + k\delta)| |(\phi - \psi)(x + k\delta)| \right)^2 dx \\ &\leq \frac{1}{4} M_\gamma^2 \sum_{n=0}^{\infty} e^{-\alpha \delta n} \int_0^\delta |\gamma(t, x + n\delta)|^2 \left(\sum_{k=0}^n (\phi - \psi)^2(x + k\delta) \right) dx. \end{aligned} \quad (2.7)$$

Hence, if $\alpha < 0$ then

$$\begin{aligned} \|F(t, \phi) - F(t, \psi)\|^2 &\leq \frac{1}{4} M_\gamma^4 \sum_{n=0}^{\infty} e^{-\alpha \delta n} \int_0^\delta \left(\sum_{k=0}^n (\phi - \psi)^2(x + k\delta) \right) dx \\ &= \frac{1}{4} M_\gamma^4 \int_0^\delta \sum_{k=0}^{\infty} (\phi - \psi)^2(x + k\delta) \sum_{n=k}^{\infty} e^{-\alpha \delta n} \\ &= \frac{M_\gamma^4}{4(1 - e^{-\alpha \delta})} \sum_{k=0}^{\infty} \int_0^\delta e^{-\alpha \delta k} (\phi - \psi)^2(x + k\delta) \\ &\leq \frac{M_\gamma^4}{4(1 - e^{-\alpha \delta})} e^{\alpha \delta} \sum_{k=0}^{\infty} \int_{k\delta}^{(k+1)\delta} e^{-\alpha x} (\phi - \psi)^2(x) dx \\ &= \frac{M_\gamma^4}{4(1 - e^{-\alpha \delta})} e^{\alpha \delta} \|\phi - \psi\|^2 \end{aligned}$$

and (2.6) follows. Assume now that $\alpha \leq 0$. Then by the first inequality in (2.7)

$$\begin{aligned} \|F(t, \phi) - F(t, \psi)\|^2 &\leq \frac{1}{4} \sum_{n=0}^{\infty} e^{-\alpha \delta n} \int_0^\delta |\gamma(t, x + n\delta)|^2 \\ &\quad \left(\sum_{k=0}^n |\gamma(t, x + k\delta)| |(\phi - \psi)(x + k\delta)| \right)^2 dx \\ &\leq \frac{1}{4} N_\gamma^2 \int_0^\delta \left(\sum_{k=0}^{\infty} |\gamma(t, x + k\delta)|^2 \right) \left(\sum_{k=0}^{\infty} e^{-\alpha \delta k} (\phi - \psi)^2(x + k\delta) \right) dx \end{aligned}$$

$$\leq \frac{1}{4} N_\gamma^4 \sum_{k=0}^{\infty} \int_{k\delta}^{(k+1)\delta} e^{-\alpha x} (\phi - \psi)^2(x) dx = \frac{1}{4} N_\gamma^4 \|\phi - \psi\|^2.$$

Finally, Theorem 7.4 in Da Prato and Zabczyk (1992) yields existence of a unique solution to equation (2.3). $\square$

3 Kolmogorov equations

The classical Black–Scholes formula for a European option price has been derived by solving a partial differential equation identified by means of heuristic arguments (cf. Black and Scholes 1973). Later on a probabilistic interpretation of the above arguments allowed the derivation to be made rigorous Harrison and Pliska (1981). Let us recall briefly the main ideas of this approach. Assume that the price $X(t)$ of a stock is a positive continuous semimartingale such that the logarithm of the stock price has a deterministic quadratic variation

$$\langle \log X \rangle_t = \sigma^2 t.$$

Then some mild technical conditions imply existence of a unique probability measure under which for every $t \geq 0$

$$X(t) = X_0 + \int_0^t r X(s) ds + \int_0^t \sigma X(s) dW(s).$$

Moreover, for a given maturity T and a strike price K we can calculate the price of a European put option by taking the conditional expectation of the discounted option payoff, i.e.,

$$V_T(t, x) = e^{-r(T-t)} E((K - X(T))^+ | X(t) = x)$$

for $t \leq T$. Since X is a strong Feller process with the infinitesimal generator

$$L = rx \frac{\partial}{\partial x} + \frac{1}{2} \sigma^2 x^2 \frac{\partial^2}{\partial x^2}$$

we can apply the Feynman–Kac formula and identify the function V_T with a unique solution of the backward Kolmogorov equation

$$\frac{\partial u}{\partial t}(t, x) + \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 u}{\partial x^2}(t, x) + rx \frac{\partial u}{\partial x}(t, x) - ru(t, x) = 0 \quad (3.1)$$

with the terminal condition $u(T, x) = (K - x)^+$.

In this section we investigate whether this strategy can be applied to interest rate options in general term structure models.

Consider a European swaption, an option with maturity T on a swap with the cashflows C_i , $i = 1, \dots, n$ at times T_i , $i = 1, 2, \dots, n$ such that $T < T_1 <$

$\dots < T_n$. Under some technical conditions the process $\{r(t, \cdot) : t \geq 0\}$ of forward curves, given by equation (1.1), is a strong Markov and Feller process in $L_\alpha^2(0, \infty)$. We will identify the form of its generator L on a class of cylindrical functions. Because the time t price of the swaption is given by the formula

$$V_T(t, \phi) = E \left(e^{-\int_t^T r(s,0) ds} \left(K - \sum_{i=1}^n C_i P(T, T_i) \right)^+ \middle| r(t, x) = \phi(x), x \geq 0 \right),$$

we can expect that in analogy with the finite dimensional case (3.1) the Feynman–Kac formula should lead to a parabolic differential equation for $V_T(\cdot, \cdot)$ of the form

$$\frac{\partial u}{\partial t}(t, \phi) + Lu(t, \phi) - \phi(0)u(t, \phi) = 0 \quad (3.2)$$

with the appropriate terminal condition $u(T, \phi)$.

We denote by δ the functional $\delta(\phi) = \phi(0)$ for $\phi \in H_\alpha^1$.

Let K be an arbitrary Hilbert space. For $p \geq 0$ we define the Banach space $C_p(K)$ of continuous functions $F : K \rightarrow \mathbb{R}$ such that

$$\|F\|_p = \sup_{k \in K} (e^{-p\|k\|} |F(k)|) < \infty.$$

Let $C_p^n(K)$ denote the subspace of $C_p(K)$ containing all functions F which are n times Fréchet continuously differentiable on K and such that

$$\|F\|_{n,p} = \sum_{j=0}^n \sup_{k \in K} (e^{-p\|k\|} \|D^j F(k)\|) < \infty,$$

where $D^j F(k)$, $j = 1, 2, \dots, n$ denotes the j -th Fréchet derivative of F and $D^0 F = F$. If $F \in C_p^1(K)$ then the derivative $DF(y)$ of F at $y \in K$ in the direction $k \in K$ may be identified with an element of the dual space K and $DF : K \rightarrow \mathbb{R}$ is continuous. If $F \in C_p^2(K)$ then the second derivative $D^2 F(k) : K \rightarrow K$ is a symmetric linear operator and the mapping $D^2 F : K \rightarrow \mathcal{L}(K)$ is continuous.

In the sequel the spaces $C_p^k(K)$ will be considered only for the two cases $K = L_\alpha^2(0, \infty)$ or $K = H_\alpha^1$.

Assume that the assumptions of Theorem 1.5 are satisfied. Then the process $r(\cdot, \zeta)$ is a strong Markov process on $L_\alpha^2(0, \infty)$ for any $\mathcal{F}_0$ -measurable initial condition ζ . Moreover, if $E\|\zeta\|^p < \infty$ for a certain $p \geq 2$ then for any $T > 0$

$$\sup_{t \leq T} E\|r(t, \zeta)\|^p \leq C_{T,p} (1 + E\|\zeta\|^p).$$

If $\tau(t, \cdot)$ is Fréchet differentiable on $L_\alpha^2(0, \infty)$ then for every $t \geq 0$ the mapping $\phi \rightarrow r(t, \phi)$ is Fréchet differentiable $\mathbb{P}$ -a.s. In general the solution to (3.5) is not a

semimartingale but for every $\psi \in \text{dom } (A^*) = \{\phi \in H^1 : \phi(0) = 0\}$

$$\begin{aligned} \langle r(t), \psi \rangle &= \langle \phi, \psi \rangle + \int_0^t \langle r(s), A^* \psi \rangle ds + \int_0^t \langle F(s, r(s)), \psi \rangle ds \\ &\quad + \int_0^t \langle G^*(s, r(s)) \psi, dW(s) \rangle \end{aligned} \quad (3.3)$$

and hence $\langle r(t), \psi \rangle$ is a semimartingale and so is the multidimensional process

$$(\langle r(t), \psi_1 \rangle, \dots, \langle r(t), \psi_n \rangle)$$

for any n and arbitrary collection of $\psi_1, \dots, \psi_n \in \text{dom } (A^*)$. It follows that the process r is an $L^2([0, T] \times \Omega, \lambda \otimes P)$ -limit of semimartingales for every $T > 0$. This property will be used later on in the discussion of the Kolmogorov equation. The following property of the process

$$R(t, \phi) = \int_0^t r(s, \phi) ds$$

will be useful.

Lemma 3.1 *For every $T > 0$ there exists $c_T > 0$ such that*

$$\sup_{t \leq T} E \| (R(t, \phi) - R(t, \psi)) \|_1 \leq c_T \|\phi - \psi\|.$$

Proof The standard proof of this lemma is omitted. □

Let us go back now to the problem of pricing interest rate dependent options. To begin with, note that in the present terminology the price of zero coupon can be rewritten as follows. Let

$$B_T(t, \phi) = e^{\langle \phi, S(t) I_{[0, T]} \rangle},$$

with $I_{[0, T]}$ denoting the indicator function of the interval $[0, T]$. It follows that $P(t, T) = B_T(t, r(t))$. Any measurable mapping $F : L^2_\alpha(0, \infty) \rightarrow \mathbb{R}$ such that

$$\sup_{t \leq T} E \left(|F(r(T))| \exp \left(- \int_t^T r(u, 0) du \right) \right) < \infty \quad (3.4)$$

represents an option with the payoff $F(r(T))$ at the maturity T . Due to the Markov property of the process r the time t ($\leq T$) price of the claim is

$$\begin{aligned} V_T(t) &= E \left(\exp \left(- \int_t^T r(u, 0) du \right) F(r(T)) \middle| \mathcal{F}_t \right) \\ &= E \left(\exp \left(- \int_t^T r(u, 0) du \right) F(r(T)) \middle| r(t) \right). \end{aligned}$$

The above can be rewritten using the function

$$V_T(t, \phi) = E \left(\exp \left(- \int_t^T \delta(r(u)) du \right) F(r(T)) \middle| r(t) = \phi \right). \quad (3.5)$$

The transformation $F \rightarrow V_T$ is closely related to the following ‘‘Feynman–Kac semigroup’’

$$P_t^\delta F(\phi) = E \left(\exp \left(- \int_0^t \delta(r(u, \phi)) du \right) F(r(t, \phi)) \right)$$

by a simple equation $V_T(t, \phi) = P_{T-t}^\delta F(\phi)$. Clearly $P_0^\delta F = F$ and the Markov property yields the semigroup property $P_{t+s}^\delta = P_t^\delta P_s^\delta$. In particular, for a constant function $F(\phi) = \mathbf{1}$ we find that

$$P_{T-t}^\delta \mathbf{1}(\phi) = E \exp \left(- \int_0^{T-t} \delta(r(s, \phi)) ds \right) = \exp \left(- \int_0^{T-t} \phi(s) ds \right) = B_T(t, \phi)$$

is the price of zero coupon if $r(t) = \phi$. It becomes obvious that in analogy to the finite dimensional case the problem of pricing interest rate dependent options is equivalent to the problem of calculating the semigroup P_t^δ for a sufficiently rich class of initial conditions F . One of the important questions in the theory of hedging is the differentiability of the price with respect to the initial yield curve. It is well known that the semigroup (P_t^δ) has poor smoothing properties and the function $\phi \rightarrow P_t^\delta F(\phi)$ need not be Fréchet differentiable for arbitrary F . However, we will show that for a large class of contingent claims containing most of the products which are traded the smoothing property takes place. In the sequel we assume for simplicity of presentation that the process r is time homogeneous, i.e., $\tau(t, \phi) = \tau(\phi)$. In view of Lemma 3.3 we use the notation $\int_0^t \delta(r(s, \phi)) ds$ instead of $\delta \left(\int_0^t r(s, \phi) ds \right)$. We will need an additional assumption.

(A) We assume $\alpha = 0$. Moreover, there exists $p \geq 0$ such that for every $t > 0$ and $a > 0$

$$\sup_{\|\phi\| \leq a} E \left(\exp \left(2p \|r(t, \phi)\| - 2 \int_0^t \delta(r(s, \phi)) ds \right) \right) < \infty.$$

If $r(t, \phi) \in H^1$ for every $t \geq 0$ and $\phi \in H^1$ then we will need a H^1 - version of (A):

(A') We assume $\alpha = 0$. Moreover, there exists $p \geq 0$ such that for every $t > 0$ and $a > 0$

$$\sup_{\|\phi\| \leq a} E \left(\exp \left(2p \|r(t, \phi)\|_1 - 2 \int_0^t \delta(r(s, \phi)) ds \right) \right) < \infty.$$

We will show that (A') holds if r is a Gaussian process. If the process r is non-negative then the results presented below are valid and the assumption (A) is not needed. In general the term $\exp\left(-\int_0^t \delta(r(s, \phi))ds\right)$ can grow exponentially.

Proposition 3.2 *If (A) holds for a certain $p \geq 0$ then putting $H = L^2(0, \infty)$, $P_t^\delta(C_p(H)) \subset C(H)$ and $P_t^\delta(C_p(H^1)) \subset C(H^1)$ for every $t \geq 0$.*

Proof We provide the proof for H^1 only. Let $F \in C_p(H^1)$ and let $(\phi_n) \subset H^1$ be a sequence converging in H^1 to ϕ . Then $F(\phi) = e^{-p\|\phi\|_1} G(\phi)$ with $G \in C_0(H^1)$ and

$$|P_t^\delta F(\phi)| \leq \|G\|_0 E \left(\exp \left(p \|r(t, \phi)\|_1 - \int_0^t \delta(r(u, \phi)) du \right) \right).$$

Hence in view of (A') $P_t^\delta F(\phi)$ is well defined. Moreover, (A') yields uniformly integrability of the family of random variables

$$\left\{ \exp \left(p \|r(t, \phi)\|_1 - \int_0^t \delta(r(u, \phi)) du \right) : \|\phi\| \leq a \right\}$$

for every $a > 0$. Hence the proposition follows from the continuity of F and Lemma (3.3). $\square$

Remark 3.3 The above theorem may be proved for any $\alpha \in \mathbb{R}$. However, the Kolmogorov equation we are going to study next is simpler in $L^2(0, \infty)$.

We shall identify the infinitesimal generator L of the Markov process r . Because the process r is not a semimartingale we can not apply the Itô formula to the function $F(r(t, \phi))$ even if $F \in C_p^2(H_\alpha)$. However, it turns out that the property (3.3) is sufficient for our needs. Let $\psi_1, \dots, \psi_n \in \text{dom}(A^*)$ and let P_n denote the orthogonal projection on the linear span H_n of the vectors $\psi_1, \dots, \psi_n$. First, let us define the space

$$\mathcal{D}_0 = \{F \in C_p(H_\alpha) : F = f \circ P_n, f \in C_p^2(\mathbb{R}^n), n = 1, \dots\}.$$

If $F \in \mathcal{D}_0$ then in view of (3.3) the process $F(r(t, \phi))$ is a semimartingale and

$$F(r(t, \phi)) = F(\phi) + \int_0^t L F(r(s, \phi)) ds + \int_0^t D F(r(s, \phi)) \tau(r(s, \phi)) dW(s), \quad (3.6)$$

where

$$L F(\phi) = \frac{1}{2} \langle D^2 F(\phi) \tau(\phi), \tau(\phi) \rangle + \langle \phi, A^* D F(\phi) \rangle + \langle G(\phi), D F(\phi) \rangle.$$

If $F \in \mathcal{D}_0$ then the function $A^* D F(\phi)$ is well-defined for all $\phi \in L^2(0, \infty)$ and therefore $L F(\phi)$ is a well-defined continuous function on $L^2(0, \infty)$. The above

considerations show that the generator of the Markov process r coincides on $\mathcal{D}_0$ with the operator L . Therefore we can expect that V_T as defined in (3.5) is a Feynman–Kac formula for the solution of the following equation

$$\begin{cases} \frac{\partial u}{\partial t}(t, \phi) + Lu(t, \phi) - \delta(\phi)u(t, \phi) = 0, \\ u(T, \phi) = F(\phi). \end{cases} \quad (3.7)$$

In other words the operator $L^\delta = L - \delta$ when considered on an appropriate domain is a generator of the semigroup P_t^δ . However, equation (3.7) is not valid in general because $V_T(t, \cdot)$ need not be differentiable.

Proposition 3.4 *Assume that τ and G are twice differentiable on H . Then for every $F \in C_p^2(H)$ the function V_T is a unique solution of the backward Kolmogorov equation (3.7) in the following sense.*

- The function $V_T : [0, \infty) \times H \rightarrow \mathbb{R}$ is bounded and continuous with respect to each variable.
- For every $t \geq 0$ we have $V_T(t, \cdot) \in C^2(H)$.
- We have $V_T \in C^1([0, T], H^1)$.
- Equation (3.7) holds for every $\phi \in \text{dom}(A)$ and $t \geq 0$. Moreover, V_T is given by (3.5).

Proof Let δ_n denote a sequence of C^2 functions on $\mathbb{R}$ such that $\langle \delta_n, \phi \rangle \rightarrow \delta(\phi)$ for every continuous ϕ and let $L_n = L - \delta_n$. If we denote by P_t^n the semigroup

$$P_t^n F(\phi) = E \left(\exp \left(- \int_0^t \langle \delta_n, r(u, \phi) \rangle du \right) F(r(t, \phi)) \right)$$

then by a simple modification of the proof of Theorem 9.17 in Da Prato and Zabczyk (1992) we can show, putting $u^n(t, \phi) = P_t^n F(\phi)$, that

$$\begin{cases} \frac{\partial u^n}{\partial t}(t, \phi) + Lu^n(t, \phi) - \langle \delta_n, \phi \rangle u^n(t, \phi) = 0, \\ u^n(T, \phi) = F(\phi), \end{cases} \quad (3.8)$$

and moreover u^n is a unique solution of (3.8). We shall show first that for every $\phi \in H$

$$\lim_{n \rightarrow \infty} P_t^n F(\phi) = P_t^\delta F(\phi). \quad (3.9)$$

Indeed,

$$\begin{aligned} |P_t^n F(\phi) - P_t^\delta F(\phi)| &\leq \|F\|_p E \left(e^{p\|r(t, \phi)\|} \left| \exp \left(- \int_0^t \langle \delta_n, r(u, \phi) \rangle du \right) \right. \right. \\ &\quad \left. \left. - \exp \left(- \int_0^t \delta(r(u, \phi)) du \right) \right| \right) \end{aligned}$$

and therefore (A) and the definition of δ_n yield (3.9). Using (3.9) and Theorem 9.16 in Da Prato and Zabczyk (1992) we obtain easily that the right-hand side of (3.8) converges (along the subsequence n_k) to the expression

$$LP_t^\delta F(\phi) - \delta(\phi)P_t^\delta F(\phi)$$

for every $\phi \in H_\alpha^1$ uniformly in $t \leq T$. Hence

$$\lim_{k \rightarrow \infty} \frac{\partial u^{n_k}}{\partial t}(t, \phi) = \frac{\partial P_t^\delta}{\partial t}(\phi)$$

and therefore $P_t^\delta F$ satisfies (3.7). $\square$

Unfortunately, this theorem has too strong assumptions to be applicable to some important contingent claims like swaptions. Stronger results can be obtained in the Gaussian case.

Proposition 3.5 *The mapping u is a solution of (3.7) if and only if $u(t, \phi) = B_T(t, \phi)R_T(t, \phi)$, $R_T(T, \phi) = F(\phi)$ and*

$$\begin{aligned} \frac{\partial R_T}{\partial t}(t, \phi) + \frac{1}{2} \left\langle D^2 R_T(t, \phi) \tau(\phi), \tau(\phi) \right\rangle + \langle DR_T(t, \phi), A\phi + G(\phi) \rangle \\ - \langle DR_T(t, \phi), \tau(\phi) \rangle \langle \tau(\phi), S(t)I_{[0, T]} \rangle = 0, \end{aligned} \quad (3.10)$$

where the solution is defined in the sense of Proposition 3.4.

Proof Let u satisfy (3.7) and define the function R_T by the formula $u(t, \phi) = B_T(t, \phi)R_T(t, \phi)$. Then R_T is smooth and

$$\frac{\partial u}{\partial t}(t, \phi) = \phi(T - t)B_T(t, \phi)R_T(t, \phi) + B_T(t, \phi)\frac{\partial R_T}{\partial t}(t, \phi), \quad (3.11)$$

$$Du(t, \phi) = -B_T(t, \phi)R_T(t, \phi)S(t)I_{[0, T]} + B_T(t, \phi)DR_T(t, \phi), \quad (3.12)$$

$$\begin{aligned} D^2 u(t, \phi) = B_T(t, \phi)R_T(t, \phi) \left(S(t)I_{[0, T]} \right) \otimes \left(S(t)I_{[0, T]} \right) \\ - 2B_T(t, \phi)DR_T(t, \phi) \otimes S(t)I_{[0, T]} + B_T(t, \phi)D^2 R_T(t, \phi). \end{aligned} \quad (3.13)$$

Hence by (3.12)

$$\begin{aligned} \langle Du(t, \phi), A\phi + G(\phi) \rangle = -B_T(t, \phi)R_T(t, \phi) \\ \left(\phi(T - t) - \phi(0) + \frac{1}{2} \langle S(t)I_{[0, T]}, \tau(\phi) \rangle^2 \right) \end{aligned} \quad (3.14)$$

and by (3.13)

$$\begin{aligned} \langle D^2 u(t, \phi) \tau(\phi), \tau(\phi) \rangle = B_T(t, \phi)R_T(t, \phi) \langle S(t)I_{[0, T]}, \tau(\phi) \rangle^2 \\ - 2B_T(t, \phi) \langle DR_T(t, \phi), \tau(\phi) \rangle \langle S(t)I_{[0, T]}, \tau(\phi) \rangle \\ + B_T(t, \phi) \langle D^2 R_T(t, \phi) \tau(\phi), \tau(\phi) \rangle. \end{aligned} \quad (3.15)$$

Finally, taking into account (3.11), (3.14) and (3.15) we find that

$$\begin{aligned} & \frac{\partial u}{\partial t}(t, \phi) + \frac{1}{2} \langle D^2 u(t, \phi) \tau(\phi), \tau(\phi) \rangle + \langle Du(t, \phi), A\phi + G(\phi) \rangle - \delta(\phi)u(t, \phi) \\ &= B_T(t, \phi) \left(\frac{\partial R_T}{\partial t}(t, \phi) + \frac{1}{2} \langle D^2 R_T(t, \phi) \tau(\phi), \tau(\phi) \rangle + \langle DR_T(t, \phi), A\phi + G(\phi) \rangle \right. \\ & \quad \left. - \langle DR_T(t, \phi), \tau(\phi), \tau(\phi) \rangle \langle S(t)I_{[0, T]}, \tau(\phi) \rangle \right) \end{aligned}$$

and (3.10) follows. Using similar arguments we show that if R_T satisfies (3.10) then $u(t, \phi) = B_T(t, \phi)R_T(t, \phi)$ is a solution to (3.7). $\square$

Remark 3.6 The proposition 3.5 describes the forward measure transformation performed at the level of the Kolmogorov equation. Note that equation (3.10) is the Kolmogorov equation for the process Y (say) defined as a solution to the stochastic differential equation

$$dY = (AY + G_\sigma(Y) - \langle \tau(Y), S(t)I_T \rangle \tau(Y)) dt + \tau(Y)dW$$

or in a more explicit form

$$\begin{aligned} dY(t, x) &= \left(\frac{\partial Y}{\partial x}(t, x) + \tau(Y(t))(x) \int_0^x \tau(Y(t))(u) du \right) dt \\ &\quad - \tau(Y(t))(x) \int_0^{T-t} \tau(Y(t))(u) du dt + \tau(Y(t))(x) dW(t). \end{aligned}$$

From this point on we assume that $\tau \in H$ is a constant vector and therefore

$$r(t) = S(t)\phi + \int_0^t S(s)G ds + \int_0^t S(t-s)\tau dW(s).$$

This case has been discussed in Musiela (1993) and Brace and Musiela (1994). For every $t \geq 0$ the random variable $r(t)$ is Gaussian with the mean

$$Er(t) = S(t)\phi + \int_0^t S(s)G ds$$

and the covariance operator

$$Q_t = \int_0^t S(s)\tau\tau^*S^*(s) ds.$$

Moreover, because $r(t, \phi)$ is Gaussian so is $R(t, \phi)(0)$. Hence, using the Hölder inequality we check by direct calculations that for $t \leq T$

$$E(\exp(2p\|r(t, \phi)\|_\alpha - 2R(t, \phi)(0))) \leq C_T \exp(\beta_T\|\phi\|)$$

for some constants $C_T, \beta_T > 0$. Therefore (A) holds. In the present framework equation (3.7) may be written in the form

$$\begin{cases} \frac{\partial u}{\partial t}(t, \phi) = \frac{1}{2} \langle D^2 u(t, \phi) \tau, \tau \rangle + \langle A\phi + G(\phi), Du(t, \phi) \rangle - \delta(\phi)u(t, \phi), \\ u(0, \phi) = F(\phi), \phi \in \text{dom}(A). \end{cases} \quad (3.16)$$

We shall need the finite dimensional parabolic PDE

$$\frac{\partial h}{\partial t}(t, x_1, \dots, x_n) + \frac{1}{2} \sum_{i,j=1}^n b_i^*(t)b_j(t)x_i x_j \frac{\partial^2 h}{\partial x_i \partial x_j}(t, x_1, \dots, x_n) = 0 \quad (3.17)$$

with the terminal condition $h(T, x_1, \dots, x_n) = h_0(x_1, \dots, x_n)$ and

$$b_i^*(t)b_j(t) = \int_{T-t}^{T_i-t} \tau^*(x) dx \int_{T-t}^{T_j-t} \tau(x) dx.$$

Equation (3.17) has a unique solution for every measurable terminal condition h_0 with linear growth. Let

$$F_{T,T_i}(t, \phi) = \exp(-\langle S(t)I_{T,T_i}, \phi \rangle),$$

where I_{T,T_i} is an indicator function of the interval $[T, T_i]$.

Theorem 3.7 *If the function $U(t, x_1, \dots, x_n)$ is a solution to (3.17) with the terminal condition $U_0(x_1, \dots, x_n)$ then the function*

$$u(t, \phi) = B_T(t, \phi)U(t, F_{T,T_1}(t, \phi), \dots, F_{T,T_n}(t, \phi))$$

is a solution to the Cauchy problem (3.6) with the terminal condition

$$u(T, \phi) = U_0(B_{T_1}(T, \phi), \dots, B_{T_n}(T, \phi)).$$

Proof It is enough to consider the case $n = d = 1$. The general argument is exactly the same. In view of Proposition 3.5 we need to show that the function

$$R(t, \phi) = U(t, F_{T,T_1}(t, \phi), \dots, F_{T,T_n}(t, \phi)) \quad (3.18)$$

is a solution to equation (3.10). Note first that

$$\frac{dF_{T,T_1}}{dt}(t, \phi) = (\phi(T_1 - t) - \phi(T - t)) F_{T,T_1}(t, \phi),$$

$$DF_{T,T_1}(t, \phi) = -F_{T,T_1}(t, \phi)l_t$$

with $l_t = I_{[T-t, T_1-t]}$ and

$$D^2 F_{T,T_1}(t, \phi) = F_{T,T_1}(t, \phi)l_t \otimes l_t.$$

Hence, denoting $l = I_{[0, T-t]}$ we find that for $\phi \in \text{dom}(A)$

$$\begin{aligned} \frac{\partial R}{\partial t}(t, \phi) &= \frac{\partial U}{\partial t}(t, F_{T, T_1}(t, \phi)) \\ &\quad + F_{T, T_1}(t, \phi)(\phi(T_1 - t) - \phi(T - t)) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \end{aligned} \quad (3.19)$$

and

$$DR(t, \phi) = -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) l_t.$$

Hence

$$\langle DR(t, \phi), \tau \rangle = -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \langle l_t, \tau \rangle \quad (3.20)$$

and

$$\begin{aligned} \langle DR(t, \phi), A\phi + G_\sigma \rangle &= -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \langle l_t, A\phi + G_\sigma \rangle \\ &= -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) (\phi(T_1 - t) - \phi(T - t)) \\ &\quad - F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \int_{T-t}^{T_1-t} \frac{1}{2} \frac{d}{dx} \left(\int_0^x \tau(u) du \right)^2 dx \\ &= -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) (\phi(T_1 - t) - \phi(T - t)) \\ &\quad - \frac{1}{2} F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \left(\left(\int_0^{T_1-t} \tau(u) du \right)^2 - \left(\int_0^{T-t} \tau(u) du \right)^2 \right). \end{aligned}$$

Thereby

$$\begin{aligned} \langle DR(t, \phi), A\phi + G \rangle &= -F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) (\phi(T_1 - t) - \phi(T - t)) \\ &\quad - \frac{1}{2} F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \langle \tau, l \rangle^2 \\ &\quad - F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \langle \tau, l \rangle \langle \tau, l_t \rangle. \end{aligned} \quad (3.21)$$

Next

$$\begin{aligned} D^2 R(t, \phi) &= F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) l_t \otimes l_t \\ &\quad + F_{T, T_1}^2(t, \phi) \frac{\partial^2 U}{\partial x^2}(t, F_T(t, \phi)) l_t \otimes l_t. \end{aligned}$$

Hence

$$\begin{aligned} \langle D^2 R(t, \phi) \tau, \tau \rangle &= F_{T, T_1}(t, \phi) \frac{\partial U}{\partial x}(t, F_{T, T_1}(t, \phi)) \langle l_t, \tau \rangle^2 \\ &+ F_{T, T_1}^2(t, \phi) \frac{\partial^2 U}{\partial x^2}(t, F_{T, T_1}(t, \phi)) \langle l_t, \tau \rangle^2. \end{aligned} \quad (3.22)$$

Now, taking into account (3.19), (3.20), (3.21) and (3.22) we find that

$$\begin{aligned} \frac{\partial R}{\partial t}(t, \phi) + \frac{1}{2} \langle D^2 R_T(t, \phi) \tau(\phi), \tau(\phi) \rangle + \langle DR_T(t, \phi), A\phi + G_\sigma(\phi) \rangle \\ - \langle DR_T(t, \phi), \tau(\phi) \rangle \langle \tau(\phi), S(t)I_T \rangle \\ = \frac{\partial U}{\partial t}(t, F_{T, T_1}(t, \phi)) + \frac{1}{2} F_{T, T_1}^2(t, \phi) \frac{\partial^2 U}{\partial x^2}(t, F_{T, T_1}(t, \phi)) \langle l_t, \tau \rangle^2, \end{aligned}$$

where $R(t, \phi)$ is defined by (3.18). Therefore, by (3.17) the function R satisfies equation (3.10) and the theorem follows. $\square$

References

- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities, *J. Political Economy* **81** 637–59
- Brace, A., Gatarek, D. and Musiela, M. (1997), The market model of interest rate dynamics, *Math. Finance* **7** 127–54
- Brace, A. and Musiela, M. (1994), A multifactor Gauss–Markov implementation of Heath, Jarrow and Morton, *Mat. Finance* **2** 259–83
- Da Prato, G. and Zabczyk, J. (1992), *Stochastic equations in infinite dimensions*, Cambridge University Press
- Goldys, B., Musiela, M. and Sondermann, D. (1995), Lognormality of rates and term structure models, preprint, UNSW
- Gatarek, D. and Święch, A. (1997), Optimal stopping in Hilbert spaces and pricing of American options, a preprint
- Hamza, K. and Klebaner, F.C. (1995), A stochastic partial differential equation for term structure of interest rates, a preprint
- Harrison, J.M. and Pliska, S.R. (1981), Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Process. Appl.* **11** 215–60
- Heath, D. Jarrow, R. and Morton, A. (1992), Bond pricing and the term structure of interest rates: a new methodology, *Econometrica* **61**(1) 77–105
- Kennedy, P.D. (1994), The term structure of interest rates as a Gaussian Markov field, *Math. Finance* **4** 247–58
- Musiela, M. (1993), Stochastic PDEs and term structure models, *Journées Internationales de Finance*, IGR-AFFI, La Baule
- Santa-Clara, P. and Sornette, D. (1997), The dynamics of the forward interest rate curve with stochastic string shocks, preprint, UCLA

Modelling of Forward Libor and Swap Rates

Marek Rutkowski

1 Introduction

The last decade was marked by a rapidly growing interest in the arbitrage-free modelling of bond market. Undoubtedly, one of the major achievements in this area was a new approach to the term structure modelling proposed by Heath, Jarrow and Morton in their work published in 1992, commonly known as the HJM methodology. One of its main features is that it covers a large variety of previously proposed models and provides a unified approach to the modelling of instantaneous interest rates and to the valuation of interest-rate sensitive derivatives. Let us give a very concise description of the HJM approach (for a detailed account we refer, for instance, to Chapter 13 in Musiela and Rutkowski (1997a)).

The HJM methodology is based on an exogenous specification of the dynamics of instantaneous, continuously compounded forward rates $f(t, T)$. For any fixed maturity $T \leq T^*$, the dynamics of the forward rate $f(t, T)$ are

$$df(t, T) = \alpha(t, T) dt + \sigma(t, T) \cdot dW_t,$$

where α and σ are adapted stochastic processes with values in $\mathbb{R}$ and $\mathbb{R}^d$, respectively, and W is a d -dimensional standard Brownian motion with respect to the underlying probability measure $\mathbb{P}$ which plays the role of the real-world probability. More formally, for every fixed $T \leq T^*$, where $T^* > 0$ is the horizon date, we have

$$f(t, T) = f(0, T) + \int_0^t \alpha(u, T) du + \int_0^t \sigma(u, T) \cdot dW_u$$

for some Borel-measurable function $f(0, \cdot) : [0, T^*] \rightarrow \mathbb{R}$ and stochastic processes applications $\alpha(\cdot, T)$ and $\sigma(\cdot, T)$. Let us notice that, for any fixed maturity date $T \leq T^*$, the initial condition $f(0, T)$ is determined by the current value of the continuously compounded forward rate for the future date T which prevails at time 0. In practical terms, the function $f(0, T)$ is determined by the current yield curve,

which can be estimated on the basis of observed market prices of bonds (and other relevant instruments).

Let us denote by $B(t, T)$ the price at time $t \leq T$ of a unit zero-coupon bond which matures at the date $T \leq T^*$. In the present setup the price $B(t, T)$ can be recovered from the formula

$$B(t, T) = \exp\left(-\int_t^T f(t, u) du\right).$$

The problem of the absence of arbitrage opportunities in the bond market can be formulated in terms of the existence of a suitably defined martingale measure. It appears that in an arbitrage-free setting – that is, under the martingale measure – the drift coefficient α in the dynamics of the instantaneous forward rate is uniquely determined by the volatility coefficient σ , and a stochastic process which can be interpreted as the market price of the interest-rate risk. If we denote by $\mathbb{P}^*$ the martingale measure for the bond market, and by W^* the associated standard Brownian motion, then

$$dB(t, T) = B(t, T)(r_t dt + b(t, T) \cdot dW_t^*),$$

where $r_t = f(t, t)$ is the short-term interest rate, and the bond price volatility $b(t, T)$ satisfies

$$b(t, T) = -\int_t^T \sigma(t, u) du. \quad (1.1)$$

Furthermore, it appears that in the special case when the coefficient σ follows a deterministic function, the valuation formulae for interest rate-sensitive derivatives are independent of the choice of the risk premium. In this sense, the choice of a particular model from the broad class of HJM models hinges uniquely on the specification of the volatility coefficient σ .

The HJM methodology appeared to be very successful both from the theoretical and practical viewpoints. Since the HJM approach to the term structure modelling is based on an arbitrage-free dynamics of the instantaneous continuously compounded forward rates, it requires a certain degree of smoothness with respect to the tenor of the bond prices and their volatilities. For this reason, working with such models is not always convenient.

An alternative construction of an arbitrage-free family of bond prices, making no reference to the instantaneous rates, is in some circumstances more suitable. The first step in this direction was done by Sandmann and Sondermann (1993), who focused on the effective annual interest rate. This approach was further developed in ground-breaking papers by Miltersen *et al.* (1997) and Brace *et al.* (1997), who proposed to model instead the family of forward Libor rates. The main goal was to produce an arbitrage-free term structure model which would support the common

practice of pricing such interest-rate derivatives as caps and swaptions through a suitable version of Black's formula. This practical requirement enforces the lognormality of the forward Libor (or swap) rate under the corresponding forward martingale measure.

It is interesting to notice that Brace *et al.* (1997) parametrize their version of the lognormal forward Libor model introduced by Miltersen *et al.* (1997) with a piecewise constant volatility function. They need to consider smooth volatility functions in order to analyse the model in the HJM framework, however. The backward induction approach to the modelling of forward Libor and swap rate developed in Musiela and Rutkowski (1997a) and Jamshidian (1997) overcomes this technical difficulty. In addition, in contrast to the previous papers, it allows also for the modelling of forward Libor (and swap) rates associated with accrual periods of differing lengths.

It should be stressed that a similar (but not identical) approach to the modelling of market rate was developed in a series of papers by Hunt *et al.* (1996, 2000) and Hunt and Kennedy (1996, 1997). Since special emphasis is put here on the existence of the underlying low-dimensional Markov process that governs directly the dynamics of interest rates, this alternative approach is termed the *Markov-functional approach*. This property leads to a considerable simplification in numerical procedures associated with the model's implementation. Another important feature of this approach is its ability of providing a perfect fit to market prices of a given family of interest-rate options.

2 Modelling of forward Libor rates

In this section, we present various approaches to the modelling of forward Libor rates. We focus here on the model's construction, its basic properties, and the valuation of the most typical derivatives. For further details, the interested reader is referred to the original papers: Musiela and Sondermann (1993), Sandmann and Sondermann (1993), Goldys *et al.* (1994), Sandmann *et al.* (1995), Brace *et al.* (1997), Jamshidian (1997), Miltersen *et al.* (1997), Musiela and Rutkowski (1997b), Rady (1997), Sandmann and Sondermann (1997), Rutkowski (1998, 1999), Glasserman and Kou (1999), and Yasuoka (1999). The issues related to the model's implementation are extensively treated in Brace (1996), Andersen and Andreasen (1997), Sidenius (1997), Brace *et al.* (1998), Musiela and Sawa (1998), Hull and White (1999), Schlögl (1999), Uratani and Utsunomiya (1999), Yasuoka (1999), Lotz and Schlögl (2000), Glasserman and Zhao (2000), Brace and Womersley (2000), and Dun *et al.* (2000).

2.1 Forward and futures Libor rates

Our first task is to examine those properties of forward and futures contracts related to the notion of the Libor rate which are universal; that is, which do not rely on specific assumptions imposed on a particular model of the term structure of interest rates. To this end, we fix an index j , and we consider various interest-rate sensitive derivatives related to the period $[T_j, T_{j+1}]$. To be more specific, we shall focus in this section on single-period forward swaps – that is, forward rate agreements.

We need to introduce some notation. We assume that we are given a prespecified collection of reset/settlement dates $0 < T_0 < T_1 < \dots < T_n = T^*$, referred to as the *tenor structure*. Also, we denote $\delta_j = T_j - T_{j-1}$ for $j = 1, \dots, n$. We write $B(t, T_j)$ to denote the price at time t of a T_j -maturity zero-coupon bond. $\mathbb{P}^*$ is the *spot martingale measure*, while for any $j = 0, \dots, n$ we write $\mathbb{P}_{T_j}$ to denote the *forward martingale measure* associated with the date T_j . The corresponding d -dimensional Brownian motions are denoted by W^* and W^{T_j} , respectively. Also, we write $F_B(t, T, U) = B(t, T)/B(t, U)$ so that

$$F_B(t, T_{j+1}, T_j) = \frac{B(t, T_{j+1})}{B(t, T_j)}, \quad \forall t \in [0, T_j],$$

is the forward price at time t of the T_{j+1} -maturity zero-coupon bond for the settlement date T_j . We use the symbol $\pi_t(X)$ to denote the value (i.e., the arbitrage price) at time t of a European contingent claim X . Finally, we shall use the letter $\mathcal{E}$ for the Doléans exponential, for instance,

$$\mathcal{E}_t \left(\int_0^\cdot \gamma_u \cdot dW_u^* \right) = \exp \left(\int_0^t \gamma_u \cdot dW_u^* - \frac{1}{2} \int_0^t |\gamma_u|^2 du \right),$$

where the dot ‘ $\cdot$ ’ and $|\cdot|$ stand for the inner product and Euclidean norm in $\mathbb{R}^d$, respectively.

2.1.1 Single-period swaps settled in arrears

Let us first consider a single-period swap agreement settled in arrears; i.e., with the *reset date* T_j and the *settlement date* T_{j+1} (multi-period interest rate swaps are examined in Section 3). By the contractual features, the long party pays $\delta_{j+1}\kappa$ and receives $B^{-1}(T_j, T_{j+1}) - 1$ at time T_{j+1} . Equivalently, he pays an amount $Y_1 = 1 + \delta_{j+1}\kappa$ and receives $Y_2 = B^{-1}(T_j, T_{j+1})$ at this date. The values at time $t \leq T_j$ of these payoffs are

$$\pi_t(Y_1) = B(t, T_{j+1})(1 + \delta_{j+1}\kappa), \quad \pi_t(Y_2) = B(t, T_j).$$

The second equality above is trivial, since the payoff Y_2 is equivalent to the unit payoff at time T_j . Consequently, for any fixed $t \leq T_j$, the value of the *forward*

swap rate, which makes the contract worthless at time t , can be found by solving for $\kappa = \kappa(t, T_j, T_{j+1})$ the following equation:

$$\pi_t(Y_2) - \pi_t(Y_1) = B(t, T_j) - B(t, T_{j+1})(1 + \delta_{j+1}\kappa) = 0.$$

It is thus apparent that

$$\kappa(t, T_j, T_{j+1}) = \frac{B(t, T_j) - B(t, T_{j+1})}{\delta_{j+1}B(t, T_{j+1})}, \quad \forall t \in [0, T_j].$$

Note that the forward swap rate $\kappa(t, T_j, T_{j+1})$ coincides with the *forward Libor rate* $L(t, T_j)$ which, by the market convention, is set to satisfy

$$1 + \delta_{j+1}L(t, T_j) = \frac{B(t, T_j)}{B(t, T_{j+1})} = \mathbb{E}_{\mathbb{P}_{T_{j+1}}}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) \quad (2.1)$$

for every $t \in [0, T_j]$. Let us notice that the last equality is a consequence of the definition of the forward measure $\mathbb{P}_{T_{j+1}}$. We conclude that in order to determine the forward Libor rate $L(\cdot, T_j)$, it is enough to find the forward price $F_X(t, T_{j+1})$ at time t of the contingent claim $X = B^{-1}(T_j, T_{j+1})$ in the forward contract that settles at time T_{j+1} . Indeed, it is well known (see, for instance, Musiela and Rutkowski (1997a)) that

$$F_X(t, T_{j+1}) = B(t, T_{j+1}) \mathbb{E}_{\mathbb{P}_{T_{j+1}}}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t).$$

Furthermore, it is evident that the process $L(\cdot, T_j)$ follows necessarily a martingale under the forward probability measure $\mathbb{P}_{T_{j+1}}$. Recall that in the Heath–Jarrow–Morton framework, we have, under $\mathbb{P}_{T_{j+1}}$,

$$dF_B(t, T_j, T_{j+1}) = F_B(t, T_j, T_{j+1})(b(t, T_j) - b(t, T_{j+1})) \cdot dW_t^{T_{j+1}}, \quad (2.2)$$

where, for each maturity date T , the process $b(\cdot, T)$ represents the price volatility of the T -maturity zero-coupon bond. On the other hand, if the process $L(\cdot, T_j)$ is strictly positive, it can be shown to admit the following representation¹

$$dL(t, T_j) = L(t, T_j)\lambda(t, T_j) \cdot dW_t^{T_{j+1}},$$

where $\lambda(\cdot, T_j)$ is an adapted stochastic process which satisfies mild integrability conditions. Combining the last two formulae with (2.1), we arrive at the following fundamental relationship, which plays an essential role in the construction of the lognormal model of forward Libor rates,

$$\frac{\delta_{j+1}L(t, T_j)}{1 + \delta_{j+1}L(t, T_j)} \lambda(t, T_j) = b(t, T_j) - b(t, T_{j+1}), \quad \forall t \in [0, T_j]. \quad (2.3)$$

¹ This representation is a consequence of the martingale representation property of the standard Brownian motion.

For instance, in the construction which is based on the backward induction, relationship (2.3) will allow us to determine the forward measure for the date T_j , provided that $\mathbb{P}_{T_{j+1}}$, $W^{T_{j+1}}$ and the volatility $\lambda(t, T_j)$ of the forward Libor rate $L(\cdot, T_{j+1})$ are known. (One may assume, for instance, that $\lambda(\cdot, T_j)$ is a prespecified deterministic function.) Recall that in the Heath–Jarrow–Morton framework² the Radon–Nikodým density of $\mathbb{P}_{T_j}$ with respect to $\mathbb{P}_{T_{j+1}}$ is known to satisfy

$$\frac{d\mathbb{P}_{T_j}}{d\mathbb{P}_{T_{j+1}}} = \mathcal{E}_{T_j} \left(\int_0^\cdot (b(t, T_j) - b(t, T_{j+1})) \cdot dW_t^{T_{j+1}} \right). \quad (2.4)$$

In view of (2.3), we thus have

$$\frac{d\mathbb{P}_{T_j}}{d\mathbb{P}_{T_{j+1}}} = \mathcal{E}_{T_j} \left(\int_0^\cdot \frac{\delta_{j+1} L(t, T_j)}{1 + \delta_{j+1} L(t, T_j)} \lambda(t, T_j) \cdot dW_t^{T_{j+1}} \right).$$

For our further purposes, it is also useful to observe that this density admits the following representation

$$\frac{d\mathbb{P}_{T_j}}{d\mathbb{P}_{T_{j+1}}} = c F_B(T_j, T_j, T_{j+1}) = c(1 + \delta_{j+1} L(T_j, T_j)), \quad \mathbb{P}_{T_{j+1}}\text{-a.s.}, \quad (2.5)$$

where $c > 0$ is the normalizing constant, and thus

$$\frac{d\mathbb{P}_{T_j}}{d\mathbb{P}_{T_{j+1}} | \mathcal{F}_t} = c F_B(t, T_j, T_{j+1}) = c(1 + \delta_{j+1} L(t, T_j)), \quad \mathbb{P}_{T_{j+1}}\text{-a.s.}$$

Finally, the dynamics of the process $L(\cdot, T_j)$ under the probability measure $\mathbb{P}_{T_j}$ are given by a somewhat involved stochastic differential equation

$$dL(t, T_j) = L(t, T_j) \left(\frac{\delta_{j+1} L(t, T_j) |\lambda(t, T_j)|^2}{1 + \delta_{j+1} L(t, T_j)} dt + \lambda(t, T_j) \cdot dW_t^{T_j} \right).$$

As we shall see in what follows, it is nevertheless not hard to determine the probability law of $L(\cdot, T_j)$ under the forward measure $\mathbb{P}_{T_j}$ – at least in the case of the deterministic volatility $\lambda(\cdot, T_j)$ of the forward Libor rate.

2.1.2 Single-period swaps settled in advance

Consider now a similar swap which is, however, settled in advance – that is, at time T_j . Our first goal is to determine the forward swap rate implied by such a contract. Note that under the present assumptions, the long party (formally) pays an amount $Y_1 = 1 + \delta_{j+1}\kappa$ and receives $Y_2 = B^{-1}(T_j, T_{j+1})$ at the settlement date T_j (which coincides here with the reset date). The values at time $t \leq T_j$ of these payoffs admit the following representations

$$\pi_t(Y_1) = B(t, T_j)(1 + \delta_{j+1}\kappa), \quad \pi_t(Y_2) = B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t).$$

² See Heath *et al.* (1992) or Chapter 13 in Musiela and Rutkowski (1997a).

The value $\kappa = \hat{\kappa}(t, T_j, T_{j+1})$ of the modified forward swap rate, which makes the swap agreement settled in advance worthless at time t , can be found from the equality

$$\pi_t(Y_2) - \pi_t(Y_1) = B(t, T_j) \left(\mathbb{E}_{\mathbb{P}_{T_j}} (B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) - (1 + \delta_{j+1}\kappa) \right) = 0.$$

It is clear that

$$\hat{\kappa}(t, T_j, T_{j+1}) = \delta_{j+1}^{-1} \left(\mathbb{E}_{\mathbb{P}_{T_j}} (B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) - 1 \right).$$

We are in a position to introduce the *modified forward Libor rate* $\tilde{L}(t, T_j)$ by setting, for every $t \in [0, T_j]$,

$$\tilde{L}(t, T_j) := \delta_{j+1}^{-1} \left(\mathbb{E}_{\mathbb{P}_{T_j}} (B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) - 1 \right).$$

Let us make two remarks. First, it is clear that finding of the modified forward Libor rate $\tilde{L}(\cdot, T_j)$ is formally equivalent to finding the forward price of the claim $B^{-1}(T_j, T_{j+1})$ for the settlement date T_j .³ Second, it is useful to observe that

$$\tilde{L}(t, T_j) = \mathbb{E}_{\mathbb{P}_{T_j}} \left(\frac{1 - B(T_j, T_{j+1})}{\delta_{j+1} B(T_j, T_{j+1})} \middle| \mathcal{F}_t \right) = \mathbb{E}_{\mathbb{P}_{T_j}} (L(T_j, T_j) | \mathcal{F}_t). \quad (2.6)$$

In particular, it is evident that at the reset date T_j the two kinds of forward Libor rates introduced above coincide, since manifestly

$$\tilde{L}(T_j, T_j) = \frac{1 - B(T_j, T_{j+1})}{\delta_{j+1} B(T_j, T_{j+1})} = L(T_j, T_j).$$

To summarize, the “standard” forward Libor rate $L(\cdot, T_j)$ satisfies

$$L(t, T_j) = \mathbb{E}_{\mathbb{P}_{T_{j+1}}} (L(T_j, T_j) | \mathcal{F}_t), \quad \forall t \in [0, T_j],$$

with the initial condition

$$L(0, T_j) = \frac{B(0, T_j) - B(0, T_{j+1})}{\delta_{j+1} B(0, T_{j+1})}.$$

On the other hand, for the modified Libor rate $\tilde{L}(\cdot, T_j)$ we have

$$\tilde{L}(t, T_j) = \mathbb{E}_{\mathbb{P}_{T_j}} (\tilde{L}(T_j, T_j) | \mathcal{F}_t), \quad \forall t \in [0, T_j],$$

with the initial condition

$$\tilde{L}(0, T_j) = \delta_{j+1}^{-1} \left(\mathbb{E}_{\mathbb{P}_{T_j}} (B^{-1}(T_j, T_{j+1})) - 1 \right).$$

The calculation of the right-hand side above involve not only on the initial term structure, but also the volatilities of bond prices (for more details, we refer to Rutkowski (1998)).

³ Recall that in the case of a forward Libor rate, the settlement date was T_{j+1} .

2.1.3 Eurodollar futures contracts

The next object of our studies is the futures Libor rate. A *Eurodollar futures contract* is a futures contract in which the Libor rate plays the role of an underlying asset. By convention, at the contract's maturity date T_j , the quoted Eurodollar futures price, denoted by $E(T_j, T_j)$, is set to satisfy

$$E(T_j, T_j) := 1 - \delta_{j+1}L(T_j, T_j).$$

Equivalently, in terms of the zero-coupon bond price we have $E(T_j, T_j) = 2 - B^{-1}(T_j, T_{j+1})$. From the general theory, it follows that the Eurodollar futures price at time $t \leq T_j$ equals

$$E(t, T_j) := \mathbb{E}_{\mathbb{P}^*}(E(T_j, T_j)) = 2 - \mathbb{E}_{\mathbb{P}^*}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) \quad (2.7)$$

(recall that $\mathbb{P}^*$ represents the spot martingale measure in a given model of the term structure). It is thus natural to introduce the concept of the *futures Libor rate*, associated with the Eurodollar futures contract, through the following definition.

Definition 2.1 Let $E(t, T_j)$ be the Eurodollar futures price at time t for the settlement date T_j . The implied *futures Libor rate* $L^f(t, T_j)$ satisfies

$$E(t, T_j) = 1 - \delta_{j+1}L^f(t, T_j), \quad \forall t \in [0, T_j]. \quad (2.8)$$

It follows immediately from (2.7)–(2.8) that the following equality is valid:

$$1 + \delta_{j+1}L^f(t, T_j) = \mathbb{E}_{\mathbb{P}^*}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t). \quad (2.9)$$

Equivalently, we have

$$L^f(t, T_j) = \mathbb{E}_{\mathbb{P}^*}(L(T_j, T_j) | \mathcal{F}_t) = \mathbb{E}_{\mathbb{P}^*}(\tilde{L}(T_j, T_j) | \mathcal{F}_t).$$

Note that in any term structure model, the futures Libor rate necessarily follows a martingale under the spot martingale measure $\mathbb{P}^*$ (provided, of course, that $\mathbb{P}^*$ is well-defined in this model).

2.2 Lognormal models of forward Libor rates

We shall now describe alternative approaches to the modelling of forward Libor rates in a continuous- and discrete-tenor setups.

2.2.1 The Miltersen–Sandmann–Sondermann approach

The first attempt to provide a rigorous construction of a lognormal model of forward Libor rates was done by Miltersen *et al.* (1997). The interested reader is referred also to Musiela and Sondermann (1993), Goldys *et al.* (1994), and Sandmann *et al.* (1995) for related previous studies. As a starting point in their

approach, Miltersen *et al.* (1997) postulate that the forward Libor rates process $L(\cdot, T)$ satisfies

$$dL(t, T) = \mu(t, T) dt + L(t, T)\lambda(t, T) \cdot dW_t^*,$$

with a deterministic volatility function $\lambda(\cdot, T) : [0, T] \rightarrow \mathbb{R}^d$. It is not difficult to deduce from the last formula that the forward price of a zero-coupon bond satisfies

$$dF(t, T + \delta, T) = -F(t, T + \delta, T)(1 - F(t, T + \delta, T))\lambda(t, T) \cdot dW_t^T.$$

Subsequently, they focus on the partial differential equation satisfied by the function $v = v(t, x)$, which expresses the forward price of the bond option in terms of the forward bond price. It is interesting to note that the PDE (2.10) was previously solved by Rady and Sandmann (1994) who worked within a different framework, however.⁴ The PDE for the option's price is

$$\frac{\partial v}{\partial t} + \frac{1}{2}|\lambda(t, T)|^2 x^2 (1 - x)^2 \frac{\partial^2 v}{\partial x^2} = 0 \quad (2.10)$$

with the terminal condition $v(T, x) = (K - x)^+$. As a result, Miltersen *et al.* (1997) obtained not only the closed-form solution for the price of a bond option (this was already achieved in Rady and Sandmann (1994)), but also the “market formula” for the caplet's price. The rigorous approach to the problem of existence of such a model was presented by Brace *et al.* (1997), who also worked within the continuous-time Heath–Jarrow–Morton framework.

2.2.2 Brace–Gątarek–Musiela approach

To formally introduce the notion of a *forward Libor rate*, we assume that we are given a family $B(t, T)$ of bond prices, and thus also the collection $F_B(t, T, U)$ of forward processes. In contrast to the previous section, we shall now assume that a strictly positive real number $\delta < T^*$, which represents the length of the accrual period, is fixed throughout. By definition, the forward δ -Libor rate $L(t, T)$ for the future date $T \leq T^* - \delta$ prevailing at time t is given by the conventional market formula

$$1 + \delta L(t, T) = F_B(t, T, T + \delta), \quad \forall t \in [0, T]. \quad (2.11)$$

The forward Libor rate $L(t, T)$ represents the add-on rate prevailing at time t over the future time interval $[T, T + \delta]$. We can also re-express $L(t, T)$ directly in terms of bond prices, as for any $T \in [0, T^* - \delta]$, we have

$$1 + \delta L(t, T) = \frac{B(t, T)}{B(t, T + \delta)}, \quad \forall t \in [0, T]. \quad (2.12)$$

⁴ In fact, they were concerned with the valuation of options on zero-coupon bonds for the term structure model put forward by Bühler and Käsler (1989).

In particular, the initial term structure of forward Libor rates satisfies

$$L(0, T) = \delta^{-1} \left(\frac{B(0, T)}{B(0, T + \delta)} - 1 \right). \quad (2.13)$$

Given a family $F_B(t, T, T^*)$ of forward processes, it is not hard to derive the dynamics of the associated family of forward Libor rates. For instance, one finds that under the forward measure $\mathbb{P}_{T+\delta}$, we have

$$dL(t, T) = \delta^{-1} F_B(t, T, T + \delta) \gamma(t, T, T + \delta) \cdot dW_t^{T+\delta},$$

where $\mathbb{P}_{T+\delta}$ is the forward measure for the date $T + \delta$, and the associated Wiener process $W^{T+\delta}$ equals

$$W_t^{T+\delta} = W_t^* - \int_0^t b(u, T + \delta) du, \quad \forall t \in [0, T + \delta].$$

Put another way, the process $L(\cdot, T)$ solves the equation

$$dL(t, T) = \delta^{-1} (1 + \delta L(t, T)) \gamma(t, T, T + \delta) \cdot dW_t^{T+\delta}, \quad (2.14)$$

subject to the initial condition (2.13). Suppose that forward Libor rates $L(t, T)$ are strictly positive. Then formula (2.14) can be rewritten as follows:

$$dL(t, T) = L(t, T) \lambda(t, T) \cdot dW_t^{T+\delta}, \quad (2.15)$$

where for any $t \in [0, T]$

$$\lambda(t, T) = \frac{1 + \delta L(t, T)}{\delta L(t, T)} \gamma(t, T, T + \delta). \quad (2.16)$$

This shows that the collection of forward processes uniquely specifies the family of forward Libor rates. The construction of a model of forward Libor rates relies on the following assumptions.

- (LR.1)** For any maturity $T \leq T^* - \delta$, we are given a $\mathbb{R}^d$ -valued, bounded deterministic function⁵ $\lambda(\cdot, T)$, which represents the volatility of the forward Libor rate process $L(\cdot, T)$.
- (LR.2)** We assume a strictly decreasing and strictly positive initial term structure $B(0, T)$, $T \in [0, T^*]$. The associated initial term structure $L(0, T)$ of forward Libor rates satisfies, for every $T \in [0, T^* - \delta]$,

$$L(0, T) = \frac{B(0, T) - B(0, T + \delta)}{\delta B(0, T + \delta)}. \quad (2.17)$$

⁵ Volatility λ could well follow an adapted stochastic process; we deliberately focus here on a *lognormal model* of forward Libor rates in which λ is deterministic.

To construct a model satisfying (LR.1)–(LR.2), Brace *et al.* (1997) place themselves in the Heath–Jarrow–Morton setup and they assume that for every $T \in [0, T^*]$, the volatility $b(t, T)$ vanishes for every $t \in [(T - \delta) \vee 0, T]$. In essence, the construction elaborated in Brace *et al.* (1997) is based on the forward induction, as opposed to the backward induction which we shall use in the next section. They start by postulating that the dynamics of $L(t, T)$ under the spot martingale measure $\mathbb{P}^*$ are governed by the following SDE:

$$dL(t, T) = \mu(t, T) dt + L(t, T)\lambda(t, T) \cdot dW_t^*,$$

where λ is a deterministic function, and the drift coefficient μ is unspecified. Recall that the arbitrage-free dynamics of the instantaneous forward rate $f(t, T)$ are

$$df(t, T) = \sigma(t, T) \cdot \sigma^*(t, T) dt + \sigma(t, T) \cdot dW_t^*,$$

where $\sigma^*(t, T) = \int_t^T \sigma(t, u) du = -b(t, T)$. On the other hand, the relationship (cf. (2.12))

$$1 + \delta L(t, T) = \exp\left(\int_T^{T+\delta} f(t, u) du\right) \quad (2.18)$$

is valid. Applying Itô's formula to both sides of (2.18), and comparing the diffusion terms, we find that

$$\sigma^*(t, T + \delta) - \sigma^*(t, T) = \int_T^{T+\delta} \sigma(t, u) du = \frac{\delta L(t, T)}{1 + \delta L(t, T)} \lambda(t, T).$$

To solve the last equation for σ^* in terms of L , it is necessary to impose some sort of initial condition on σ^* . For instance, by setting $\sigma(t, T) = 0$ for $0 \leq t \leq T \leq t + \delta$, we obtain the following relationship:

$$b(t, T) = -\sigma^*(t, T) = - \sum_{k=1}^{[\delta^{-1}(T-t)]} \frac{\delta L(t, T - k\delta)}{1 + \delta L(t, T - k\delta)} \lambda(t, T - k\delta). \quad (2.19)$$

The existence and uniqueness of solutions to SDEs which govern the instantaneous forward rate $f(t, T)$ and the forward Libor rate $L(t, T)$ for σ^* given by (2.19) can be shown using forward induction. Taking this result for granted, we conclude that $L(t, T)$ satisfies, under the spot martingale measure $\mathbb{P}^*$,

$$dL(t, T) = L(t, T)\sigma^*(t, T + \delta) \cdot \lambda(t, T) dt + L(t, T)\lambda(t, T) \cdot dW_t^*.$$

In this way, Brace *et al.* (1997) are able to completely specify their model of forward Libor rates.

2.2.3 Musiela–Rutkowski approach

In this section, we describe an alternative approach to the modelling of forward Libor rates; the construction presented below is a slight modification of that given by Musiela and Rutkowski (1997b). Let us start by introducing some notation. We assume that we are given a prespecified collection of reset/settlement dates $0 < T_0 < T_1 < \dots < T_n = T^*$, referred to as the *tenor structure* (by convention, $T_{-1} = 0$). Let us denote $\delta_j = T_j - T_{j-1}$ for $j = 0, \dots, n$. Then obviously $T_j = \sum_{i=0}^j \delta_i$ for every $j = 0, \dots, n$. We find it convenient to denote, for $m = 0, \dots, n$,

$$T_m^* = T^* - \sum_{j=n-m+1}^n \delta_j = T_{n-m}.$$

For any $j = 0, \dots, n-1$, we define the forward Libor rate $L(\cdot, T_j)$ by setting

$$L(t, T_j) = \frac{B(t, T_j) - B(t, T_{j+1})}{\delta_{j+1} B(t, T_{j+1})}, \quad \forall t \in [0, T_j].$$

Definition 2.2 For any $j = 0, \dots, n$, a probability measure $\mathbb{P}_{T_j}$ on $(\Omega, \mathcal{F}_{T_j})$, equivalent to $\mathbb{P}$, is said to be the *forward Libor measure* for the date T_j if, for every $k = 0, \dots, n$ the relative bond price

$$U_{n-j+1}(t, T_k) := \frac{B(t, T_k)}{\delta_j B(t, T_j)}, \quad \forall t \in [0, T_k \wedge T_j],$$

follows a local martingale under $\mathbb{P}_{T_j}$.

It is clear that the notion of forward Libor measure is in fact identical with that of a forward probability measure for a given date. Also, it is trivial to observe that the forward Libor rate $L(\cdot, T_j)$ necessarily follows a local martingale under the forward Libor measure for the date T_{j+1} . If, in addition, it is a strictly positive process, the existence of the associated volatility process can be justified by standard arguments.

In our further development, we shall go the other way around; that is, we will assume that for any date T_j , the volatility $\lambda(\cdot, T_j)$ of the forward Libor rate $L(\cdot, T_j)$ is exogenously given. In principle, it can be a deterministic $\mathbb{R}^d$ -valued function of time, a $\mathbb{R}^d$ -valued function of the underlying forward Libor rates, or it can follow a d -dimensional adapted stochastic process. For simplicity, we assume throughout that the volatilities of forward Libor rates are bounded processes (or functions). To be more specific, we make the following standing assumptions.

Assumptions (LR) We are given a family of bounded adapted processes $\lambda(\cdot, T_j)$, $j = 0, \dots, n-1$, which represent the volatilities of forward Libor rates $L(\cdot, T_j)$. In addition, we are given an initial term structure of interest rates, specified by a

family $B(0, T_j)$, $j = 0, \dots, n$, of bond prices. We assume here that $B(0, T_j) > B(0, T_{j+1})$ for $j = 0, \dots, n-1$.

Our aim is to construct a family $L(\cdot, T_j)$, $j = 0, \dots, n-1$ of forward Libor rates, a collection of mutually equivalent probability measures $\mathbb{P}_{T_j}$, $j = 1, \dots, n$, and a family W^{T_j} , $j = 1, \dots, n$ of processes in such a way that: (i) for any $j = 1, \dots, n$ the process W^{T_j} follows a d -dimensional standard Brownian motion under the probability measure $\mathbb{P}_{T_j}$, (ii) for any $j = 0, \dots, n-1$, the forward Libor rate $L(\cdot, T_j)$ satisfies the SDE

$$dL(t, T_j) = L(t, T_j) \lambda(t, T_j) \cdot dW_t^{T_{j+1}}, \quad \forall t \in [0, T_j], \quad (2.20)$$

with the initial condition

$$L(0, T_j) = \frac{B(0, T_j) - B(0, T_{j+1})}{\delta_{j+1} B(0, T_{j+1})}.$$

As already mentioned, the construction of the model is based on backward induction, therefore we start by defining the forward Libor rate with the longest maturity, i.e., T_{n-1} . We postulate that $L(\cdot, T_{n-1}) = L(\cdot, T_1^*)$ is governed under the underlying probability measure $\mathbb{P}$ by the following SDE⁶

$$dL(t, T_1^*) = L(t, T_1^*) \lambda(t, T_1^*) \cdot dW_t,$$

with the initial condition

$$L(0, T_1^*) = \frac{B(0, T_1^*) - B(0, T^*)}{\delta_n B(0, T^*)}.$$

Put another way, we have

$$L(t, T_1^*) = \frac{B(0, T_1^*) - B(0, T^*)}{\delta_n B(0, T^*)} \mathcal{E}_t \left(\int_0^\cdot \lambda(u, T_1^*) \cdot dW_u \right).$$

Since $B(0, T_1^*) > B(0, T^*)$, it is clear that the $L(\cdot, T_1^*)$ follows a strictly positive martingale under $\mathbb{P}_{T^*} = \mathbb{P}$. The next step is to define the forward Libor rate for the date T_2^* . For this purpose, we need to introduce first the forward probability measure for the date T_1^* . By definition, it is a probability measure $\mathbb{Q}$, which is equivalent to $\mathbb{P}$, and such that processes

$$U_2(t, T_k^*) = \frac{B(t, T_k^*)}{\delta_{n-1} B(t, T_1^*)}$$

⁶ Notice that, for simplicity, we have chosen the underlying probability measure $\mathbb{P}$ to play the role of the forward Libor measure for the date T^* . This choice is not essential, however.

are $\mathbb{Q}$ -local martingales. It is important to observe that the process $U_2(\cdot, T_k^*)$ admits the following representation:

$$U_2(t, T_k^*) = \frac{\delta_{n-1}\delta_n U_1(t, T_k^*)}{\delta_n L(t, T_1^*) + 1}.$$

Let us formulate an auxiliary result, which is a straightforward consequence of Itô's rule.

Lemma 2.3 *Let G and H be real-valued adapted processes, such that*

$$dG_t = \alpha_t \cdot dW_t, \quad dH_t = \beta_t \cdot dW_t.$$

Assume, in addition, that $H_t > -1$ for every t and denote $Y_t = (1 + H_t)^{-1}$. Then

$$d(Y_t G_t) = Y_t(\alpha_t - Y_t G_t \beta_t) \cdot (dW_t - Y_t \beta_t dt).$$

It follows immediately from Lemma 2.3 that

$$dU_2(t, T_k^*) = \eta_t^k \cdot \left(dW_t - \frac{\delta_n L(t, T_1^*)}{1 + \delta_n L(t, T_1^*)} \lambda(t, T_1^*) dt \right)$$

for a certain process η^k . Therefore it is enough to find a probability measure under which the process

$$W_t^{T_1^*} := W_t - \int_0^t \frac{\delta_n L(u, T_1^*)}{1 + \delta_n L(u, T_1^*)} \lambda(u, T_1^*) du = W_t - \int_0^t \gamma(u, T_1^*) du,$$

$t \in [0, T_1^*]$, follows a standard Brownian motion (the definition of $\gamma(\cdot, T_1^*)$ is clear from the context). This can be easily achieved using Girsanov's theorem, as we may put

$$\frac{d\mathbb{P}_{T_1^*}}{d\mathbb{P}} = \mathcal{E}_{T_1^*} \left(\int_0^\cdot \gamma(u, T_1^*) \cdot dW_u \right), \quad \mathbb{P}\text{-a.s.}$$

We are in a position to specify the dynamics of the forward Libor rate for the date T_2^* under $\mathbb{P}_{T_1^*}$, i.e. we postulate that

$$dL(t, T_2^*) = L(t, T_2^*) \lambda(t, T_2^*) \cdot dW_t^{T_1^*},$$

with the initial condition

$$L(0, T_2^*) = \frac{B(0, T_2^*) - B(0, T_1^*)}{\delta_{n-1} B(0, T_1^*)}.$$

Let us now assume that we have found processes $L(\cdot, T_1^*), \dots, L(\cdot, T_m^*)$. This means, in particular, that the forward Libor measure $\mathbb{P}_{T_{m-1}^*}$ and the associated

Brownian motion $W^{T_m^*}$ are already specified. Our aim is to determine the forward Libor measure $\mathbb{P}_{T_m^*}$. It is easy to check that

$$U_{m+1}(t, T_k^*) = \frac{\delta_{n-m-1} \delta_{n-m} U_m(t, T_k^*)}{\delta_{n-m} L(t, T_m^*) + 1}.$$

Using Lemma 2.3, we obtain the following relationship:

$$W_t^{T_m^*} = W_t^{T_{m-1}^*} - \int_0^t \frac{\delta_{n-m} L(u, T_m^*)}{1 + \delta_{n-m} L(u, T_m^*)} \lambda(u, T_m^*) du$$

for $t \in [0, T_m^*]$. The forward Libor measure $\mathbb{P}_{T_m^*}$ can thus be easily found using Girsanov's theorem. Finally, we define the process $L(\cdot, T_{m+1}^*)$ as the solution to the SDE

$$dL(t, T_{m+1}^*) = L(t, T_{m+1}^*) \lambda(t, T_{m+1}^*) \cdot dW_t^{T_m^*},$$

with the initial condition

$$L(0, T_{m+1}^*) = \frac{B(0, T_{m+1}^*) - B(0, T_m^*)}{\delta_{n-m} B(0, T_m^*)}.$$

Remarks If the volatility coefficient $\lambda(\cdot, T_m) : [0, T_n] \rightarrow \mathbb{R}^d$ is a deterministic function, then for each date $t \in [0, T_m]$ the random variable $L(t, T_m)$ has a lognormal probability law under the forward probability measure $\mathbb{P}_{T_{m+1}}$.

Let us now examine the existence and uniqueness of the implied savings account,⁷ in a discrete-time setup. Intuitively, the value B_t^* of a savings account at time t can be interpreted as the cash amount accumulated up to time t by rolling over a series of zero-coupon bonds with the shortest maturities available. To find the process B^* in a discrete-tenor framework, we do not have to specify explicitly all bond prices; the knowledge of forward bond prices is sufficient. Indeed, it is clear that

$$F_B(t, T_j, T_{j+1}) = \frac{F_B(t, T_j, T^*)}{F_B(t, T_{j+1}, T^*)} = \frac{B(t, T_j)}{B(t, T_{j+1})}.$$

This in turn yields, upon setting $t = T_j$

$$F_B(T_j, T_j, T_{j+1}) = 1/B(T_j, T_{j+1}), \quad (2.21)$$

so that the price $B(T_j, T_{j+1})$ of a single-period bond is uniquely specified for every j . Though the bond that matures at time T_j does not physically exist after this date, it seems justifiable to consider $F_B(T_j, T_j, T_{j+1})$ as its forward value at time T_j for the next future date T_{j+1} . In other words, the spot value at time T_{j+1} of one cash

⁷ The interested reader is referred to Musiela and Rutkowski (1997b) for the definition of an implied savings account in a continuous-time setup. See also Döberlein and Schweizer (1998) and Döberlein *et al.* (2000) for further developments and the general uniqueness result.

unit received at time T_j equals $B^{-1}(T_j, T_{j+1})$. The discrete-time savings account B^* thus equals (recall that $T_{-1} = 0$)

$$B_{T_k}^* = \prod_{j=0}^k F_B(T_{j-1}, T_{j-1}, T_j) = \left(\prod_{j=0}^k B(T_{j-1}, T_j) \right)^{-1}$$

for $k = 0, \dots, n$, since, by convention, we set $B_0^* = 1$. Note that

$$F_B(T_{j-1}, T_{j-1}, T_j) = 1 + \delta L(T_{j-1}, T_j) > 1$$

for $j = 0, \dots, n$, and since

$$B_{T_j}^* = F_B(T_{j-1}, T_{j-1}, T_j) B_{T_{j-1}}^*,$$

we find that $B_{T_j}^* > B_{T_{j-1}}^*$ for every $j = 0, \dots, n$. We conclude that the implied savings account B^* follows a strictly increasing discrete-time process. Let us define the probability measure $\mathbb{P}^*$ equivalent to $\mathbb{P}$ on $(\Omega, \mathcal{F}_{T^*})$ by the formula⁸

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = B_{T^*}^* B(0, T^*), \quad \mathbb{P}\text{-a.s.} \quad (2.22)$$

The probability measure $\mathbb{P}^*$ appears to be a plausible candidate for a spot martingale measure. Indeed, if we set

$$B(T_l, T_k) = \mathbb{E}_{\mathbb{P}^*}(B_{T_l}^* (B_{T_k}^*)^{-1} \mid \mathcal{F}_{T_l}) \quad (2.23)$$

for every $l \leq k \leq n$, then in the case of $l = k - 1$, equality (2.23) coincides with (2.21). Let us observe that it is not possible to uniquely determine the continuous-time dynamics of a bond price $B(t, T_j)$ within the framework of the discrete-tenor model of forward Libor rates (the specification of forward Libor rates for all maturities is necessary for this purpose).

2.2.4 Jamshidian's approach

The backward induction approach to modelling of forward Libor rates presented in the preceding section was re-examined and essentially generalized by Jamshidian (1997). In this section, we present briefly his approach to the modelling of forward Libor rates. As made apparent in the preceding section, in the direct modelling of Libor rates, no explicit reference is made to the bond price processes, which are used to formally define a forward Libor rate through equality (2.12). Nevertheless, to explain the idea that underpins Jamshidian's approach, we shall temporarily assume that we are given a family of bond prices $B(t, T_j)$ for the future dates T_j , $j = 1, \dots, n$. By definition, the *spot Libor measure* is that probability measure equivalent to $\mathbb{P}$, under which all relative bond prices are local martingales, when the

⁸ Recall that $\mathbb{P}$ plays the role of the forward Libor measure for the date T^* . Therefore, formula (2.22) is a consequence of the standard definition of a forward measure.

price process obtained by rolling over single-period bonds is taken as a numeraire. The existence of such a measure can be either postulated or derived from other conditions.⁹ Let us put, for $t \in [0, T^*]$ (as before $T_{-1} = 0$)

$$G_t = B(t, T_{m(t)}) \prod_{j=0}^{m(t)} B^{-1}(T_{j-1}, T_j), \quad (2.24)$$

where

$$m(t) = \inf \left\{ k = 0, 1, \dots \mid \sum_{i=0}^k \delta_i \geq t \right\} = \inf \{k = 0, 1, \dots \mid T_k \geq t\}.$$

It is easily seen that G_t represents the wealth at time t of a portfolio which starts at time 0 with one unit of cash invested in a zero-coupon bond of maturity T_0 , and whose wealth is then reinvested at each date T_j , $j = 0, \dots, n-1$, in zero-coupon bonds which mature at the next date; that is, T_{j+1} .

Definition 2.4 A *spot Libor measure*, denoted by $\mathbb{P}^L$, is a probability measure on $(\Omega, \mathcal{F}_{T^*})$ which is equivalent to $\mathbb{P}$, and such that for any $j = 0, \dots, n$ the relative bond price $B(t, T_j)/G_t$ follows a local martingale under $\mathbb{P}^L$.

Note that

$$B(t, T_{k+1})/G_t = \prod_{j=0}^{m(t)} (1 + \delta_j L(T_{j-1}, T_j))^{-1} \prod_{j=m(t)+1}^k (1 + \delta_j L(t, T_{j-1})),$$

so that all relative bond prices $B(t, T_j)/G_t$, $j = 0, \dots, n$ are uniquely determined by a collection of forward Libor rates. In this sense, G is the correct choice of the reference price process in the present setting. We shall now concentrate on the derivation of the dynamics under $\mathbb{P}^L$ of forward Libor rates $L(\cdot, T_j)$, $j = 0, \dots, n-1$. Our aim is to show that these dynamics involve only the volatilities of forward Libor rates (as opposed to volatilities of bond prices or other processes). Therefore, it is possible to define the whole family of forward Libor rates simultaneously under one probability measure (of course, this feature can also be deduced from the preceding construction). To facilitate the derivation of the dynamics of $L(\cdot, T_j)$, we postulate temporarily that bond prices $B(t, T_j)$ follow Itô processes under the underlying probability measure $\mathbb{P}$, more explicitly

$$dB(t, T_j) = B(t, T_j)(a(t, T_j)dt + b(t, T_j) \cdot dW_t) \quad (2.25)$$

⁹ One may assume, e.g., that bond prices $B(t, T_j)$ satisfy the *weak no-arbitrage condition*, meaning that there exists a probability measure $\tilde{\mathbb{P}}$, equivalent to $\mathbb{P}$, and such that all processes $B(t, T_k)/B(t, T^*)$ are $\tilde{\mathbb{P}}$ -local martingales.

for every $j = 0, \dots, n$, where, as before, W is a d -dimensional standard Brownian motion under an underlying probability measure $\mathbb{P}$ (it should be stressed, however, that we do not assume here that $\mathbb{P}$ is a forward (or spot) martingale measure). Combining (2.24) with (2.25), we obtain

$$dG_t = G_t(a(t, T_{m(t)}) dt + b(t, T_{m(t)}) \cdot dW_t). \quad (2.26)$$

Furthermore, by applying Itô's rule to the equality

$$1 + \delta_{j+1}L(t, T_j) = \frac{B(t, T_j)}{B(t, T_{j+1})}, \quad (2.27)$$

we find that

$$dL(t, T_j) = \mu(t, T_j) dt + \zeta(t, T_j) \cdot dW_t,$$

where

$$\mu(t, T_j) = \frac{B(t, T_j)}{\delta_{j+1}B(t, T_{j+1})} (a(t, T_j) - a(t, T_{j+1})) - \zeta(t, T_j)b(t, T_{j+1})$$

and

$$\zeta(t, T_j) = \frac{B(t, T_j)}{\delta_{j+1}B(t, T_{j+1})} (b(t, T_j) - b(t, T_{j+1})). \quad (2.28)$$

Using (2.27) and the last formula, we arrive at the following relationship:

$$b(t, T_{m(t)}) - b(t, T_{j+1}) = \sum_{k=m(t)}^j \frac{\delta_{k+1}\zeta(t, T_k)}{1 + \delta_{k+1}L(t, T_k)}. \quad (2.29)$$

By definition of a spot Libor measure $\mathbb{P}^L$, each relative price $B(t, T_j)/G_t$ follows a local martingale under $\mathbb{P}^L$. Since, in addition, $\mathbb{P}^L$ is assumed to be equivalent to $\mathbb{P}$, it is clear that it is given by the Doléans exponential, that is

$$\frac{d\mathbb{P}^L}{d\mathbb{P}} = \mathcal{E}_{T^*} \left(\int_0^\cdot h_u \cdot dW_u \right), \quad \mathbb{P}\text{-a.s.}$$

for some adapted process h . It is not hard to check, using Itô's rule, that h necessarily satisfies, for $t \in [0, T_j]$,

$$a(t, T_j) - a(t, T_{m(t)}) = (b(t, T_{m(t)}) - h_t) \cdot (b(t, T_j) - b(t, T_{m(t)}))$$

for every $j = 0, \dots, n$. Combining (2.28) with the last formula, we obtain

$$\frac{B(t, T_j)}{\delta_{j+1}B(t, T_{j+1})} (a(t, T_j) - a(t, T_{j+1})) = \zeta(t, T_j) \cdot (b(t, T_{m(t)}) - h_t),$$

and this in turn yields

$$dL(t, T_j) = \zeta(t, T_j) \cdot \left((b(t, T_{m(t)}) - b(t, T_{j+1}) - h_t) dt + dW_t \right).$$

Using (2.29), we conclude that process $L(\cdot, T_j)$ satisfies

$$dL(t, T_j) = \sum_{k=m(t)}^j \frac{\delta_{k+1} \zeta(t, T_k) \cdot \zeta(t, T_j)}{1 + \delta_{k+1} L(t, T_k)} dt + \zeta(t, T_j) \cdot dW_t^L,$$

where the process $W_t^L = W_t - \int_0^t h_u du$ follows a d -dimensional standard Brownian motion under the spot Libor measure $\mathbb{P}^L$. To further specify the model, we assume that processes $\zeta(t, T_j)$, $j = 0, \dots, n-1$, have the following form, for $t \in [0, T_j]$,

$$\zeta(t, T_j) = \lambda_j(t, L(t, T_j), L(t, T_{j+1}), \dots, L(t, T_n)),$$

where $\lambda_j : [0, T_j] \times \mathbb{R}^{n-j+1} \rightarrow \mathbb{R}^d$ are given functions. In this way, we obtain a system of SDEs

$$dL(t, T_j) = \sum_{k=m(t)}^j \frac{\delta_{k+1} \lambda_k(t, L_k(t)) \cdot \lambda_j(t, L_j(t))}{1 + \delta_{k+1} L(t, T_k)} dt + \lambda_j(t, L_j(t)) \cdot dW_t^L,$$

where we write $L_j(t) = (L(t, T_j), L(t, T_{j+1}), \dots, L(t, T_n))$. Under mild regularity assumptions, this system can be solved recursively, starting from $L(\cdot, T_{n-1})$. The lognormal model of forward Libor rates corresponds to the choice of $\zeta(t, T_j) = \lambda(t, T_j) L(t, T_j)$, where $\lambda(\cdot, T_j) : [0, T_j] \rightarrow \mathbb{R}^d$ is a deterministic function for every j .

2.3 Dynamics of Libor rates and bond prices

We assume that the volatilities of processes $L(\cdot, T_j)$ follow deterministic functions. Put another way, we place ourselves within the framework of the lognormal model of forward Libor rates. It is interesting to note that in all approaches, there is a uniquely determined correspondence between forward measures (and forward Brownian motions) associated with different dates $T_0, \dots, T_n$. On the other hand, however, there is a considerable degree of ambiguity in the way in which the spot martingale measure is specified (in some instances, it is not introduced at all). Consequently, the futures Libor rate $L^f(\cdot, T_j)$, which equals (cf. Section 2.1.3)

$$L^f(t, T_j) = \mathbb{E}_{\mathbb{P}^*}(L(T_j, T_j) | \mathcal{F}_t) = \mathbb{E}_{\mathbb{P}^*}(\tilde{L}(T_j, T_j) | \mathcal{F}_t), \quad (2.30)$$

is not necessarily specified in the same way in various approaches to the lognormal model of forward Libor rates. For this reason, we start by examining the distributional properties of forward Libor rates, which are identical in all abovementioned models.

For a given function $g : \mathbb{R} \rightarrow \mathbb{R}$ and a fixed date $u \leq T_j$, we are interested in the following payoff of the form $X = g(L(u, T_j))$ which settles at time T_j . Particular

cases of such payoffs are

$$X_1 = g(B^{-1}(T_j, T_{j+1})), \quad X_2 = g(B(T_j, T_{j+1})), \quad X_3 = g(F_B(u, T_{j+1}, T_j)).$$

Recall that

$$B^{-1}(T_j, T_{j+1}) = 1 + \delta_{j+1}L(T_j, T_j) = 1 + \delta_{j+1}\tilde{L}(T_j, T_j) = 1 + \delta_{j+1}L^f(T_j, T_j).$$

The choice of the “pricing measure” is thus largely the matter of convenience. Similarly, we have

$$B(T_j, T_{j+1}) = \frac{1}{1 + \delta_{j+1}L(T_j, T_j)} = F_B(T_j, T_{j+1}, T_j). \quad (2.31)$$

More generally, the forward price of a T_{j+1} -maturity bond for the settlement date T_j equals

$$F_B(u, T_{j+1}, T_j) = \frac{B(u, T_{j+1})}{B(u, T_j)} = \frac{1}{1 + \delta_{j+1}L(u, T_j)}. \quad (2.32)$$

Generally speaking, to value the claim $X = g(L(u, T_j)) = \tilde{g}(F_B(u, T_{j+1}, T_j))$ which settles at time T_j we may use the formula

$$\pi_t(X) = B(t, T_j)\mathbb{E}_{\mathbb{P}_{T_j}}(X | \mathcal{F}_t), \quad \forall t \in [0, T_j].$$

It is thus clear that to value a claim in the case $u \leq T_j$, it is enough to know the dynamics of either $L(\cdot, T_j)$ or $F_B(\cdot, T_{j+1}, T_j)$ under the forward probability measure $\mathbb{P}_{T_j}$. If $u = T_j$, we may equally well use the the dynamics, under $\mathbb{P}_{T_j}$, of either $\tilde{L}(\cdot, T_j)$ or $L^f(\cdot, T_j)$. For instance,

$$\begin{aligned} \pi_t(X_1) &= B(t, T_j)\mathbb{E}_{\mathbb{P}_{T_j}}(B^{-1}(T_j, T_{j+1}) | \mathcal{F}_t) \\ &= B(t, T_j)\mathbb{E}_{\mathbb{P}_{T_j}}(F_B^{-1}(T_j, T_{j+1}, T_j) | \mathcal{F}_t), \end{aligned}$$

but also

$$\pi_t(X_1) = B(t, T_j)(1 + \delta_{j+1}\mathbb{E}_{\mathbb{P}_{T_j}}(Z(T_j) | \mathcal{F}_t)),$$

where $Z(T_j) = L(T_j, T_j) = \tilde{L}(T_j, T_j) = L^f(T_j, T_j)$.

2.3.1 Dynamics of $L(\cdot, T_j)$ under $\mathbb{P}_{T_j}$

We shall now derive the transition probability density function (p.d.f.) of the process $L(\cdot, T_j)$ under the forward probability measure $\mathbb{P}_{T_j}$. Let us first prove the following related result, due to Jamshidian (1997).

Proposition 2.5 *Let $t \leq u \leq T_j$. Then*

$$\mathbb{E}_{\mathbb{P}_{T_j}}(L(u, T_j) | \mathcal{F}_t) = L(t, T_j) + \frac{\delta_{j+1}\text{Var}_{\mathbb{P}_{T_{j+1}}}(L(u, T_j) | \mathcal{F}_t)}{1 + \delta_{j+1}L(t, T_j)}. \quad (2.33)$$

In the case of the lognormal model of Libor rates, we have

$$\mathbb{E}_{\mathbb{P}_{T_j}}(L(u, T_j) | \mathcal{F}_t) = L(t, T_j) \left(1 + \frac{\delta_{j+1} L(t, T_j) (e^{v_j^2(t, u)} - 1)}{1 + \delta_{j+1} L(t, T_j)} \right), \quad (2.34)$$

where

$$v_j^2(t, u) = \text{Var}_{\mathbb{P}_{T_{j+1}}} \left(\int_t^u \lambda(s, T_j) \cdot dW_s^{T_{j+1}} \right) = \int_t^u |\lambda(s, T_j)|^2 ds. \quad (2.35)$$

In particular, the modified Libor rate $\tilde{L}(t, T_j)$ satisfies¹⁰

$$\tilde{L}(t, T_j) = \mathbb{E}_{\mathbb{P}_{T_j}}(L(T_j, T_j) | \mathcal{F}_t) = L(t, T_j) \left(1 + \frac{\delta_{j+1} L(t, T_j) (e^{v_j^2(t, T_j)} - 1)}{1 + \delta_{j+1} L(t, T_j)} \right).$$

Proof Combining (2.5) with the martingale property of the process $L(\cdot, T_j)$ under $\mathbb{P}_{T_{j+1}}$, we obtain

$$\mathbb{E}_{\mathbb{P}_{T_j}}(L(u, T_j) | \mathcal{F}_t) = \frac{\mathbb{E}_{\mathbb{P}_{T_{j+1}}}((1 + \delta_{j+1} L(u, T_j)) L(u, T_j) | \mathcal{F}_t)}{1 + \delta_{j+1} L(t, T_j)}$$

so that

$$\mathbb{E}_{\mathbb{P}_{T_j}}(L(u, T_j) | \mathcal{F}_t) = L(t, T_j) + \frac{\delta_{j+1} \mathbb{E}_{\mathbb{P}_{T_{j+1}}}((L(u, T_j) - L(t, T_j))^2 | \mathcal{F}_t)}{1 + \delta_{j+1} L(t, T_j)}.$$

In the case of the lognormal model, we have

$$L(u, T_j) = L(t, T_j) e^{\eta_j(t, u) - \frac{1}{2} v_j^2(t, u)},$$

where

$$\eta_j(t, u) = \int_t^u \lambda(s, T_j) dW_s^{T_{j+1}}. \quad (2.36)$$

Consequently,

$$\mathbb{E}_{\mathbb{P}_{T_{j+1}}}((L(u, T_j) - L(t, T_j))^2 | \mathcal{F}_t) = L^2(t, T_j) (e^{v_j^2(t, u)} - 1).$$

This gives the desired equality (2.34). The last asserted equality is a consequence of (2.6). $\square$

To derive the transition probability density function (p.d.f.) of the process $L(\cdot, T_j)$, notice that for any $t \leq u \leq T_j$, and any bounded Borel measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$ we have

$$\mathbb{E}_{\mathbb{P}_{T_j}}(g(L(u, T_j)) | \mathcal{F}_t) = \frac{\mathbb{E}_{\mathbb{P}_{T_{j+1}}}(g(L(u, T_j)) (1 + \delta_{j+1} L(u, T_j)) | \mathcal{F}_t)}{1 + \delta_{j+1} L(t, T_j)}.$$

¹⁰ This equality can be referred to as the *convexity correction*.

The following simple lemma appears to be useful.

Lemma 2.6 *Let ζ be a nonnegative random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the probability density function $f_{\mathbb{P}}$. Let $\mathbb{Q}$ be a probability measure equivalent to $\mathbb{P}$. Suppose that for any bounded Borel measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$ we have*

$$\mathbb{E}_{\mathbb{P}}(g(\zeta)) = \mathbb{E}_{\mathbb{Q}}((1 + \zeta)g(\zeta)).$$

Then the p.d.f. $f_{\mathbb{Q}}$ of ζ under $\mathbb{Q}$ satisfies $f_{\mathbb{P}}(y) = (1 + y)f_{\mathbb{Q}}(y)$.

Proof The assertion is in fact trivial since, by assumption,

$$\int_{-\infty}^{\infty} g(y) f_{\mathbb{P}}(y) dy = \int_{-\infty}^{\infty} g(y)(1 + y) f_{\mathbb{Q}}(y) dy$$

for any bounded Borel measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$. □

Assume the lognormal model of Libor rates and fix $x \in \mathbb{R}$. Recall that for any $t \geq u$ we have

$$L(u, T_j) = L(t, T_j) e^{\eta_j(t, u) - \frac{1}{2} \text{Var}_{\mathbb{P}_{T_{j+1}}}(\eta_j(t, u))},$$

where $\eta_j(t, u)$ is given by (2.36) (so that it is independent of the σ -field $\mathcal{F}_t$). The Markov property of $L(\cdot, T_j)$ under the forward measure $\mathbb{P}_{T_{j+1}}$ is thus apparent. Denote by $p_L(t, x; u, y)$ the transition p.d.f. under $\mathbb{P}_{T_{j+1}}$ of the process $L(\cdot, T_j)$. Elementary calculations involving Gaussian densities yield

$$\begin{aligned} p_L(t, x; u, y) &= \mathbb{P}_{T_{j+1}}\{L(u, T_j) = y \mid L(t, T_j) = x\} \\ &= \frac{1}{\sqrt{2\pi} v_j(t, u) y} \exp \left\{ -\frac{(\ln(y/x) + \frac{1}{2} v_j^2(t, u))^2}{2 v_j^2(t, u)} \right\} \end{aligned}$$

for any $x, y > 0$ and $t < u$. Taking into account Lemma 2.6, we conclude that the transition p.d.f. of the process¹¹ $L(\cdot, T_j)$, under the forward probability measure $\mathbb{P}_{T_j}$, satisfies

$$\tilde{p}_L(t, x; u, y) = \mathbb{P}_{T_j}\{L(u, T_j) = y \mid L(t, T_j) = x\} = \frac{1 + \delta_{j+1}y}{1 + \delta_{j+1}x} p_L(t, x; u, y).$$

We are in a position to state the following result, which can be used, for instance, to value a contingent claim of the form $X = h(L(T_j))$ which settles at time T_j (see Schmidt (1996)).

¹¹ The Markov property of $L(\cdot, T_j)$ under $\mathbb{P}_{T_j}$ can be easily deduced from the Markovian features of the forward price $F_B(\cdot, T_j, T_{j+1})$ under $\mathbb{P}_{T_j}$ (see formulae (2.37)–(2.38)).

Corollary 2.7 *The transition p.d.f. under $\mathbb{P}_{T_j}$ of the forward Libor rate $L(\cdot, T_j)$ equals, for any $t < u$ and $x, y > 0$,*

$$\tilde{p}_L(t, x; u, y) = \frac{1 + \delta_{j+1}y}{\sqrt{2\pi} v_j(t, u) y(1 + \delta_{j+1}x)} \exp \left\{ -\frac{(\ln(y/x) + \frac{1}{2}v_j^2(t, u))^2}{2v_j^2(t, u)} \right\}.$$

2.3.2 Dynamics of $F_B(\cdot, T_{j+1}, T_j)$ under $\mathbb{P}_{T_j}$

Observe that the forward bond price $F_B(\cdot, T_{j+1}, T_j)$ satisfies

$$F_B(t, T_{j+1}, T_j) = \frac{B(t, T_{j+1})}{B(t, T_j)} = \frac{1}{1 + \delta_{j+1}L(t, T_j)}. \quad (2.37)$$

First, this implies that in the lognormal model of Libor rates, the dynamics of the forward bond price $F_B(\cdot, T_{j+1}, T_j)$ are governed by the following stochastic differential equation, under $\mathbb{P}_{T_j}$,

$$dF_B(t) = -F_B(t)(1 - F_B(t))\lambda(t, T_j) \cdot dW_t^{T_j}, \quad (2.38)$$

where we write $F_B(t) = F_B(t, T_{j+1}, T_j)$. If the initial condition satisfies $0 < F_B(0) < 1$, this equation can be shown to admit a unique strong solution (it satisfies $0 < F_B(t) < 1$ for every $t > 0$). This makes clear that the process $F_B(\cdot, T_{j+1}, T_j)$ – and thus also the process $L(\cdot, T_j)$ – are Markovian under $\mathbb{P}_{T_j}$. Using Corollary 2.7 and relationship (2.37), one can find the transition p.d.f. of the Markov process $F_B(\cdot, T_{j+1}, T_j)$ under $\mathbb{P}_{T_j}$; that is,

$$p_B(t, x; u, y) = \mathbb{P}_{T_j}\{F_B(u, T_{j+1}, T_j) = y \mid F_B(t, T_{j+1}, T_j) = x\}.$$

We have the following result (see Rady and Sandmann (1994), Miltersen *et al.* (1997), and Jamshidian (1997)).

Corollary 2.8 *The transition p.d.f. under $\mathbb{P}_{T_j}$ of the forward bond price $F_B(\cdot, T_{j+1}, T_j)$ equals, for any $t < u$ and arbitrary $0 < x, y < 1$,*

$$p_B(t, x; u, y) = \frac{x}{\sqrt{2\pi} v_j(t, u) y^2(1 - y)} \exp \left\{ -\frac{\left(\ln \frac{x(1-y)}{y(1-x)} + \frac{1}{2}v_j^2(t, u)\right)^2}{2v_j^2(t, u)} \right\}.$$

Proof Let us fix $x \in (0, 1)$. Using (2.37), it is easy to show that

$$p_B(t, x; u, y) = \delta^{-1} y^{-2} \tilde{p}_L\left(t, \frac{1-x}{\delta x}; u, \frac{1-y}{\delta y}\right),$$

where $\delta = \delta_{j+1}$. The formula now follows from Corollary 2.7. $\square$

Let us observe that the results of this section can be applied to value the so-called irregular cash flows, such as caps or floors settled in advance (for more details on this issue we refer to Schmidt (1996)).

2.4 Caps and floors

An *interest rate cap* (known also as a *ceiling rate agreement*) is a contractual arrangement where the grantor (seller) has an obligation to pay cash to the holder (buyer) if a particular interest rate exceeds a mutually agreed level at some future date or dates. Similarly, in an *interest rate floor*, the grantor has an obligation to pay cash to the holder if the interest rate is below a preassigned level. When cash is paid to the holder, the holder's net position is equivalent to borrowing (or depositing) at a rate fixed at that agreed level. This assumes that the holder of a cap (or floor) agreement also holds an underlying asset (such as a deposit) or an underlying liability (such as a loan). Finally, the holder is not affected by the agreement if the interest rate is ultimately more favorable to him than the agreed level. This feature of a cap (or floor) agreement makes it similar to an option. Specifically, a *forward start cap* (or a *forward start floor*) is a strip of caplets (floorlets), each of which is a call (put) option on a forward rate, respectively. Let us denote by κ and by δ_j the cap strike rate and the length of the accrual period, respectively. We shall check that an interest rate caplet (i.e., one leg of a cap) may also be seen as a put option with strike price 1 (per dollar of notional principal) which expires at the caplet start day on a discount bond with face value $1 + \kappa\delta_j$ which matures at the caplet end date.

Similarly to swap agreements, interest rate caps and floors may be settled either *in arrears* or *in advance*. In a forward cap or floor, which starts at time T_0 , and is settled in arrears at dates T_j , $j = 1, \dots, n$, the cash flows at times T_j are $N_p(L(T_{j-1}) - \kappa)^+\delta_j$ and $N_p(\kappa - L(T_{j-1}))^+\delta_j$, respectively, where N_p stands for the notional principal (recall that $\delta_j = T_j - T_{j-1}$). As usual, the rate $L(T_{j-1}) = L(T_{j-1}, T_{j-1})$ is determined at the reset date T_{j-1} , and it satisfies

$$B(T_{j-1}, T_j)^{-1} = 1 + \delta_j L(T_{j-1}). \quad (2.39)$$

The price at time $t \leq T_0$ of a *forward cap*, denoted by $\mathbf{FC}_t$, is (we set $N_p = 1$)

$$\begin{aligned} \mathbf{FC}_t &= \sum_{j=1}^n \mathbb{E}_{\mathbb{P}^*} \left(\frac{B_t}{B_{T_j}} (L(T_{j-1}) - \kappa)^+ \delta_j \middle| \mathcal{F}_t \right) \\ &= \sum_{j=1}^n B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} \left((L(T_{j-1}) - \kappa)^+ \delta_j \middle| \mathcal{F}_t \right). \end{aligned} \quad (2.40)$$

On the other hand, since the cash flow of the j^{th} caplet at time T_j is manifestly an

$\mathcal{F}_{T_{j-1}}$ -measurable random variable, we may directly express the value of the cap in terms of expectations under forward measures $\mathbb{P}_{T_{j-1}}$, $j = 1, \dots, n$. Indeed, we have

$$\mathbf{FC}_t = \sum_{j=1}^n B(t, T_{j-1}) \mathbb{E}_{\mathbb{P}_{T_{j-1}}} \left(B(T_{j-1}, T_j) (L(T_{j-1}) - \kappa)^+ \delta_j \mid \mathcal{F}_t \right). \quad (2.41)$$

Consequently, using (2.39) we get the equality

$$\mathbf{FC}_t = \sum_{j=1}^n B(t, T_{j-1}) \mathbb{E}_{\mathbb{P}_{T_{j-1}}} \left((1 - \tilde{\delta}_j B(T_{j-1}, T_j))^+ \mid \mathcal{F}_t \right), \quad (2.42)$$

which is valid for every $t \in [0, T]$. It is apparent that a caplet is essentially equivalent to a put option on a zero-coupon bond; it may also be seen as an option on a single-period swap.

The equivalence of a cap and a put option on a zero-coupon bond can be explained in an intuitive way. For this purpose, it is enough to examine two basic features of both contracts: the exercise set and the payoff value. Let us consider the j^{th} caplet. A caplet is exercised at time T_{j-1} if and only if $L(T_{j-1}) - \kappa > 0$, or, equivalently, if

$$B(T_{j-1}, T_j)^{-1} = 1 + L(T_{j-1})(T_j - T_{j-1}) > 1 + \kappa \delta_j = \tilde{\delta}_j.$$

The last inequality holds whenever $\tilde{\delta}_j B(T_{j-1}, T_j) < 1$. This shows that both of the considered options are exercised in the same circumstances. If exercised, the caplet pays $\delta_j (L(T_{j-1}) - \kappa)$ at time T_j , or equivalently

$$\delta_j B(T_{j-1}, T_j) (L(T_{j-1}) - \kappa) = 1 - \tilde{\delta}_j B(T_{j-1}, T_j) = \tilde{\delta}_j (\tilde{\delta}_j^{-1} - B(T_{j-1}, T_j))$$

at time T_{j-1} . This shows once again that the j^{th} caplet, with strike level κ and nominal value 1, is essentially equivalent to a put option with strike price $(1 + \kappa \delta_j)^{-1}$ and nominal value $\tilde{\delta}_j = (1 + \kappa \delta_j)$ written on the corresponding zero-coupon bond with maturity T_j .

The analysis of a floor contract can be done along similar lines. By definition, the j^{th} floorlet pays $(\kappa - L(T_{j-1}))^+$ at time T_j . Therefore,

$$\mathbf{FF}_t = \sum_{j=1}^n \mathbb{E}_{\mathbb{P}^*} \left(\frac{B_t}{B_{T_j}} (\kappa - L(T_{j-1}))^+ \delta_j \mid \mathcal{F}_t \right), \quad (2.43)$$

but also

$$\mathbf{FF}_t = \sum_{j=1}^n B(t, T_{j-1}) \mathbb{E}_{\mathbb{P}_{T_{j-1}}} \left((1 - \tilde{\delta}_j B(T_{j-1}, T_j))^+ \mid \mathcal{F}_t \right). \quad (2.44)$$

Combining (2.40) with (2.43) (or (2.42) with (2.44)), we obtain the following cap–floor parity relationship

$$\mathbf{FC}_t - \mathbf{FF}_t = \sum_{j=1}^n (B(t, T_{j-1}) - \tilde{\delta}_j B(t, T_j)), \quad (2.45)$$

which is also an immediate consequence of the no-arbitrage property, so that it does not depend on the model's choice.

2.4.1 Market valuation formula for caps and floors

The main motivation for the introduction of a lognormal model of Libor rates was the market practice of pricing caps and swaptions by means of Black–Scholes-like formulae. For this reason, we shall first describe how market practitioners value caps. The formulae commonly used by practitioners assume that the underlying instrument follows a geometric Brownian motion under some probability measure, $\mathbb{Q}$ say. Since the formal definition of this probability measure is not available, we shall informally refer to $\mathbb{Q}$ as the *market probability*.

Let us consider an interest rate cap with expiry date T and fixed strike level κ . Market practice is to price the option assuming that the underlying forward interest rate process is lognormally distributed with zero drift. Let us first consider a caplet – that is, one leg of a cap. Assume that the forward Libor rate $L(t, T)$, $t \in [0, T]$, for the accrual period of length δ follows a geometric Brownian motion under the “market probability”, $\mathbb{Q}$ say. More specifically,

$$dL(t, T) = L(t, T)\sigma dW_t, \quad (2.46)$$

where W follows a one-dimensional standard Brownian motion under $\mathbb{Q}$, and σ is a strictly positive constant. The unique solution of (2.46) is

$$L(t, T) = L(0, T) \exp\left(\sigma W_t - \frac{1}{2}\sigma^2 t^2\right), \quad \forall t \in [0, T], \quad (2.47)$$

where the initial condition is derived from the yield curve $Y(0, T)$, namely

$$1 + \delta L(0, T) = \frac{B(0, T)}{B(0, T + \delta)} = \exp((T + \delta)Y(0, T + \delta) - TY(0, T)).$$

The “market price” at time t of a caplet with expiry date T and strike level κ is calculated by means of the formula

$$\mathbf{FC}_t = \delta B(t, T + \delta) \mathbb{E}_{\mathbb{Q}}\left((L(T, T) - \kappa)^+ \mid \mathcal{F}_t\right).$$

More explicitly, for any $t \in [0, T]$ we have

$$\mathbf{FC}_t = \delta B(t, T + \delta) \left(L(t, T) N(\hat{e}_1(t, T)) - \kappa N(\hat{e}_2(t, T)) \right), \quad (2.48)$$

where N is the standard Gaussian cumulative distribution function

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-z^2/2} dz, \quad \forall x \in \mathbb{R},$$

and

$$\hat{e}_{1,2}(t, T) = \frac{\ln(L(t, T)/\kappa) \pm \frac{1}{2} \hat{v}_0^2(t, T)}{\hat{v}_0(t, T)}$$

with $\hat{v}_0^2(t, T) = \sigma^2(T - t)$. This means that market practitioners price caplets using Black's formula, with discount from the settlement date $T + \delta$.

A cap settled in arrears at times T_j , $j = 1, \dots, n$, where $T_j - T_{j-1} = \delta_j$, $T_0 = T$, is priced by the formula

$$\mathbf{FC}_t = \sum_{j=1}^n \delta_j B(t, T_j) \left(L(t, T_{j-1}) N(\hat{e}_1^j(t)) - \kappa N(\hat{e}_2^j(t)) \right), \quad (2.49)$$

where for every $j = 0, \dots, n - 1$

$$\hat{e}_{1,2}^j(t) = \frac{\ln(L(t, T_{j-1})/\kappa) \pm \frac{1}{2} \hat{v}_j^2(t)}{\hat{v}_j(t)} \quad (2.50)$$

and $\hat{v}_j^2(t) = (T_{j-1} - t)\sigma_j^2$ for some constants σ_j , $j = 1, \dots, n$. Apparently, the market assumes that for any maturity T_j , the corresponding forward Libor rate has a lognormal probability law under the “market probability”. The value of a floor can be easily derived by combining (2.49)–(2.50) with the cap–floor parity relationship (2.45). As we shall see in what follows, the valuation formulae obtained for caps and floors in the lognormal model of forward Libor rates agree with the market practice.

2.4.2 Valuation in the lognormal model of forward Libor rates

We shall now examine the valuation of caps within the lognormal model of forward Libor rates of Section 2.2.3. The dynamics of the forward Libor rate $L(t, T_{j-1})$ under the forward probability measure $\mathbb{P}_{T_j}$ are

$$dL(t, T_{j-1}) = L(t, T_{j-1}) \lambda(t, T_{j-1}) \cdot dW_t^{T_j}, \quad (2.51)$$

where W^{T_j} follows a d -dimensional Brownian motion under the forward measure $\mathbb{P}_{T_j}$, and $\lambda(\cdot, T_{j-1}) : [0, T_{j-1}] \rightarrow \mathbb{R}^d$ is a deterministic function. Consequently, for every $t \in [0, T_{j-1}]$ we have

$$L(t, T_{j-1}) = L(0, T_{j-1}) \mathcal{E}_t \left(\int_0^t \lambda(u, T_{j-1}) \cdot dW_u^{T_j} \right).$$

In the present setup, the cap valuation formula (2.52) was first established by Miltersen *et al.* (1997), who focused on the dynamics of the forward Libor rate

for a given date. Equality (2.52) was subsequently rederived through a probabilistic approach in Goldys (1997) and Rady (1997). Finally, the same result was established by means of the forward measure approach in Brace *et al.* (1997). The following proposition is a consequence of formula (2.41), combined with the dynamics (2.51). As before, N is the standard Gaussian probability distribution function.

Proposition 2.9 *Consider an interest rate cap with strike level κ , settled in arrears at times T_j , $j = 1, \dots, n$. Assuming the lognormal model of Libor rates, the price of a cap at time $t \in [0, T]$ equals*

$$\mathbf{FC}_t = \sum_{j=1}^n \delta_j B(t, T_j) \left(L(t, T_{j-1}) N(\tilde{e}_1^j(t)) - \kappa N(\tilde{e}_2^j(t)) \right) = \sum_{j=1}^n \mathbf{FC}_t^j, \quad (2.52)$$

where $\mathbf{FC}_t^j$ stands for the price at time t of the j^{th} caplet for $j = 1, \dots, n$,

$$\tilde{e}_{1,2}^j(t) = \frac{\ln(L(t, T_{j-1})/\kappa) \pm \frac{1}{2} \tilde{v}_j^2(t)}{\tilde{v}_j(t)}$$

and

$$\tilde{v}_j^2(t) = \int_t^{T_{j-1}} |\lambda(u, T_{j-1})|^2 du.$$

Proof We fix j and we consider the j^{th} caplet. It is clear that its payoff at time T_j admits the representation

$$\mathbf{FC}_{T_j}^j = \delta_j (L(T_{j-1}) - \kappa)^+ = \delta_j L(T_{j-1}) \mathbb{1}_D - \delta_j \kappa \mathbb{1}_D, \quad (2.53)$$

where $D = \{L(T_{j-1}) > \kappa\}$ is the exercise set. Since the caplet settles at time T_j , it is convenient to use the forward measure $\mathbb{P}_{T_j}$ to find its arbitrage price. We have

$$\mathbf{FC}_t^j = B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} (\mathbf{FC}_{T_j}^j | \mathcal{F}_t), \quad \forall t \in [0, T_j].$$

Obviously, it is enough to find the value of a caplet for $t \in [0, T_{j-1}]$. In view of (2.53), it is clear that we need to evaluate the following conditional expectations:

$$\begin{aligned} \mathbf{FC}_t^j &= \delta_j B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} (L(T_{j-1}) \mathbb{1}_D | \mathcal{F}_t) - \kappa \delta_j B(t, T_j) \mathbb{P}_{T_j}(D | \mathcal{F}_t) \\ &= \delta_j B(t, T_j) (I_1 - I_2), \end{aligned}$$

where the meaning of I_1 and I_2 is obvious from the context. Recall that $L(T_{j-1})$ is given by the formula

$$L(T_{j-1}) = L(t, T_{j-1}) \exp \left(\int_t^{T_{j-1}} \lambda(u, T_{j-1}) \cdot dW_u^{T_j} - \frac{1}{2} \int_t^{T_{j-1}} |\lambda(u, T_{j-1})|^2 du \right).$$

Since $\lambda(\cdot, T_{j-1})$ is a deterministic function, the probability law under $\mathbb{P}_{T_j}$ of the Itô integral

$$\zeta(t, T_{j-1}) = \int_t^{T_{j-1}} \lambda(u, T_{j-1}) \cdot dW_u^{T_j}$$

is Gaussian, with zero mean and the variance

$$\text{Var}_{\mathbb{P}_{T_j}}(\zeta(t, T_{j-1})) = \int_t^{T_{j-1}} |\lambda(u, T_{j-1})|^2 du.$$

Therefore, it is straightforward to show that¹²

$$I_2 = \kappa N \left(\frac{\ln L(t, T_{j-1}) - \ln \kappa - \frac{1}{2} v_j^2(t)}{v_j(t)} \right).$$

To evaluate I_1 , we introduce an auxiliary probability measure $\hat{\mathbb{P}}_{T_j}$, equivalent to $\mathbb{P}_{T_j}$ on $(\Omega, \mathcal{F}_{T_{j-1}})$, by setting

$$\frac{d\hat{\mathbb{P}}_{T_j}}{d\mathbb{P}_{T_j}} = \mathcal{E}_{T_{j-1}} \left(\int_0^\cdot \lambda(u, T_{j-1}) \cdot dW_u^{T_j} \right).$$

Then the process $\hat{W}^{T_j}$ given by the formula

$$\hat{W}_t^{T_j} = W_t^{T_j} - \int_0^t \lambda(u, T_{j-1}) du, \quad \forall t \in [0, T_{j-1}],$$

follows the d -dimensional standard Brownian motion under $\hat{\mathbb{P}}_{T_j}$. Furthermore, the forward price $L(T_{j-1})$ admits the representation under $\hat{\mathbb{P}}_{T_j}$, for $t \in [0, T_{j-1}]$,

$$L(T_{j-1}) = L(t, T_{j-1}) \exp \left(\int_t^{T_{j-1}} \lambda_{j-1}(u) \cdot d\hat{W}_u^{T_j} + \frac{1}{2} \int_t^{T_{j-1}} |\lambda_{j-1}(u)|^2 du \right)$$

where we set $\lambda_{j-1}(u) = \lambda(u, T_{j-1})$. Since

$$I_1 = L(t, T_{j-1}) \mathbb{E}_{\mathbb{P}_{T_j}} \left(\mathbb{1}_D \exp \left(\int_t^{T_{j-1}} \lambda_{j-1}(u) \cdot dW_u^{T_j} - \frac{1}{2} \int_t^{T_{j-1}} |\lambda_{j-1}(u)|^2 du \right) \middle| \mathcal{F}_t \right)$$

from the abstract Bayes rule, we get $I_1 = L(t, T_{j-1}) \hat{\mathbb{P}}_{T_j}(D | \mathcal{F}_t)$. Arguing in much the same way as for I_2 , we thus obtain

$$I_1 = L(t, T_{j-1}) N \left(\frac{\ln L(t, T_{j-1}) - \ln \kappa + \frac{1}{2} v_j^2(t)}{v_j(t)} \right).$$

This completes the proof of the proposition. □

¹² See, for instance, the proof of the Black–Scholes formula in Musiela and Rutkowski (1997a).

Once again, to derive the floors valuation formula, it is enough to make use of the cap–floor parity (2.45).

2.4.3 Hedging of caps and floors

It is clear that the replicating strategy for a cap is a simple sum of replicating strategies for caplets. Therefore, it is enough to focus on a particular caplet. Let us denote by $F_C(t, T_j)$ the forward price of the j^{th} caplet for the settlement date T_j . From (2.52), it is clear that

$$F_C(t, T_j) = \delta_j L(t, T_{j-1}) N(\tilde{e}_1^j(t)) - \kappa N(\tilde{e}_2^j(t)),$$

so that an application of Itô's formula yields¹³

$$dF_C(t, T_j) = \delta_j N(\tilde{e}_1^j(t)) dL(t, T_{j-1}). \quad (2.54)$$

Let us consider the following self-financing trading strategy in the T_j -forward market. We start our trade at time 0 with $F_C(0, T_j)$ units of zero-coupon bonds.¹⁴ At any time $t \leq T_{j-1}$ we assume $\psi_t^j = N(\tilde{e}_1^j(t))$ positions in forward rate agreements (that is, single-period forward swaps) over the period $[T_{j-1}, T_j]$. The associated gains/losses process V , in the T_j forward market,¹⁵ satisfies¹⁶

$$dV_t = \delta_j \psi_t^j dL(t, T_{j-1}) = \delta_j N(\tilde{e}_1^j(t)) dL(t, T_{j-1}) = dF_C(t, T)$$

with $V_0 = 0$. Consequently,

$$F_C(T_{j-1}, T_j) = F_C(0, T_j) + \int_0^{T_{j-1}} \delta_j \psi_t^j dL(t, T_{j-1}) = F_C(0, T_j) + V_{T_{j-1}}.$$

It should be stressed that dynamic trading takes place on the interval $[0, T_{j-1}]$ only, the gains/losses (involving the initial investment) are incurred at time T_j , however. All quantities in the last formula are expressed in units of T_j -maturity zero-coupon bonds. Also, the caplet's payoff is known already at time T_{j-1} , so that it is completely specified by its forward price $F_C(T_{j-1}, T_j) = \mathbf{FC}_{T_{j-1}}^j / B(T_{j-1}, T_j)$. Therefore the last equality makes it clear that the strategy ψ introduced above does indeed replicate the j^{th} caplet.

It should be observed that formally the replicating strategy has also second component, η_t^j say, which represents the number of forward contracts on a T_j -maturity bond, with the settlement date T_j . Since obviously $F_B(t, T_j, T_j) = 1$ for every $t \leq T_j$, so that $dF_B(t, T_j, T_j) = 0$, for the T_j -forward value of our strategy, we get

¹³ The calculations here are essentially the same as in the classic Black–Scholes model.

¹⁴ We need thus to invest $\mathbf{FC}_0^j = F_C(0, T_j)B(0, T_j)$ of cash at time 0.

¹⁵ That is, with the value expressed in units of T_j -maturity zero-coupon bonds.

¹⁶ To get a more intuitive insight in this formula, it is advisable to consider first a discretized version of ψ .

$\tilde{V}_t(\psi^j, \eta^j) = \eta_t^j = F_C(t, T_j)$ and

$$d\tilde{V}_t(\psi^j, \eta^j) = \psi_t^j \delta_j dL(t, T_{j-1}) + \eta_t^j dF_B(t, T_j, T_j) = \delta_j N(\tilde{e}_1^j(t)) dL(t, T_{j-1}).$$

It should be stressed, however, with the exception for the initial investment at time 0 in T_j -maturity bonds, no bonds trading is required for the caplet's replication. In practical terms, the hedging of a cap within the framework of the lognormal model of forward Libor rates is done exclusively through dynamic trading in the underlying single-period swaps. Of course, the same remarks (and similar calculations) apply also to floors. In this interpretation, the component η^j simply represents the future (i.e., as of time T_{j-1}) effects of a continuous trading in forward contracts.

Alternatively, the hedging of a cap can be done in the spot (i.e., cash) market, using two simple portfolios of bonds. Indeed, it is easily seen that for the process

$$V_t(\psi^j, \eta^j) = B(t, T_{j-1})\tilde{V}_t(\psi^j, \eta^j) = \mathbf{FC}_t^j$$

we have

$$V_t(\psi^j, \eta^j) = \psi_t^j (B(t, T_{j-1}) - B(t, T_j)) + \eta_t^j dF_B(t, T_j, T_j)$$

and

$$\begin{aligned} dV_t(\psi^j, \eta^j) &= \psi_t^j d(B(t, T_{j-1}) - B(t, T_j)) + \eta_t^j dB(t, T_j) \\ &= N(\tilde{e}_1^j(t)) d(B(t, T_{j-1}) - B(t, T_j)) + \eta_t^j dB(t, T_j). \end{aligned}$$

This means that the components ψ^j and η^j now represent the number of units of portfolios $B(t, T_{j-1}) - B(t, T_j)$ and $B(t, T_j)$ held at time t .

2.4.4 Bond options

We shall now give the bond option valuation formula within the framework of the lognormal model of forward Libor rates. This result was first obtained by Rady and Sandmann (1994), who adopted the PDE approach and who worked in a different setup (see also Goldys (1997), Miltersen *et al.* (1997), and Rady (1997)). In the present framework, it is an immediate consequence of (2.52) combined with (2.42).

Proposition 2.10 *The price $\mathbf{C}_t$ at time $t \leq T_{j-1}$ of a European call option, with expiration date T_{j-1} and strike price $0 < K < 1$, written on a zero-coupon bond maturing at $T_j = T_{j-1} + \delta_j$, equals*

$$\mathbf{C}_t = (1 - K)B(t, T_j)N(l_1^j(t)) - K(B(t, T_{j-1}) - B(t, T_j))N(l_2^j(t)), \quad (2.55)$$

where

$$l_{1,2}^j(t) = \frac{\ln((1 - K)B(t, T_j)) - \ln(K(B(t, T_{j-1}) - B(t, T_j))) \pm \frac{1}{2}\tilde{v}_j(t)}{\tilde{v}_j(t)}$$

and

$$\tilde{v}_j^2(t) = \int_t^{T_{j-1}} |\lambda(u, T_{j-1})|^2 du.$$

In view of (2.55), it is apparent that the replication of the bond option using the underlying bonds of maturity T_{j-1} and T_j is rather involved. This should be contrasted with the case of the Gaussian Heath–Jarrow–Morton model¹⁷ in which hedging of bond options with the use of the underlying bonds is straightforward. This illustrates the general feature that each particular way of modelling the term structure is tailored to the specific class of derivatives and hedging instruments.

3 Modelling of forward swap rates

We shall first describe the most typical swap contracts and related options (the so-called *swaptions*). Subsequently, we shall present a model of forward swap rates put forward by Jamshidian (1996, 1997). For the sake of expositional convenience, we shall follow the backward induction approach due to Rutkowski (1999), however.

3.1 Interest rate swaps

Let us consider a *forward (start) payer swap* (that is, *fixed-for-floating interest rate swap*) settled in arrears, with notional principal N_p . As before, we consider a finite collection of dates $0 < T_0 < T_1 < \dots < T_n$ so that $\delta_j = T_j - T_{j-1} > 0$ for every $j = 1, \dots, n$. The floating rate $L(T_{j-1})$ received at time T_j is set at time T_{j-1} by reference to the price of a zero-coupon bond over the period $[T_{j-1}, T_j]$. More specifically, $L(T_{j-1})$ is the spot Libor rate prevailing at time T_{j-1} , so that it satisfies

$$B(T_{j-1}, T_j)^{-1} = 1 + (T_j - T_{j-1})L(T_{j-1}) = 1 + \delta_j L(T_{j-1}). \quad (3.1)$$

Recall that in general, the forward Libor rate $L(t, T_{j-1})$ for the future time period $[T_{j-1}, T_j]$ of length δ_j satisfies

$$1 + \delta_j L(t, T_{j-1}) = \frac{B(t, T_{j-1})}{B(t, T_j)} = F_B(t, T_{j-1}, T_j), \quad (3.2)$$

so that $L(T_{j-1})$ coincides with $L(T_{j-1}, T_{j-1})$. At any date T_j , $j = 1, \dots, n$, the cash flows of a forward payer swap are $N_p L(T_{j-1})\delta_j$ and $-N_p \kappa \delta_j$, where κ is a preassigned fixed rate of interest (the cash flows of a forward receiver swap have the same size, but opposite signs). The number n , which coincides with the number of payments, is referred to as the *length* of a swap, (for instance, the length of a

¹⁷ In such a model the forward prices of bonds follow lognormal processes.

three-year swap with quarterly settlement equals $n = 12$). The dates $T_0, \dots, T_{n-1}$ are known as *reset dates*, and the dates $T_1, \dots, T_n$ as *settlement dates*. We shall refer to the first reset date T_0 as the *start date* of a swap. Finally, the time interval $[T_{j-1}, T_j]$ is referred to as the j^{th} *accrual period*. We may and do assume, without loss of generality, that the notional principal $N_p = 1$.

The value at time t of a forward payer swap, which is denoted by $\mathbf{FS}_t$ or $\mathbf{FS}_t(\kappa)$, equals

$$\mathbf{FS}_t(\kappa) = \mathbb{E}_{\mathbb{P}^*} \left\{ \sum_{j=1}^n \frac{B_t}{B_{T_j}} (L(T_{j-1}) - \kappa) \delta_j \mid \mathcal{F}_t \right\}. \quad (3.3)$$

Since

$$L(t, T_{j-1}) = \frac{B(t, T_{j-1}) - B(t, T_j)}{\delta_j B(t, T_j)},$$

it is clear that the process $L(\cdot, T_{j-1})$ follows a martingale under the forward martingale measure $\mathbb{P}_{T_j}$. Therefore

$$\begin{aligned} \mathbf{FS}_t(\kappa) &= \sum_{j=1}^n B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} ((L(T_{j-1}) - \kappa) \delta_j \mid \mathcal{F}_t) \\ &= \sum_{j=1}^n B(t, T_j) ((L(t, T_{j-1}) - \kappa) \delta_j) \\ &= \sum_{j=1}^n (B(t, T_{j-1}) - B(t, T_j) - \kappa \delta_j B(t, T_j)). \end{aligned}$$

After rearranging, this yields

$$\mathbf{FS}_t(\kappa) = B(t, T_0) - \sum_{j=1}^n c_j B(t, T_j) \quad (3.4)$$

for every $t \in [0, T]$, where $c_j = \kappa \delta_j$ for $j = 1, \dots, n-1$, and $c_n = \tilde{\delta}_n = 1 + \kappa \delta_n$. The last equality makes clear that a forward payer swap settled in arrears is, essentially, a contract to deliver a specific coupon-bearing bond and to receive at the same time a zero-coupon bond. Relationship (3.4) may also be established through a straightforward comparison of the future cash flows from these bonds. Note that (3.4) provides a simple method for the replication of a swap contract, independent of the term structure model.

In the forward payer swap settled in advance – that is, in which each reset date is also a settlement date – the discounting method varies from country to country. In the U.S. and in many European markets, the cash flows of a swap settled in advance at reset dates T_j , $j = 0, \dots, n-1$, are $L(T_j) \delta_{j+1} (1 + L(T_j) \delta_{j+1})^{-1}$ and

$-\kappa\delta_{j+1}(1 + L(T_j)\delta_{j+1})^{-1}$. Therefore the value $\mathbf{FS}_t^{**}(\kappa)$ at time t of this swap is

$$\begin{aligned}\mathbf{FS}_t^{**}(\kappa) &= \mathbb{E}_{\mathbb{P}^*} \left\{ \sum_{j=0}^{n-1} \frac{B_t}{B_{T_j}} \frac{\delta_{j+1}(L(T_j) - \kappa)}{1 + \delta_{j+1}L(T_j)} \middle| \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \sum_{j=0}^{n-1} \frac{B_t}{B_{T_j}} (L(T_j) - \kappa) \delta_{j+1} B(T_j, T_{j+1}) \middle| \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \sum_{j=0}^{n-1} \frac{B_t}{B_{T_{j+1}}} (L(T_j) - \kappa) \delta_{j+1} \middle| \mathcal{F}_t \right\},\end{aligned}$$

which coincides with the value of the swap settled in arrears. Once again, this is by no means surprising, since the payoffs $L(T_j)\delta_{j+1}(1 + L(T_j)\delta_{j+1})^{-1}$ and $-\kappa\delta_{j+1}(1 + L(T_j)\delta_{j+1})^{-1}$ at time T_j are easily seen to be equivalent to payoffs $L(T_j)\delta_{j+1}$ and $-\kappa\delta_{j+1}$ respectively at time T_{j+1} (recall that $1 + L(T_j)\delta_{j+1} = B^{-1}(T_j, T_{j+1})$).

In what follows, we shall restrict our attention to interest rate swaps settled in arrears. As mentioned, a swap agreement is worthless at initiation. This important feature of a swap leads to the following definition, which refers in fact to the more general concept of a forward swap. Basically, a forward swap rate is that fixed rate of interest which makes a forward swap worthless.

Definition 3.1 The *forward swap rate* $\kappa(t, T_0, n)$ at time t for the date T_0 is that value of the fixed rate κ which makes the value of the forward swap zero, i.e., that value of κ for which $\mathbf{FS}_t(\kappa) = 0$. Using (3.4), we obtain

$$\kappa(t, T_0, n) = (B(t, T_0) - B(t, T_n)) \left(\sum_{j=1}^n \delta_j B(t, T_j) \right)^{-1}. \quad (3.5)$$

A *swap* (*swap rate*, respectively) is the forward swap (forward swap rate, respectively) with $t = T$. The swap rate, $\kappa(T_0, T_0, n)$, equals

$$\kappa(T_0, T_0, n) = (1 - B(T_0, T_n)) \left(\sum_{j=1}^n \delta_j B(T_0, T_j) \right)^{-1}. \quad (3.6)$$

Note that the definition of a forward swap rate implicitly refers to a swap contract of length n which starts at time T_0 . It would thus be more correct to refer to $\kappa(t, T_0, n)$ as the *n-period forward swap rate* prevailing at time t , for the future date T_0 . A forward swap rate is a rather theoretical concept, as opposed to swap rates, which are quoted daily (subject to an appropriate bid–ask spread) by financial institutions who offer interest rate swap contracts to their institutional clients. In practice, swap agreements of various lengths are offered. Also, typically, the length of the reference period varies over time; for instance, a five-year swap may be

settled quarterly during the first three years, and semi-annually during the last two. Swap rates also play an important role as a basis for several derivative instruments. For instance, an appropriate swap rate is commonly used as a strike level for an option written on the value of a swap; that is, a *swaption*.

Finally, it will be useful to express that value at time t of a given forward swap with fixed rate κ in terms of the current value of the forward swap rate. Since obviously $\mathbf{FS}_t(\kappa(t, T_0, n)) = 0$, using (3.4), we get

$$\mathbf{FS}_t(\kappa) = \mathbf{FS}_t(\kappa) - \mathbf{FS}_t(\kappa(t, T_0, n)) = \sum_{j=1}^n (\kappa(t, T_0, n) - \kappa) B(t, T_j). \quad (3.7)$$

3.2 The lognormal model of forward swap rates

The lognormal model of forward swap rates was developed by Jamshidian (1996, 1997). In this section, we follow Rutkowski (1999). We assume, as before, that the tenor structure $0 < T_0 < T_1 < \dots < T_n = T^*$ is given. Recall that $\delta_j = T_j - T_{j-1}$ for $j = 1, \dots, n$, and thus $T_j = \sum_{i=0}^j \delta_i$ for every $j = 0, \dots, n$. For any fixed j , we consider a fixed-for-floating forward (payer) swap which starts at time T_j and has $n - j$ accrual periods, whose consecutive lengths are $\delta_{j+1}, \dots, \delta_n$. The fixed interest rate paid at each of the reset dates T_l for $l = j + 1, \dots, n$ equals κ , and the corresponding floating rate, $L(T_l)$, is found using the formula

$$B(T_l, T_{l+1})^{-1} = 1 + (T_{l+1} - T_l)L(T_l) = 1 + \delta_{l+1}L(T_l),$$

i.e., it coincides with the Libor rate $L(T_l, T_l)$. It is not difficult to check, using no-arbitrage arguments, that the value of such a swap equals, for $t \in [0, T_j]$ (by convention, the notional principal equals 1)

$$\mathbf{FS}_t(\kappa) = B(t, T_j) - \sum_{l=j+1}^n c_l B(t, T_l),$$

where $c_l = \kappa \delta_l$ for $l = j + 1, \dots, n - 1$, and $c_n = 1 + \kappa \delta_n$. Consequently, the associated *forward swap rate*, $\kappa(t, T_j, n - j)$, that is, that value of a fixed rate κ for which such a swap is worthless at time t , is given by the formula

$$\kappa(t, T_j, n - j) = \frac{B(t, T_j) - B(t, T_n)}{\delta_{j+1} B(t, T_{j+1}) + \dots + \delta_n B(t, T_n)} \quad (3.8)$$

for every $t \in [0, T_j]$, $j = 0, \dots, n - 1$. In this section, we consider the family of forward swap rates $\tilde{\kappa}(t, T_j) = \kappa(t, T_j, n - j)$ for $j = 0, \dots, n - 1$. Let us stress that the underlying swap agreements differ in length, however, they all have a common expiration date, $T^* = T_n$.

Suppose momentarily that we are given a family of bond prices $B(t, T_m)$, $m = 1, \dots, n$, on a filtered probability space $(\Omega, \mathbb{F}, \mathbb{P})$ equipped with a Brownian

motion W . As in Section 2.1, we find it convenient to postulate that $\mathbb{P} = \mathbb{P}_{T^*}$ is the forward measure for the date T^* , and the process $W = W^{T^*}$ is the corresponding Brownian motion. For any $m = 1, \dots, n-1$, we introduce the fixed-maturity coupon process $G(m)$ by setting (recall that $T_l^* = T_{n-l}$, in particular, $T_0^* = T_n$)

$$G_t(m) = \sum_{l=n-m+1}^n \delta_l B(t, T_l) = \sum_{k=0}^{m-1} \delta_{n-k} B(t, T_k^*) \quad (3.9)$$

for $t \in [0, T_{n-m+1}]$. A *forward swap measure* is that probability measure, equivalent to $\mathbb{P}$, which corresponds to the choice of the fixed-maturity coupon process as a numeraire asset. We have the following definition.

Definition 3.2 For $j = 0, \dots, n$, a probability measure $\tilde{\mathbb{P}}_{T_j}$ on $(\Omega, \mathcal{F}_{T_j})$, equivalent to $\mathbb{P}$, is said to be the *fixed-maturity forward swap measure* for the date T_j if, for every $k = 0, \dots, n$, the relative bond price

$$Z_{n-j+1}(t, T_k) := \frac{B(t, T_k)}{G_t(n-j+1)} = \frac{B(t, T_k)}{\delta_j B(t, T_j) + \dots + \delta_n B(t, T_n)},$$

$t \in [0, T_k \wedge T_j]$, follows a local martingale under $\tilde{\mathbb{P}}_{T_j}$.

Put another way, for any fixed $m = 1, \dots, n+1$, the relative bond prices

$$Z_m(t, T_k^*) = \frac{B(t, T_k^*)}{G_t(m)} = \frac{B(t, T_k^*)}{\delta_{n-m+1} B(t, T_{m-1}^*) + \dots + \delta_n B(t, T^*)},$$

$t \in [0, T_k^* \wedge T_{m-1}^*]$, are bound to follow local martingales under the forward swap measure $\tilde{\mathbb{P}}_{T_{m-1}^*}$. It follows immediately from (3.8) that the forward swap rate for the date T_m^* equals, for $t \in [0, T_m^*]$,

$$\tilde{\kappa}(t, T_m^*) = \frac{B(t, T_m^*) - B(t, T^*)}{\delta_{n-m+1} B(t, T_{m-1}^*) + \dots + \delta_n B(t, T^*)},$$

or, equivalently,

$$\tilde{\kappa}(t, T_m^*) = Z_m(t, T_m^*) - Z_m(t, T^*).$$

Therefore $\tilde{\kappa}(\cdot, T_m^*)$ also follows a local martingale under the forward swap measure $\tilde{\mathbb{P}}_{T_{m-1}^*}$. Moreover, since obviously $G_t(1) = \delta_n B(t, T^*)$, it is evident that $Z_1(t, T_k^*) = \delta_n^{-1} F_B(t, T_k^*, T^*)$, and thus the probability measure $\tilde{\mathbb{P}}_{T^*}$ can be chosen to coincide with the forward martingale measure $\mathbb{P}_{T^*}$. Our aim is to construct a model of forward swap rates through backward induction. As one might expect, the underlying bond price processes will not be explicitly specified. We make the following standing assumptions.

Assumptions (SR) We assume that we are given a family of bounded adapted processes $v(\cdot, T_j)$, $j = 0, \dots, n-1$, which represent the volatilities of forward swap rates $\tilde{\kappa}(\cdot, T_j)$. In addition, we are given an initial term structure of interest rates, specified by a family $B(0, T_j)$, $j = 0, \dots, n$, of bond prices. We assume that $B(0, T_j) > B(0, T_{j+1})$ for $j = 0, \dots, n-1$.

We wish to construct a family of forward swap rates in such a way that

$$d\tau\kappa(t, T_j) = \tilde{\kappa}(t, T_j)v(t, T_j) \cdot d\tilde{W}_t^{T_{j+1}} \quad (3.10)$$

for any $j = 0, \dots, n-1$, where each process $\tilde{W}^{T_{j+1}}$ follows a standard Brownian motion under the corresponding forward swap measure $\tilde{\mathbb{P}}_{T_{j+1}}$. The model should also be consistent with the initial term structure of interest rates, meaning that

$$\tilde{\kappa}(0, T_j) = \frac{B(0, T_j) - B(0, T^*)}{\delta_{j+1}B(0, T_{j+1}) + \dots + \delta_n B(0, T_n)}. \quad (3.11)$$

We proceed by backward induction. The first step is to introduce the forward swap rate for the date T_1^* by postulating that the forward swap rate $\tilde{\kappa}(\cdot, T_1^*)$ solves the SDE

$$d\tau\kappa(t, T_1^*) = \tilde{\kappa}(t, T_1^*)v(t, T_1^*) \cdot d\tau W_t^{T^*}, \quad \forall t \in [0, T_1^*], \quad (3.12)$$

where $\tilde{W}^{T^*} = W^{T^*} = W$, with the initial condition

$$\tilde{\kappa}(0, T_1^*) = \frac{B(0, T_1^*) - B(0, T^*)}{\delta_n B(0, T^*)}.$$

To specify the process $\tilde{\kappa}(\cdot, T_2^*)$, we need first to introduce a forward swap measure $\tilde{\mathbb{P}}_{T_1^*}$ and an associated Brownian motion $\tilde{W}^{T_1^*}$. To this end, notice that each process $Z_1(\cdot, T_k^*) = B(\cdot, T_k^*)/\delta_n B(\cdot, T^*)$, follows a strictly positive local martingale under $\tilde{\mathbb{P}}_{T^*} = \mathbb{P}_{T^*}$. More specifically, we have

$$dZ_1(t, T_k^*) = Z_1(t, T_k^*)\gamma_1(t, T_k^*) \cdot d\tau W_t^{T^*} \quad (3.13)$$

for some adapted process $\gamma_1(\cdot, T_k^*)$. According to the definition of a fixed-maturity forward swap measure, we postulate that for every k the process

$$Z_2(t, T_k^*) = \frac{B(t, T_k^*)}{\delta_{n-1}B(t, T_1^*) + \delta_n B(t, T^*)} = \frac{Z_1(t, T_k^*)}{1 + \delta_{n-1}Z_1(t, T_1^*)}$$

follows a local martingale under $\tilde{\mathbb{P}}_{T_1^*}$. Applying Lemma 2.3 to processes $G = Z_1(\cdot, T_k^*)$ and $H = \delta_{n-1}Z_1(\cdot, T_1^*)$, it is easy to see that for this property to hold, it suffices to assume that the process $\tilde{W}^{T_1^*}$, which is given by the formula

$$\tilde{W}_t^{T_1^*} = \tilde{W}_t^{T^*} - \int_0^t \frac{\delta_{n-1}Z_1(u, T_1^*)}{1 + \delta_{n-1}Z_1(u, T_1^*)} \gamma_1(u, T_1^*) du,$$

$t \in [0, T_1^*]$, follows a Brownian motion under $\tilde{\mathbb{P}}_{T_1^*}$, (the probability measure $\tilde{\mathbb{P}}_{T_1^*}$ is yet unspecified, but will be soon found through Girsanov's theorem). Note that

$$Z_1(t, T_1^*) = \frac{B(t, T_1^*)}{\delta_n B(t, T^*)} = \tilde{\kappa}(t, T_1^*) + Z_1(t, T^*) = \tilde{\kappa}(t, T_1^*) + \delta_n^{-1}.$$

Differentiating both sides of the last equality, we get (cf. (3.12) and (3.13))

$$Z_1(t, T_1^*) \gamma_1(t, T_1^*) = \tilde{\kappa}(t, T_1^*) \nu(t, T_1^*).$$

Consequently, $\tilde{W}^{T_1^*}$ is explicitly given by the formula

$$\tilde{W}_t^{T_1^*} = \tilde{W}_t^{T^*} - \int_0^t \frac{\delta_{n-1} \tilde{\kappa}(u, T_1^*)}{1 + \delta_{n-1} \delta_n^{-1} + \delta_{n-1} \tilde{\kappa}(u, T_1^*)} \nu(u, T_1^*) du$$

for $t \in [0, T_1^*]$. We are in a position to define, using Girsanov's theorem, the associated forward swap measure $\tilde{\mathbb{P}}_{T_1^*}$. Subsequently, we introduce the process $\tilde{\kappa}(\cdot, T_2^*)$, by postulating that it solves the SDE

$$d\tau \kappa(t, T_2^*) = \tilde{\kappa}(t, T_2^*) \nu(t, T_2^*) \cdot d\tilde{W}_t^{T_1^*}$$

with the initial condition

$$\tilde{\kappa}(0, T_2^*) = \frac{B(0, T_2^*) - B(0, T^*)}{\delta_{n-1} B(0, T_1^*) + \delta_n B(0, T^*)}.$$

For the reader's convenience, let us consider one more inductive step, in which we are looking for $\tilde{\kappa}(t, T_3^*)$. We now consider processes

$$Z_3(t, T_k^*) = \frac{B(t, T_k^*)}{\delta_{n-2} B(t, T_2^*) + \delta_{n-1} B(t, T_1^*) + \delta_n B(t, T^*)} = \frac{Z_2(t, T_k^*)}{1 + \delta_{n-2} Z_2(t, T_2^*)},$$

so that

$$\tilde{W}_t^{T_2^*} = \tilde{W}_t^{T_1^*} - \int_0^t \frac{\delta_{n-2} Z_2(u, T_2^*)}{1 + \delta_{n-2} Z_2(u, T_2^*)} \gamma_2(u, T_2^*) du$$

for $t \in [0, T_2^*]$. It is useful to note that

$$Z_2(t, T_2^*) = \frac{B(t, T_2^*)}{\delta_{n-1} B(t, T_1^*) + \delta_n B(t, T^*)} = \tilde{\kappa}(t, T_2^*) + Z_2(t, T^*),$$

where in turn

$$Z_2(t, T^*) = \frac{Z_1(t, T^*)}{1 + \delta_{n-1} Z_1(t, T^*) + \delta_{n-1} \tilde{\kappa}(t, T_1^*)}$$

and the process $Z_1(\cdot, T^*)$ is already known from the previous step (clearly, $Z_1(\cdot, T^*) = 1/d_n$). Differentiating the last equality, we may thus find the volatility of the process $Z_2(\cdot, T^*)$, and consequently, define $\tilde{\mathbb{P}}_{T_2^*}$.

We now examine the general case. We proceed by induction with respect to m . Suppose that we have found forward swap rates $\tilde{\kappa}(\cdot, T_1^*), \dots, \tilde{\kappa}(\cdot, T_m^*)$, the forward swap measure $\tilde{\mathbb{P}}_{T_{m-1}^*}$ and the associated Brownian motion $\tilde{W}^{T_{m-1}^*}$. Our aim is to determine the forward swap measure $\tilde{\mathbb{P}}_{T_m^*}$, the associated Brownian motion $\tilde{W}^{T_m^*}$, and the forward swap rate $\tilde{\kappa}(\cdot, T_{m+1}^*)$. To this end, we postulate that processes

$$\begin{aligned} Z_{m+1}(t, T_k^*) &= \frac{B(t, T_k^*)}{G_t(m+1)} = \frac{B(t, T_k^*)}{\delta_{n-m}B(t, T_m^*) + \dots + \delta_n B(t, T^*)} \\ &= \frac{Z_m(t, T_k^*)}{1 + \delta_{n-m}Z_m(t, T_m^*)} \end{aligned}$$

follow local martingales under $\tilde{\mathbb{P}}_{T_m^*}$. In view of Lemma 2.3, applied to processes $G = Z_m(\cdot, T_k^*)$ and $H = Z_m(\cdot, T_m^*)$, it is clear that we may set

$$\tilde{W}_t^{T_m^*} = \tilde{W}_t^{T^*} - \int_0^t \frac{\delta_{n-m}Z_m(u, T_m^*)}{1 + \delta_{n-m}Z_m(u, T_m^*)} \gamma_m(u, T_m^*) du, \quad (3.14)$$

for $t \in [0, T_m^*]$. Therefore it is sufficient to analyse the process

$$Z_m(t, T_m^*) = \frac{B(t, T_m^*)}{\delta_{n-m+1}B(t, T_{m-1}^*) + \dots + \delta_n B(t, T^*)} = \tilde{\kappa}(t, T_m^*) + Z_m(t, T^*).$$

To conclude, it is enough to notice that

$$Z_m(t, T^*) = \frac{Z_{m-1}(t, T^*)}{1 + \delta_{n-m+1}Z_{m-1}(t, T^*) + \delta_{n-m+1}\tilde{\kappa}(t, T_{m-1}^*)}.$$

Indeed, from the preceding step, we know that the process $Z_{m-1}(\cdot, T^*)$ is a (rational) function of forward swap rates $\tilde{\kappa}(\cdot, T_1^*), \dots, \tilde{\kappa}(\cdot, T_{m-1}^*)$. Consequently, the process under the integral sign on the right-hand side of (3.14) can be expressed using the terms $\tilde{\kappa}(\cdot, T_1^*), \dots, \tilde{\kappa}(\cdot, T_{m-1}^*)$ and their volatilities (since the explicit formula is rather lengthy, it is not reported here). Having found the process $\tilde{W}_t^{T_m^*}$ and probability measure $\tilde{\mathbb{P}}_{T_m^*}$, we introduce the forward swap rate $\tilde{\kappa}(\cdot, T_{m+1}^*)$ through (3.10)–(3.11), and so forth. If all volatilities are deterministic, the model is termed the *lognormal model* of fixed-maturity forward swap rates.

3.3 Valuation of swaptions

For a long time, Black's swaptions formula was merely a (widely used) practical tool to value swaptions. Indeed, the use of this formula was not supported by the existence of a reliable term structure model. Valuation and hedging of swaptions based on the suitable version of Black's formula was analysed, for instance, in Neuberger (1990). The formal derivation of this heuristic results within the framework of a well established term structure model was first achieved in Jamshidian (1997).

3.3.1 Payer and receiver swaptions

The owner of a *payer* (*receiver*, respectively) *swaption* with strike rate κ , maturing at time $T = T_0$, has the right to enter at time T the underlying forward payer (*receiver*, respectively) swap settled in arrears.¹⁸ Because $\mathbf{FS}_T(\kappa)$ is the value at time T of the payer swap with the fixed interest rate κ , it is clear that the price of the payer swaption at time t equals

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} (\mathbf{FS}_T(\kappa))^+ \mid \mathcal{F}_t \right\}.$$

Using (3.3), we obtain

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (L(T_{j-1}) - \kappa) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\}. \quad (3.15)$$

On the other hand, in view of (3.7) we also have

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa(T, T, n) - \kappa) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\} \quad (3.16)$$

The last equality yields

$$\begin{aligned} \mathbf{PS}_t &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa(T, T, n) - \kappa) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa(T, T, n) - \kappa)^+ \delta_j \mid \mathcal{F}_T \right) \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \sum_{j=1}^n \delta_j B(T, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} ((\kappa(T, T, n) - \kappa)^+ \mid \mathcal{F}_T) \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \sum_{j=1}^n \delta_j B(T, T_j) (\kappa(T, T, n) - \kappa)^+ \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(1 - \sum_{j=1}^n c_j B(T, T_j) \right)^+ \mid \mathcal{F}_t \right\}. \end{aligned}$$

Similarly, for the receiver swaption, we have

$$\mathbf{RS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} (-\mathbf{FS}_T(\kappa))^+ \mid \mathcal{F}_t \right\},$$

¹⁸ By convention, the notional principal of the underlying swap (and thus also the notional principal of the swaption) equals $N_p = 1$.

that is

$$\mathbf{RS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa - L(T_{j-1})) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\}, \quad (3.17)$$

where we write $\mathbf{RS}_t$ to denote the price at time t of a receiver swaption. Consequently, reasoning in much the same way as in the case of a payer swaption, we get

$$\begin{aligned} \mathbf{RS}_t &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa - \kappa(T, T, n)) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa - \kappa(T, T, n))^+ \delta_j \mid \mathcal{F}_T \right) \mid \mathcal{F}_t \right\} \\ &= \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\sum_{j=1}^n c_j B(T, T_j) - 1 \right)^+ \mid \mathcal{F}_t \right\}. \end{aligned}$$

We shall first focus on a payer swaption. In view of (3.15), it is apparent that a payer swaption is exercised at time T if and only if the value of the underlying swap is positive at this date. It should be made clear that a swaption may be exercised by its owner only at its maturity date T . If exercised, a swaption gives rise to a sequence of cash flows at prescribed future dates. By considering the future cash flows from a swaption and from the corresponding *market swap*¹⁹ available at time T , it is easily seen that the owner of a swaption is protected against the adverse movements of the swap rate that may occur before time T . Suppose, for instance, that the swap rate at time T is greater than κ . Then by combining the swaption with a market swap, the owner of a swaption with exercise rate κ is entitled to enter at time T , at no additional cost, a swap contract in which the fixed rate is κ . If, on the contrary, the swap rate at time T is less than κ , the swaption is worthless, but its owner is, of course, able to enter a market swap contract based on the current swap rate $\kappa(T, T, n) \leq \kappa$. Concluding, the fixed rate paid by the owner of a swaption who intends to initiate a swap contract at time T will never be above the preassigned level κ .

Notice that we that we have shown, in particular, that

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (\kappa(T, T, n) - \kappa)^+ \delta_j \mid \mathcal{F}_T \right) \mid \mathcal{F}_t \right\}. \quad (3.18)$$

This shows that a payer swaption is essentially equivalent to a sequence of fixed payments $d_j^p = \delta_j (\kappa(T, T, n) - \kappa)^+$ which are received at settlement dates

¹⁹ At any time t , a *market swap* is that swap whose current value equals zero. Put more explicitly, it is the swap in which the fixed rate κ equals the current swap rate.

$T_1, \dots, T_n$, but whose value is known already at the expiry date T . In words, a payer swaption can be seen as a specific call option on a forward swap rate, with fixed strike level κ . The exercise date of the option is T , but the payoff takes place at each date $T_1, \dots, T_n$. This equivalence may also be derived by directly verifying that the future cash flows from the following portfolios established at time T are identical: portfolio A – a swaption and a market swap; and portfolio B – a just described call option on a swap rate and a market swap. Indeed, both portfolios correspond to a payer swap with the fixed rate equal to κ .

Finally, the equality

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(1 - \sum_{j=1}^n c_j B(T, T_j) \right)^+ \middle| \mathcal{F}_t \right\} \quad (3.19)$$

shows that the payer swaption may also be seen as a standard put option on a coupon-bearing bond with the coupon rate κ , with exercise date T and strike price 1.

Similar remarks are valid for the receiver swaption. In particular, a receiver swaption can also be viewed as a sequence of put options on a swap rate which are not allowed to be exercised separately. At time T the long party receives the value of a sequence of cash flows, discounted from time T_j , $j = 1, \dots, n$, to the date T , defined by $\delta_j (\kappa - \kappa(T, T, n))^+$. On the other hand, a receiver swaption may be seen as a call option, with strike price 1 and expiry date T , written on a coupon bond with coupon rate equal to the strike rate κ of the underlying forward swap.

Let us finally mention the put–call parity relationship for swaptions. It follows easily from (3.15)–(3.17) that $\mathbf{PS}_t - \mathbf{RS}_t = \mathbf{FS}_t$, i.e.,

$$\text{payer swaption}(t) - \text{receiver swaption}(t) = \text{forward swap}(t)$$

provided that both swaptions expire at the same date T (and have the same contractual features).

3.3.2 Forward swaptions

Let us now consider a *forward swaption*. In this case, we assume that the expiry date $\hat{T}$ of the swaption precedes the initiation date T of the underlying payer swap – that is, $\hat{T} \leq T$. Recall that

$$\mathbf{FS}_t(\kappa) = \sum_{j=1}^n (\kappa(t, T, n) - \kappa) B(t, T_j)$$

for $t \in [0, T]$. It is thus clear that the payoff $\mathbf{PS}_{\hat{T}}$ at expiry $\hat{T}$ of the forward swaption (with strike 0) is either 0, if $\kappa \geq \kappa(\hat{T}, T, n)$, or

$$\mathbf{PS}_{\hat{T}} = \sum_{j=1}^n (\kappa(\hat{T}, T, n) - \kappa) B(\hat{T}, T_j)$$

if, on the contrary, inequality $\kappa(\hat{T}, T, n) > \kappa$ holds. We conclude that the payoff $\mathbf{PS}_{\hat{T}}$ of the forward swaption can be represented in the following way:

$$\mathbf{PS}_{\hat{T}} = \sum_{j=1}^n (\kappa(\hat{T}, T, n) - \kappa)^+ B(\hat{T}, T_j). \quad (3.20)$$

This means that, if exercised, the forward swaption gives rise to a sequence of equal payments $\kappa(\hat{T}, T, n) - \kappa$ at each settlement date $T_1, \dots, T_n$. By substituting $\hat{T} = T$ we recover, in a more intuitive way and in a more general setting, the previously observed dual nature of the swaption: it may be seen either as an option on the value of a particular (forward) swap or, equivalently, as an option on the corresponding (forward) swap rate. It is also clear that the owner of a forward swaption is able to enter at time $\hat{T}$ (at no additional cost) into a forward payer swap with preassigned fixed interest rate κ .

3.3.3 Valuation in the lognormal model of forward Libor rates

Recall that within the general framework, the price at time $t \in [0, T_0]$ of a payer swaption²⁰ with expiry date $T = T_0$ and strike level κ equals

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} \left(\mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (L(T_{j-1}) - \kappa) \delta_j \mid \mathcal{F}_T \right) \right)^+ \mid \mathcal{F}_t \right\}.$$

Let $D \in \mathcal{F}_T$ be the exercise set of a swaption; that is

$$D = \{\omega \in \Omega \mid (\kappa(T, T, n) - \kappa)^+ > 0\} = \{\omega \in \Omega \mid \sum_{j=1}^n c_j B(T, T_j) < 1\}.$$

Lemma 3.3 *The following equality holds for every $t \in [0, T]$:*

$$\mathbf{PS}_t = \sum_{j=1}^n \delta_j B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} \left((L(T, T_{j-1}) - \kappa) I_D \mid \mathcal{F}_t \right). \quad (3.21)$$

Proof Since

$$\mathbf{PS}_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \frac{B_t}{B_T} I_D \mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_T}{B_{T_j}} (L(T_{j-1}) - \kappa) \delta_j \mid \mathcal{F}_T \right) \mid \mathcal{F}_t \right\},$$

²⁰ Since the relationship $\mathbf{PS}_t - \mathbf{RS}_t = \mathbf{FS}_t$ is always valid, and the value of a forward swap is given by (3.4), it is enough to examine the case of a payer swaption.

we have

$$\begin{aligned} \mathbf{PS}_t &= \mathbb{E}_{\mathbb{P}^*} \left\{ \mathbb{E}_{\mathbb{P}^*} \left(\sum_{j=1}^n \frac{B_t}{B_{T_j}} (L(T_{j-1}) - \kappa) \delta_j \mathbf{I}_D \mid \mathcal{F}_T \right) \mid \mathcal{F}_t \right\} \\ &= \sum_{j=1}^n B(t, T_j) \mathbb{E}_{\mathbb{P}_{T_j}} \left((L(T_{j-1}) - \kappa) \delta_j \mathbf{I}_D \mid \mathcal{F}_t \right), \end{aligned}$$

where $L(T_{j-1}) = L(T_{j-1}, T_{j-1})$. For any $j = 1, \dots, n$, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_{T_j}} \left((L(T_{j-1}) - \kappa) \mathbf{I}_D \mid \mathcal{F}_t \right) &= \mathbb{E}_{\mathbb{P}_{T_j}} \left(\mathbb{E}_{\mathbb{P}_{T_j}} (L(T_{j-1}) - \kappa \mid \mathcal{F}_T) \mathbf{I}_D \mid \mathcal{F}_t \right) \\ &= \mathbb{E}_{\mathbb{P}_{T_j}} \left((L(T, T_{j-1}) - \kappa) \mathbf{I}_D \mid \mathcal{F}_t \right), \end{aligned}$$

since $\mathcal{F}_t \subset \mathcal{F}_T$ and the process $L(t, T_{j-1})$ is a $\mathbb{P}_{T_j}$ -martingale. $\square$

For any $k = 1, \dots, n$, we define the random variable $\zeta_k(t)$ by setting

$$\zeta_k(t) = \int_t^T \lambda(u, T_{k-1}) \cdot dW_u^{T_k}, \quad \forall t \in [0, T], \quad (3.22)$$

and we write

$$\lambda_k^2(t) = \int_t^T |\lambda(u, T_{k-1})|^2 du, \quad \forall t \in [0, T]. \quad (3.23)$$

Note that for every $k = 1, \dots, n$ and $t \in [0, T]$, we have

$$L(T, T_{k-1}) = L(t, T_{k-1}) e^{\zeta_k(t) - \lambda_k^2(t)/2}.$$

Recall also that the processes W^{T_k} satisfy the following relationship:

$$W_t^{T_{k+1}} = W_t^{T_k} + \int_0^t \frac{\delta_{k+1} L(u, T_k)}{1 + \delta_{k+1} L(u, T_k)} \lambda(u, T_k) du$$

for $t \in [0, T_k]$ and $k = 0, \dots, n-1$. For ease of notation, we formulate the next result for $t = 0$ only; a general case can be treated along the same lines. For any fixed j , we denote by G_j the joint probability distribution function of the n -dimensional random variable $(\zeta_1(0), \dots, \zeta_n(0))$ under the forward measure $\mathbb{P}_{T_j}$.

Proposition 3.4 *Assume the lognormal model of Libor rates. The price at time 0 of a payer swaption with expiry date $T = T_0$ and strike level κ equals*

$$\mathbf{PS}_0 = \sum_{j=1}^n \delta_j B(0, T_j) \int_{\mathbb{R}^n} \left(L(0, T_{j-1}) e^{y_j - \lambda_j^2(0)/2} - \kappa \right) \mathbf{I}_{\bar{D}} dG_j(y_1, \dots, y_n),$$

where $I_{\tilde{D}} = I_{\tilde{D}}(y_1, \dots, y_n)$, and $\tilde{D}$ stands for the set

$$\tilde{D} = \left\{ (y_1, \dots, y_n) \in \mathbb{R}^n \mid \sum_{j=1}^n c_j \prod_{k=1}^j \left(1 + \delta_k L(0, T_{k-1}) e^{y_k - \lambda_k^2(0)/2} \right)^{-1} < 1 \right\}.$$

Proof Let us start by considering arbitrary $t \in [0, T]$. Notice that

$$\frac{B(t, T_j)}{B(t, T)} = \prod_{k=1}^j \frac{B(t, T_k)}{B(t, T_{k-1})} = \prod_{k=1}^j (F_B(t, T_{k-1}, T_k))^{-1},$$

and thus, in view of (2.12), we have

$$B(T, T_j) = \prod_{k=1}^j (1 + \delta_k L(T, T_{k-1}))^{-1}.$$

Consequently, the exercise set D can be re-expressed in terms of forward Libor rates. Indeed, we have

$$D = \left\{ \omega \in \Omega \mid \sum_{j=1}^n c_j \prod_{k=1}^j (1 + \delta_k L(T, T_{k-1}))^{-1} < 1 \right\},$$

or more explicitly

$$D = \left\{ \omega \in \Omega \mid \sum_{j=1}^n c_j \prod_{k=1}^j \left(1 + \delta_k L(t, T_{k-1}) e^{\zeta_k(t) - \lambda_k^2(t)/2} \right)^{-1} < 1 \right\}.$$

Let us put $t = 0$. In view of Lemma 3.3, to find the arbitrage price of a swaption at time 0, it is sufficient to determine the joint law under the forward measure $\mathbb{P}_{T_j}$ of the random variable $(\zeta_1(0), \dots, \zeta_n(0))$, where $\zeta_1(0), \dots, \zeta_n(0)$ are given by (3.22). Note also that

$$D = \left\{ \omega \in \Omega \mid \sum_{j=1}^n c_j \prod_{k=1}^j \left(1 + \delta_k L(0, T_{k-1}) e^{\zeta_k(0) - \lambda_k^2(0)/2} \right)^{-1} < 1 \right\}.$$

This shows the validity of the valuation formula for $t = 0$. It is clear that it admits a rather straightforward generalization to arbitrary $0 < t \leq T$. $\square$

3.3.4 Market valuation formula for swaptions

The commonly used formula for pricing swaptions, based on the assumption that the underlying swap rate follows a geometric Brownian motion under the intuitively perceived “market probability” $\mathbb{Q}$, is given by *Black’s swaption formula* (see Neuberger (1990))

$$\mathbf{PS}_t = \sum_{j=1}^n B(t, T_j) \delta_j \left(\kappa(t, T, n) N(h_1(t, T)) - \kappa N(h_2(t, T)) \right), \quad (3.24)$$

where $T = T_0$ is the swaption's expiry date, and

$$h_{1,2}(t, T) = \frac{\ln(\kappa(t, T, n)/\kappa) \pm \frac{1}{2} \sigma^2 (T - t)}{\sigma \sqrt{T - t}}$$

for some constant $\sigma > 0$. To examine formula (3.24) in an intuitive way, let us assume, for simplicity, that $t = 0$. In this case, using general valuation results, we obtain the following equality

$$\mathbf{PS}_0 = \sum_{j=1}^n \delta_j B(0, T_j) \mathbb{E}_{\mathbb{P}_{T_j}}((\kappa(T, T, n) - \kappa)^+).$$

Apparently, market practitioners assume a lognormal probability law for the swap rate $\kappa(T, T, n)$ under $\mathbb{P}_{T_j}$. The swaption valuation formula obtained in the framework of the lognormal model of Libor rates appears to be more involved. It reduces to the “market formula” (3.24) only in very special circumstances. On the other hand, the swaption price derived within the lognormal model of forward swap rates (see Section 3.2 below) agrees with (3.24). More precisely, this holds for a specific family of swaptions. This is by no means surprising, as the model was exactly tailored to handle a particular family of swaptions, or rather, to analyse certain path-dependent swaptions (such as *Bermudan swaptions*). The price of a cap in the lognormal model of swap rates is not given by a closed-form expression, however.

3.3.5 Valuation in the lognormal model of forward swap rates

For a fixed, but otherwise arbitrary, date T_j , $j = 0, \dots, n - 1$, we consider a swaption with expiry date T_j , written on a forward payer swap settled in arrears. The underlying forward payer swap starts at date T_j , has the fixed rate κ and $n - j$ accrual periods. Such a swaption is referred to as the j^{th} swaption in what follows. Notice that the j^{th} swaption can be seen as a contract which pays to its owner the amount $\delta_k(\kappa(T_j, T_j, n - j) - \kappa)^+$ at each settlement date T_k , where $k = j + 1, \dots, n$ (recall that we assume that the notional principal $N_p = 1$). Equivalently, the j^{th} swaption pays an amount

$$\tilde{Y} = \sum_{k=j+1}^n \delta_k B(T_j, T_k) (\tilde{\kappa}(T_j, T_j) - \kappa)^+$$

at maturity date T_j . It is useful to observe that $\tilde{Y}$ admits the following representation in terms of the numeraire process $G(n - j)$ introduced in Section 3.2 (cf. formula (3.9))

$$\tilde{Y} = G_{T_j}(n - j) (\tilde{\kappa}(T_j, T_j) - \kappa)^+.$$

Recall that the model of fixed-maturity forward swap rates presented in Section 3.2 specifies the dynamics of the process $\tilde{\kappa}(\cdot, T_j)$ through the following SDE:

$$d\tau\kappa(t, T_j) = \tilde{\kappa}(t, T_j)v(t, T_j) \cdot d\tilde{W}_t^{T_{j+1}},$$

where $\tilde{W}^{T_{j+1}}$ follows a standard d -dimensional Brownian motion under the corresponding forward swap measure $\tilde{\mathbb{P}}_{T_{j+1}}$. Recall that the definition of $\tilde{\mathbb{P}}_{T_{j+1}}$ implies that any process of the form $B(t, T_k)/G_t(n-j)$, $k = 0, \dots, n$, is a local martingale under $\tilde{\mathbb{P}}_{T_{j+1}}$. Furthermore, from the general considerations concerning the choice of a numeraire (see, e.g. Geman *et al.* (1995) or Musiela and Rutkowski (1997a)) it is easy to see that the arbitrage price $\pi_t(X)$ of an attainable contingent claim $X = g(B(T_j, T_{j+1}), \dots, B(T_j, T_n))$ equals, for $t \in [0, T_j]$,

$$\pi_t(X) = G_t(n-j) \mathbb{E}_{\tilde{\mathbb{P}}_{T_{j+1}}} (G_{T_j}^{-1}(n-j)X \mid \mathcal{F}_t),$$

provided that X settles at time T_j . Applying the last formula to the swaption's payoff $\tilde{Y}$, we obtain the following representation for the arbitrage price $\mathbf{PS}_t^j$ at time $t \in [0, T_j]$ of the j^{th} swaption:

$$\mathbf{PS}_t^j = \pi_t(\tilde{Y}) = G_t(n-j) \mathbb{E}_{\tilde{\mathbb{P}}_{T_{j+1}}} ((\tilde{\kappa}(T_j, T_j) - \kappa)^+ \mid \mathcal{F}_t).$$

We assume from now on that $v(\cdot, T_j) : [0, T_j] \rightarrow \mathbb{R}^d$ is a bounded deterministic function. In other words, we place ourselves within the framework of the lognormal model of fixed-maturity forward swap rates. The proof of following result, due to Jamshidian (1996, 1997), is straightforward.

Proposition 3.5 *For any $j = 1, \dots, n-1$, the arbitrage price at time $t \in [0, T_j]$ of the j^{th} swaption equals*

$$\mathbf{PS}_t^j = \sum_{k=j+1}^n \delta_k B(t, T_k) \left(\tilde{\kappa}(t, T_j) N(\tilde{h}_1(t, T_j)) - \kappa N(\tilde{h}_2(t, T_j)) \right),$$

where N denotes the standard Gaussian cumulative distribution function, and

$$\tilde{h}_{1,2}(t, T_j) = \frac{\ln(\tilde{\kappa}(t, T_j)/\kappa) \pm \frac{1}{2} v^2(t, T_j)}{v(t, T_j)},$$

with $v^2(t, T_j) = \int_t^{T_j} |v(u, T_j)|^2 du$.

Proof The proof of the proposition is quite similar to that of Proposition 2.9 and thus it is omitted. $\square$

3.3.6 Hedging of swaptions

The replicating strategy for a swaption within the present framework has similar features to the replicating strategy for a cap in the lognormal model of forward Libor rates. Therefore, we shall focus mainly on differences between these two cases. Let us fix j , and let us denote by $F_{Sj}(t, T)$ the relative price at time $t \leq T_j$ of the j^{th} swaption, when the value process

$$G_t(n-j) = \sum_{k=j+1}^n \delta_k B(t, T_k)$$

is chosen as a numeraire asset. From Proposition 3.5, we find easily that for every $t \leq T_j$

$$F_{Sj}(t, T_j) = \tilde{\kappa}(t, T_j) N(\tilde{h}_1(t, T_j)) - \kappa N(\tilde{h}_2(t, T_j)).$$

Applying Itô's formula to the last expression, we obtain

$$dF_{Sj}(t, T_j) = N(\tilde{h}_1(t, T_j)) d\tau\kappa(t, T_j). \quad (3.25)$$

Let us consider the following self-financing trading strategy. We start our trade at time 0 with the amount $\mathbf{PS}_0^j$ of cash, which is then immediately invested in the portfolio $G(n-j)$.²¹ At any time $t \leq T_j$ we assume $\psi_t^j = N(h_1(t, T_j))$ positions in market forward swaps (of course, these swaps have the same starting date and tenor structure as the underlying forward swap). The associated gains/losses process V , expressed in units of the numeraire asset $G(n-j)$, satisfies

$$dV_t = \psi_t^j d\tau\kappa(t, T_j) = N(\tilde{h}_1(t, T_j)) d\tau\kappa(t, T_j) = dF_{Sj}(t, T_j)$$

with $V_0 = 0$. Consequently,

$$F_{Sj}(T_j, T_j) = F_{Sj}(0, T_j) + \int_0^{T_j} \psi_t^j d\tau\kappa(t, T_j) = F_{Sj}(0, T_j) + V_{T_j}.$$

Here the dynamic trading in market forward swaps takes place at any date $t \in [0, T_j]$, and all gains/losses from trading (involving the initial investment) are expressed in units of $G(n-j)$. The last equality makes it clear that the strategy ψ^j introduced above does indeed replicate the j^{th} swaption.

3.4 Choice of numeraire portfolio

Let us summarize briefly the theoretic results which underpin the recent approaches to term structure modelling. For the reader's convenience, we shall restrict our attention here to the case of bond portfolios.

²¹ One unit of portfolio $G(n-j)$ costs $\sum_{k=j+1}^n \delta_k B(0, T_k)$ at time 0.

Let us consider two particular portfolio of zero-coupon bonds, with value processes V_t^1 and V_t^2 . Typically, we are interested in options to exchange one of this portfolios for another, at a given date T . Let us write

$$C_T = (V_T^1 - K V_T^2) = V_T^1 \mathbb{1}_D - K V_T^2 \mathbb{1}_D, \quad (3.26)$$

where $K > 0$ is a constant, and $D = \{V_T^1 > K V_T^2\}$ is the exercise set. It is easy to check using the abstract Bayes rule that the equality

$$\frac{d\mathbb{P}^1}{d\mathbb{P}^2} = \frac{V_0^2}{V_0^1} \frac{V_T^1}{V_T^2}, \quad \mathbb{P}^2\text{-a.s.}, \quad (3.27)$$

links the martingale measures $\mathbb{P}^1$ and $\mathbb{P}^2$ associated with the choice of value processes V^1 and V^2 as discount factors, respectively (both probability measures are considered here on $(\Omega, \mathcal{F}_T)$). Furthermore, the arbitrage price of the option admits the following representation

$$C_t = V_t^1 \mathbb{P}^1(D | \mathcal{F}_t) - K V_t^2 \mathbb{P}^2(D | \mathcal{F}_t), \quad \forall t \in [0, T], \quad (3.28)$$

where $D = \{V_T^1 > K V_T^2\}$. To obtain the Black–Scholes-like formula for the option's price C_t , it is enough to assume that the relative price V^1/V^2 follows a lognormal martingale under $\mathbb{P}^2$, so that

$$d(V_t^1/V_t^2) = (V_t^1/V_t^2) \gamma_t^{1,2} \cdot dW_t^{1,2} \quad (3.29)$$

for a deterministic function $\gamma^{1,2} : [0, T] \rightarrow \mathbb{R}^d$ (for simplicity, we also assume that the function $\gamma^{1,2}$ is bounded). In view of (3.27), the Radon–Nikodým density of $\mathbb{P}^1$ with respect to $\mathbb{P}^2$ equals

$$\frac{d\mathbb{P}^1}{d\mathbb{P}^2} = \mathcal{E}_T \left(\int_0^\cdot \gamma_u^{1,2} \cdot dW_u^{1,2} \right), \quad \mathbb{P}^2\text{-a.s.}, \quad (3.30)$$

and thus the process

$$W_t^{2,1} = W_t^{1,2} - \int_0^t \gamma_u^{1,2} du, \quad \forall t \in [0, T],$$

is a standard Brownian motion under $\mathbb{P}^2$. Reasoning in the much the same way as in the proof of the classic Black–Scholes formula (see, for instance, the proof of Theorem 5.1.1 in Musiela and Rutkowski (1997a)), we obtain

$$C_t = V_t^1 N(d_1(t, T)) - K V_t^2 N(d_2(t, T)), \quad (3.31)$$

where

$$d_{1,2}(t, T) = \frac{\ln(V_t^1/V_t^2) - \ln K \pm \frac{1}{2} v_{1,2}^2(t, T)}{v_{1,2}(t, T)}$$

and

$$v_{1,2}^2(t, T) = \int_t^T |\gamma_u^{1,2}|^2 du, \quad \forall t \in [0, T].$$

Of course, the caps and swaptions²² valuation formulae in lognormal models described above can be seen as special cases of (3.31). The idea can be, of course, applied to other interest rate derivatives.

It is worthwhile noting that in order to get the valuation result (3.31) for $t = 0$, it is enough to assume that the random variable V_T^1 / V_T^2 has a lognormal probability law under the martingale measure $\mathbb{P}^2$. This simple observation underpins the construction of the so-called *Markov-functional interest rate models* – this alternative approach to term structure modelling is briefly reviewed in the next section.

A more straightforward generalization of lognormal models of the term structure was developed by Andersen and Andreasen (1997). In this case, the assumption that the volatility is deterministic is replaced by a suitable functional form of the volatility. The resulting models are capable of handling the so-called *volatility skew* in observed option prices (empirical studies have shown that the implied volatilities of observed caps and swaptions prices tend to be decreasing functions of the strike level). The main focus in Andersen and Andreasen (1997) is on the use of the CEV process²³ as a model of the forward Libor rate. Put more explicitly, they generalize equality (2.20) by postulating that

$$dL(t, T_j) = L^\alpha(t, T_j) \lambda(t, T_j) \cdot dW_t^{T_{j+1}}, \quad \forall t \in [0, T_j],$$

where $\alpha > 0$ is a strictly positive constant. They derive closed-form solutions for caplet prices under the above specification of the dynamics of Libor rates with $\alpha \neq 1$, in terms of the cumulative distribution function of a non-central χ^2 probability law. It appears that, depending on the choice of the parameter α , the implied Black's volatilities of caplet prices, considered as a function of the strike level $\kappa > 0$, exhibit downward- or upward-sloping skew.

4 Markov-functional models

As shown in Section 2.2.4, the forward Libor or swap²⁴ rates follow a multi-dimensional Markov process under any of the associated forward measures. In principle, lognormal models can be easily calibrated to market prices of caps (or

²² For the j^{th} caplet, we take $V_t^1 = B(t, T_j) - B(t, T_{j+1})$ and $V_t^2 = \delta_{j+1} B(t, T_{j+1})$. In the case of the j^{th} swaption, we have $V_t^1 = B(t, T_j) - B(t, T_n)$ and $V_t^2 = \sum_{k=j+1}^n \delta_k B(t, T_k)$.

²³ In the context of equity options, the CEV (*constant elasticity of variance*) process was first introduced in Cox and Ross (1976).

²⁴ The multi-dimensional SDE which governs the dynamics of the family of forward swap rates is more involved than the SDE for the family of Libor rates, and thus it is not reported here. The interested reader is referred to Jamshidian (1997).

swaptions), which is, of course, a nice feature of this class of term structure models, as opposed to the classic models based on the specification of the dynamics of (spot or forward) instantaneous rates. On the other hand, however, due to the high dimensionality of the underlying Markov process, the efficient implementation of these models appears to be rather difficult.

To circumvent this obstacle, an alternative approach was recently developed in a series of papers by Hunt and Kennedy (1997, 1998) and Hunt *et al.* (1996, 2000).²⁵ It is based on the introduction of a low-dimensional Markov process which (by assumption) governs, through a simple functional dependence, the dynamics of all other relevant stochastic processes. For this reason, these class of term structure models is referred to as *Markov-functional interest rate models*. In economical interpretation, the underlying Markov process is assumed to represent the state of the economy; it is thus justified to refer to its components as “state variables”.

Formally, one starts by introducing a one- or multi-dimensional process M , which possesses the Markov property under the *terminal measure*, where the generic term *terminal measure* is intended to cover not only cases considered in previous sections, but also other suitable choices of the numeraire portfolio. As already mentioned, the relevant processes, such as in particular the value process of the numeraire portfolio and zero-coupon bond prices, are assumed to be functions of M . For instance, if $T^* > 0$ is the horizon date, then for any $t \leq s \leq T$ we have

$$\frac{B(t, T, M_t)}{V_t(M_t)} = \mathbb{E}_{\hat{\mathbb{P}}} \left(\frac{B(s, T, M_s)}{V_s(M_s)} \middle| \mathcal{F}_t \right),$$

where $V_t(M_t)$, $t \leq T^*$, is the value process of the numeraire portfolio, and $\hat{\mathbb{P}}$ is the associated martingale measure. The notation $B(t, T, M_t)$ emphasizes the direct dependence of the bond price on time variables, t and T , as well as on the state variable represented by the random variable M_t . Note that the functional form of $B(t, T, M_t)$ is not explicitly known, except for some very special choices of dates t and T . In some instances, it may appear convenient to postulate that²⁶

$$\frac{B(T, S, M_T)}{V_T(M_T)} = A + B(S)M_T$$

and to derive further properties from the martingale feature of relative prices. In the next section, we shall present a particular example of such an approach, in which we focus on the derivation of a simple formula for the so-called *convexity correction*. Then, in Section 4.2, we shall discuss the problem of calibration of the Markov-functional model.

²⁵ We present here only few examples of their approach. The interested reader is referred to the original papers and to Hunt and Kennedy (2000) for a more detailed account.

²⁶ See Hunt *et al.* (1996) for alternative kinds of the functional dependence, including *exponential* and *geometric*.

4.1 Terminal swap rate model

The terminal swap rate model – put forward by Hunt *et al.* (1996) – was primarily designed for the purpose of the comparative pricing of non-standard swap contracts *vis-à-vis* plain vanilla swaps (informally, this is referred to as *convexity correction*; see Schmidt (1996)). Let us consider, as usual, a given collection of reset/settlement dates $T_0, \dots, T_n$. We assume that the market price at time 0 of the (plain vanilla) fixed-for-floating swaption is known. We postulate, in addition, that it is given by Black's formula for swaptions. Let us consider the family of bond prices $B(T, S)$, where the maturity date $S \geq T$ belongs to some set $\mathcal{S}$ of dates. We postulate that there exist constants A and B_S such that for any $S \in \mathcal{S}$

$$D(T, S) := B(T, S)G_T^{-1}(n) = A + B_S\kappa(T, T, n), \quad (4.1)$$

where $G_t(n) = \sum_{j=1}^n \delta_j B(t, T_j)$, and (cf. (3.8))

$$\kappa(t, T, n) = \frac{B(t, T) - B(t, T_n)}{\delta_1 B(t, T_1) + \dots + \delta_n B(t, T_n)} = \frac{B(t, T) - B(t, T_n)}{G_t(n)}.$$

Using the martingale property of discounted bond price $D(\cdot, S)$ and forward swap rate $\kappa(\cdot, T, n)$ under the corresponding forward swap measure associated with the choice of $G(n)$ as a numeraire, we get

$$D(t, S) = A + B_S\kappa(t, T, n),$$

or equivalently

$$B(t, S) = A(1 - B(t, T_n)) + B_S G_t(n)$$

for every $t \in [0, T]$. We thus see that condition (4.1) is rather stringent; it implies that the price of any bond of maturity S from $\mathcal{S}$ can be represented as a linear combination of values of two particular portfolios of bonds, with one coefficient independent of maturity date S . The problem of whether such an assumption can be supported by an arbitrage-free model of the term structure is not addressed in Hunt *et al.* (1996).

Let us now focus on the derivation of values of constants A and B_S . To this end, we assume that equality (4.1) holds, in particular, for any $S = T_j$, $j = 1, \dots, n$. Then

$$A \sum_{j=1}^n \delta_j + \sum_{j=1}^n \delta_j B_{T_j} \kappa(T, T, n) = A(T_n - T_0) + \sum_{j=1}^n \delta_j B_{T_j} \kappa(T, T, n) = 1,$$

and thus

$$A = (T_n - T_0)^{-1}, \quad \sum_{j=1}^n \delta_j B_{T_j} = 0. \quad (4.2)$$

Consequently, using the first equality above and the martingale property of $D(\cdot, S)$ and $\kappa(\cdot, T, n)$, we obtain

$$B(0, S)G_0^{-1}(n) = (T_n - T_0)^{-1} + B_S\kappa(0, T, n), \quad (4.3)$$

so that for each maturity in question the constant B_S is also uniquely determined. Notice that the second equality in (4.2) is also satisfied for this choice of B_S .

Hunt and Kennedy (2000) argue that under (4.1) the problem of pricing irregular cashflows becomes relatively easy to handle. To illustrate this point, assume that we wish to value the claim X which settles at time T and admits the following representation:

$$X = \sum_{i=1}^m c_i B(T, S_i) F,$$

where the c_i are constants, and $S_i \in \mathcal{S}$ for $i = 1, \dots, m$. We assume that the $\mathcal{F}_T$ -measurable random variable F has the form $F = \tilde{F}(B(T, S_1), \dots, B(T, S_m))$ for some function $\tilde{F} : \mathbb{R}_+^m \rightarrow \mathbb{R}$. To be in line with the notation introduced in Section 3.4, we denote

$$V_t^1 = B(t, T) - B(t, T_n), \quad V_t^2 = \sum_{j=1}^n \delta_j B(t, T_j) = G_t(n).$$

Using (4.1) and (4.2)–(4.3), we obtain

$$X = \sum_{i=1}^m c_i (A(1 - B(T, T_n)) + B_{S_i} G_T(n)) F = w_1 V_T^1 F + w_2 V_T^2 F,$$

where $w_1 = \sum_{i=1}^m c_i A$ and $w_2 = \sum_{i=1}^m c_i B_{S_i}$. In view of the discussion in Section 3.4, it is clear that

$$\pi_t(X) = w_1 V_t^1 \mathbb{E}_{\mathbb{P}^1}(F | \mathcal{F}_t) + w_2 V_t^2 \mathbb{E}_{\mathbb{P}^2}(F | \mathcal{F}_t). \quad (4.4)$$

Under the assumption that the forward rate $\kappa(\cdot, T, n)$ follows a geometric Brownian motion under the forward swap measure $\mathbb{P}^2$, it follows also a lognormally distributed process under $\mathbb{P}^1$ (see the discussion in Section 3.4). Consequently, under (4.1), the joint (conditional) probability law of random variables $B(T, S_1), \dots, B(T, S_m)$ under probability measures $\mathbb{P}^1$ and $\mathbb{P}^2$ are explicitly known. We conclude that the conditional expectations in (4.4) can be, in principle, evaluated.

Consider, for instance, a fixed-for-floating constant maturity swap.²⁷ To value one leg of the floating side of a constant maturity swap, consider a cashflow proportional to $\kappa(T, T, n)$, which takes place at some date $M > T$. Ignoring the constant,

²⁷ Similarly as in the case of a plain vanilla fixed-for-floating swap, in a constant maturity swap the fixed and floating payments occur at regularly spaced dates. The amounts of floating payments are based not on a Libor rate, but on some other swap rate, however.

such a payoff is equivalent to the claim $X = B(T, M)\kappa(T, T, n)$ which settles at time T . Using (4.4), we obtain

$$\pi_t(X) = B_M V_t^1 \mathbb{E}_{\mathbb{P}^1}(\kappa(T, T, n) | \mathcal{F}_t) + A V_t^2 \mathbb{E}_{\mathbb{P}^2}(\kappa(T, T, n) | \mathcal{F}_t).$$

Consequently, at time 0 we have

$$\pi_0(X) = B_M(B(0, T) - B(0, T_n))\kappa(0, T, n)e^{\sigma^2 T} + AG_0(n)\kappa(0, T, n),$$

where σ is the implied volatility of the traded swaption with maturity date T . Using the formula for B_M , we get

$$\pi_0(X) = (B(0, M) - AG_0(n))\kappa(0, T, n)e^{\sigma^2 T} + AG_0(n)\kappa(0, T, n),$$

or finally

$$\pi_0(X) = B(0, M)\kappa(0, T, n)(1 + (1 - w)e^{\sigma^2 T}), \quad (4.5)$$

where we write $w = AG_0(n)B^{-1}(0, M)$. It should be stressed that the simple valuation result (4.5) hinges on the strong assumption (4.1).

4.2 Calibration of Markov-functional models

The most important feature of Markov-functional models is the fact that their calibration to market prices of plain vanilla derivatives is relatively easy to perform. For convenience, we shall focus here on the calibration of the Markov-functional model of fixed-maturity forward swap rates. The case of forward Libor rates can be dealt with in an analogous way. A more extensive discussion of this issue can be found in Hunt *et al.* (2000).

First, we assume that the forward swap rate for the date T_{n-1} follows a lognormal martingale under the corresponding forward measure $\mathbb{P}_{T_n}$. More specifically, we postulate that the process $\tilde{\kappa}(\cdot, T_{n-1}) = \kappa(\cdot, T_{n-1}, 1)$ satisfies

$$d\tau\kappa(t, T_{n-1}) = \tilde{\kappa}(t, T_{n-1})v(t, T_{n-1})dW_t, \quad (4.6)$$

where W is a Brownian motion under $\mathbb{P}_{T_n}$ and $v(\cdot, T_{n-1})$ is a strictly positive deterministic function. If we take the process

$$M_t = \int_0^t v(u, T_{n-1}) dW_u$$

as the driving Markov process for our model, then clearly

$$\tilde{\kappa}(T_{n-1}, T_{n-1}) = \tilde{\kappa}(0, T_{n-1}) e^{M_{T_{n-1}} - \frac{1}{2} \int_0^{T_{n-1}} v^2(u, T_{n-1}) du} \quad (4.7)$$

and

$$B(T_{n-1}, T_n, M_{T_{n-1}}) = \left(1 + \delta_n \tilde{\kappa}(0, T_{n-1}) e^{M_{T_{n-1}} - \frac{1}{2} \int_0^{T_{n-1}} v^2(u, T_{n-1}) du} \right)^{-1}. \quad (4.8)$$

Suppose that we are given (digital) swaptions prices for all strikes $\kappa > 0$ and all expiration dates $T_0, \dots, T_{n-1}$. Our goal is to find the joint probability law of $(\tilde{\kappa}(T_0, T_0), \dots, \tilde{\kappa}(T_{n-1}, T_{n-1}))$ under $\mathbb{P}_{T_n}$. This can be achieved by deriving the functional dependence of each rate $\tilde{\kappa}(T_j, T_j)$ on the underlying Markov process; more specifically, we search for the function $h_j : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\tilde{\kappa}(T_j, T_j) = h_j(M_{T_j})$. To this end, we assume that for any $j = 0, \dots, n-1$ there exists a strictly increasing function h_j such that this holds (in view of (4.7), this statement is valid for $j = n-1$).

By the definition of the probability measure $\mathbb{P}_{T_n}$, for $i = j+1, \dots, n$

$$\frac{B(T_j, T_i)}{B(T_j, T_n)} = \mathbb{E}_{\mathbb{P}_{T_n}} \left(\frac{B(T_i, T_i)}{B(T_i, T_n)} \mid \mathcal{F}_{T_i} \right) = \mathbb{E}_{\mathbb{P}_{T_n}} \left(\frac{B(T_i, T_i)}{B(T_i, T_n)} \mid M_{T_j} \right)$$

since $\mathcal{F}_{T_i} = \mathcal{F}_{T_i}^W = \mathcal{F}_{T_i}^M$. Therefore, if $B(T_i, T_n) = B(T_i, T_n, M_{T_i})$ we obtain

$$\frac{B(T_j, T_i)}{B(T_j, T_n)} = \mathbb{E}_{\mathbb{P}_{T_n}} \left(\frac{1}{B(T_i, T_n, M_{T_i})} \mid M_{T_j} \right),$$

so that the right-hand side in the formula above is a function of M_{T_j} . Consequently, for

$$G_{T_j}(n-j) = \sum_{i=j+1}^n \delta_i B(T_j, T_i)$$

we get

$$\frac{G_{T_j}(n-j)}{B(T_j, T_n)} = \sum_{i=j+1}^n \mathbb{E}_{\mathbb{P}_{T_n}} \left(\frac{\delta_i}{B(T_i, T_n, M_{T_i})} \mid M_{T_j} \right) = g_j(M_{T_j}), \quad (4.9)$$

where $g_j : \mathbb{R} \rightarrow \mathbb{R}$ is a measurable function with strictly positive values. The right-hand side in (4.9) can be evaluated using the transition p.d.f. $p_M(t, m; u, x)$ of the Markov process M , provided that the functional form of $B(T_i, T_n, M_{T_i})$ is known for every $i = j+1, \dots, n$. To put it more explicitly,

$$g_j(m) = \sum_{i=j+1}^n \int_{\mathbb{R}} \frac{\delta_i p_M(T_j, m; T_i, x)}{B(T_i, T_n, x)} dx. \quad (4.10)$$

We work back iteratively from the last relevant date T_{n-1} . In the first step, i.e., when $j = n-2$, the functional form of $B(T_{n-1}, T_n, M_{T_{n-1}})$ is given by (4.8). Assume now that the functional forms of $B(T_i, T_n, M_{T_i})$ were already found for

$i = j + 1, \dots, n - 1$. In order to determine $B(T_j, T_n, M_{T_j})$, it is enough to find the functional form of the swap rate $\tilde{\kappa}(T_j, T_j)$. Indeed, we have

$$\tilde{\kappa}(T_j, T_j) = \frac{1 - B(T_j, T_n)}{G_{T_j}(n - j)}$$

and thus

$$B^{-1}(T_j, T_n) = 1 + \tilde{\kappa}(T_j, T_j) \frac{G_{T_j}(n - j)}{B(T_j, T_n)} = 1 + h_j(M_{T_j}) g_j(M_{T_j}). \quad (4.11)$$

Our next goal is to show how to find the function h_j , under the assumption that the functional forms of bonds prices $B(T_i, T_n, M_{T_i})$ are known for every $i = j + 1, \dots, n$. To this end, we assume that we are given all market prices of digital swaptions with expiration date T_j and any strictly positive strike level κ . We find it convenient to represent the price at time 0 of the j^{th} digital swaption, with strike κ and expiration date T_j , in the following way:²⁸

$$\mathbf{DS}_0^j(\kappa) = B(0, T_n) \mathbb{E}_{\mathbb{P}_{T_n}} \left(\frac{G_{T_j}(n - j)}{B(T_j, T_n)} \mathbb{1}_{\{\tilde{\kappa}(T_j, T_j) > \kappa\}} \right)$$

for $j = 0, \dots, n - 2$. Under the present assumptions, we obtain

$$\mathbf{DS}_0^j(\kappa) = B(0, T_n) \mathbb{E}_{\mathbb{P}_{T_n}} (g_j(M_{T_j}) \mathbb{1}_{\{h_j(M_{T_j}) > \kappa\}}),$$

or equivalently,

$$\mathbf{DS}_0^j(\kappa) = B(0, T_n) \mathbb{E}_{\mathbb{P}_{T_n}} (g_j(M_{T_j}) \mathbb{1}_{\{M_{T_j} > h_j^{-1}(\kappa)\}}).$$

Finally, if we denote by $f_M(x) = p_M(0, 0; T_j, x)$ the p.d.f. of M_{T_j} under $\mathbb{P}_{T_n}$, then

$$\mathbf{DS}_0^j(\kappa) = B(0, T_n) \int_{\mathbb{R}} g_j(x) \mathbb{1}_{\{x > \tilde{h}_j(\kappa)\}} f_M(x) dx, \quad (4.12)$$

where we write $\hat{h}_j = h_j^{-1}$. It is natural to assume that the function²⁹ $\mathbf{DS}_0^j : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is strictly decreasing as a function of the strike level κ , with

$$\mathbf{DS}_0^j(0) = \sum_{i=j+1}^n \delta_i B(0, T_i) = G_0(n - j)$$

and $\mathbf{DS}_0^j(+\infty) = 0$. Since

$$\mathbb{E}_{\mathbb{P}_{T_n}} (g_j(M_{T_j})) = G_0(n - j) B^{-1}(0, T_n)$$

²⁸ By definition, the j^{th} digital swaption, with unit notional principal, pays the amount δ_i at time T_i for $i = j + 1, \dots, n$ whenever the inequality $\tilde{\kappa}(T_j, T_j) > \kappa$ holds.

²⁹ Recall that the function $\mathbf{DS}_0^j$ represents the observed market prices of digital swaptions. Therefore, the foregoing assumptions about the behaviour of this function are indeed quite natural.

it can be deduced from (4.12) that $\hat{h}_j(0) = -\infty$. On the other hand, condition $\mathbf{DS}_0^j(+\infty) = 0$ implies that $\hat{h}_j(+\infty) = +\infty$. Finally, the function $\hat{h}_j$ implicitly defined through equality (4.12) is strictly increasing, so that it admits an inverse function h_j with desired properties. To wit, for $h_j = \hat{h}_j^{-1}$ we have: $h_j : \mathbb{R} \rightarrow \mathbb{R}_+$ is strictly increasing, with $h_j(-\infty) = 0$ and $h_j(+\infty) = +\infty$. This shows that the procedure above leads to a reasonable specification of the functional form $\tilde{\kappa}(T_j, T_j) = h_j(M_{T_j})$.

For the reader's convenience, we shall recapitulate the main steps of the calibration procedure. In the first step, we numerically find the function h_{n-2} which expresses $\tilde{\kappa}(T_{n-2}, T_{n-2})$ in terms of $M_{T_{n-2}}$. To this end, we need first to evaluate the function g_{n-2} using formula (4.10) with $B(T_n, T_n, x) = 1$ and $B(T_{n-1}, T_n, x)$ given by (4.8).

In the second step, we first determine $B(T_{n-2}, T_n, x)$ using relationship (4.11), that is,

$$B^{-1}(T_{n-2}, T_n, x) = 1 + h_{n-2}(x)g_{n-2}(x).$$

Then, we find g_{n-3} using (4.10), and subsequently we determine the rate $\tilde{\kappa}(T_{n-3}, T_{n-3})$, or rather the corresponding function h_{n-3} .

Continuing this procedure, we end up with the following representation of the finite family of swap rates:

$$(\tilde{\kappa}(T_0, T_0), \dots, \tilde{\kappa}(T_{n-1}, T_{n-1})) = (g_0(M_{T_0}), \dots, g_{n-1}(M_{T_{n-1}})).$$

This representation uniquely specifies the probability law of the considered family of swap rates under the terminal forward measure $\mathbb{P}_{T_n}$.

Remarks In view of (4.6), the price at time $t \leq T_{n-1}$ of the $(n-1)^{\text{th}}$ digital swaption equals

$$\mathbf{DS}_t^{n-1}(\kappa) = \delta_n B(t, T_n) \mathbb{P}_{T_n} \{ \tilde{\kappa}(T_{n-1}, T_{n-1}) > \kappa \mid \mathcal{F}_t \},$$

that is,

$$\mathbf{DS}_t^{n-1}(\kappa) = \delta_n B(t, T_n) N(\tilde{h}_2(t, T_{n-1})), \quad (4.13)$$

where N denotes the standard Gaussian cumulative distribution function, and the coefficient $\tilde{h}_2$ is given in the formulation of Proposition 3.5. Needless to say that formula (4.13) is not valid in the present setup, even for $t = 0$, for any digital swaption with maturity $T_0, \dots, T_{n-2}$. Moreover, it is clear that assumption (4.6) is not necessary; we need only assume that the functional form of the swap rate $\tilde{\kappa}(T_{n-1}, T_{n-1})$ with respect to some underlying Markov process M is explicitly known (and is a monotone function of $M_{T_{n-1}}$).

References

- Andersen, L. (2000), A simple approach to the pricing of Bermudan swaptions in the multifactor LIBOR market model, *Journal of Computational Finance* **3**(2), 5–32.
- Andersen, L. and Andreasen, J. (1997), Volatility skews and extensions of the Libor market model, working paper, National Australia Bank and University of New South Wales.
- Brace, A. (1996), Dual swap and swaption formulae in the normal and lognormal models, working paper, University of New South Wales.
- Brace, A., Gątarek, D. and Musiela, M. (1997), The market model of interest rate dynamics, *Mathematical Finance* **7**, 127–54.
- Brace, A., Musiela, M. and Schlögl, E. (1998), A simulation algorithm based on measure relationships in the lognormal market model, working paper, University of New South Wales.
- Brace, A. and Womersley, R.S. (2000), Exact fit to the swaption volatility matrix using semidefinite programming, working paper, National Australia Bank and University of New South Wales.
- Bühler, W. and Käsler, J. (1989), Konsistente Anleihenpreise und Optionen auf Anleihen, working paper, University of Dortmund.
- Cox, J. and Ross, S. (1976), The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–66.
- Döberlein, F. and Schweizer, M. (1998), On term structure models generated by semimartingales, working paper, Technische Universität Berlin.
- Döberlein, F., Schweizer, M. and Stricker, C. (2000), Implied savings accounts are unique, *Finance and Stochastics* **4**, 431–42.
- Dun, T., Schlögl, E. and Barton, G. (2000), Simulated swaption delta-hedging in the lognormal forward LIBOR model, working paper, University of Sydney and University of Technology, Sydney.
- Flesaker, B. (1993), Arbitrage free pricing of interest rate futures and forward contracts, *Journal of Futures Markets* **13**, 77–91.
- Flesaker, B. and Hughston, L. (1996a), Positive interest, *Risk* **9**(1), 46–9.
- Flesaker, B. and Hughston, L. (1996b), Positive interest: foreign exchange, in: *Vasicek and Beyond*, L. Hughston, ed., Risk Publications, London, pp. 351–67.
- Flesaker, B. and Hughston, L. (1997), Dynamic models of yield curve evolution, in: *Mathematics of Derivative Securities*, M.A.H. Dempster and S.R. Pliska, eds., Cambridge University Press, Cambridge, pp. 294–314.
- Geman, H., El Karoui, N. and Rochet, J.C. (1995), Changes of numeraire, changes of probability measures and pricing of options, *Journal of Applied Probability* **32**, 443–58.
- Glasserman, P. and Kou, S.G. (1999), The term structure of simple forward rates with jump risk, working paper, Columbia University.
- Glasserman, P. and Zhao, X. (1999), Fast greeks by simulation in forward LIBOR models, *Journal of Computational Finance* **3**(1), 5–39.
- Glasserman, P. and Zhao, X. (2000), Arbitrage-free discretization of lognormal forward Libor and swap rate model, *Finance and Stochastics* **4**, 35–68.
- Goldys, B. (1997), A note on pricing interest rate derivatives when Libor rates are lognormal, *Finance and Stochastics* **1**, 345–52.
- Goldys, B., Musiela, M. and Sondermann, D. (1994), Lognormality of rates and term structure models, working paper, University of New South Wales.
- Heath, D., Jarrow, R. and Morton, A. (1992), Bond pricing and the term structure of

- interest rates: a new methodology for contingent claim valuation, *Econometrica* **60**, 77–105.
- Hull, J.C. and White, A. (1999), Forward rate volatilities, swap rate volatilities, and the implementation of the LIBOR market model, working paper, University of Toronto.
- Hunt, P.J. and Kennedy, J.E. (1997), On convexity corrections, working paper, ABN-Amro Bank and University of Warwick.
- Hunt, P.J. and Kennedy, J.E. (1998), Implied interest rate pricing model, *Finance and Stochastics* **2**, 275–93.
- Hunt, P.J. and Kennedy, J.E. (2000) *Financial Derivatives in Theory and Practice*, John Wiley & Sons, Chichester.
- Hunt, P.J., Kennedy, J.E. and Pelsser, A. (2000), Markov-functional interest rate models, *Finance and Stochastics* **4**, 391–408.
- Hunt, P.J., Kennedy, J.E. and Scott, E.M. (1996), Terminal swap-rate models, working paper, ABN-Amro Bank and University of Warwick.
- Jamshidian, F. (1996), Pricing and hedging European swaptions with deterministic (lognormal) forward swap rate volatility, working paper, Sakura Global Capital.
- Jamshidian, F. (1997), Libor and swap market models and measures, *Finance and Stochastics* **1**, 293–330.
- Jamshidian, F. (1999), Libor market model with semimartingales, working paper, NetAnalytic Limited.
- Jin, Y. and Glasserman, P. (1997), Equilibrium positive interest rates: a unified view, forthcoming in *Review of Financial Studies*.
- Lotz, C. and Schlögl, L. (2000), Default risk in a market model, *Journal of Banking and Finance* **24**, 301–27.
- Miltersen, K., Sandmann, K. and Sondermann, D. (1997), Closed form solutions for term structure derivatives with log-normal interest rates, *Journal of Finance* **52**, 409–30.
- Musiela, M. (1994), Nominal annual rates and lognormal volatility structure, working paper, University of New South Wales.
- Musiela, M. and Rutkowski, M. (1997a) *Martingale Methods in Financial Modelling*, Springer-Verlag, Berlin.
- Musiela, M. and Rutkowski, M. (1997b), Continuous-time term structure models: forward measure approach, *Finance and Stochastics* **1**, 261–91.
- Musiela, M. and Sawa, J. (1998), Interpolation and modelling term structure, working paper, University of New South Wales.
- Musiela, M. and Sondermann, D. (1993), Different dynamical specifications of the term structure of initial rates and their implications, working paper, University of Bonn.
- Neuberger, A. (1990), Pricing swap options using the forward swap market, working paper, London Business School.
- Rady, S. (1997), Option pricing in the presence of natural boundaries and a quadratic diffusion term, *Finance and Stochastics* **1**, 331–44.
- Rady, S. and Sandmann, K. (1994), The direct approach to debt option pricing, *Review of Futures Markets* **13**, 461–514.
- Rebonato, R. (1999), On the pricing implications of the joint lognormal assumption for the swaption and cap markets, *Journal of Computational Finance* **2(3)**, 57–76.
- Rebonato, R. (2000), On the simultaneous calibration of multifactor lognormal interest rate models to Black volatilities and to the correlation matrix, *Journal of Computational Finance* **2(4)**, 5–27.
- Rutkowski, M. (1997), A note on the Flesaker-Hughston model of term structure of interest rates, *Applied Mathematical Finance* **4**, 151–63.
- Rutkowski, M. (1998), Dynamics of spot, forward, and futures Libor rates, *International*

Journal of Theoretical and Applied Finance **1**, 425–45.

Rutkowski, M. (1999), Models of forward Libor and swap rates, *Applied Mathematical Finance* **6**, 29–60.

Sandmann, K. and Sondermann, D. (1993), On the stability of lognormal interest rate models, working paper, University of Bonn.

Sandmann, K. and Sondermann, D. (1997), A note on the stability of lognormal interest rate models and the pricing of Eurodollar futures, *Mathematical Finance* **7**, 119–25.

Sandmann, K., Sondermann, D. and Miltersen, K.R. (1995), Closed form term structure derivatives in a Heath–Jarrow–Morton model with log-normal annually compounded interest rates, in: *Proceedings of the Seventh Annual European Futures Research Symposium Bonn, 1994*, Chicago Board of Trade, pp. 145–65.

Schlögl, E. (1999), A multicurrency extension of the lognormal interest rate market model, working paper, University of Technology, Sydney.

Schmidt, W.M. (1996), Pricing irregular interest cash flows, working paper, Deutsche Morgan Grenfell.

Schoenmakers, J. and Coffey, B. (1999), Libor rates models, related derivatives and model calibration, working paper.

Sidenius, J. (1997), Libor market models in practice, *Journal of Computational Finance* **3**(3), 5–26.

Uratani, T. and Utsunomiya, M. (1999), Lattice calculation for forward LIBOR model, working paper, Hosei University.

Yasuoka, T. (1998), No arbitrage relation between a swaption and a cap/floor in the framework of Brace, Gatarek and Musiela, working paper, Fuji Research Institute Corporation.

Yasuoka, T. (1999), Mathematical pseudo-completion of the BGM model, working paper, Fuji Research Institute Corporation.

Part three

Risk Management and Hedging

Credit Risk Modelling: Intensity Based Approach

Tomasz R. Bielecki and Marek Rutkowski

1 Introduction

Let $B(t, T)$ and $D(t, T)$ denote prices at time t of *default-free* and *default-risky* (or *defaultable*) zero coupon bonds maturing at time T , respectively. The default-free bond pays \$1 at time T . The (recovery) payment for the default-risky bond needs to be modelled. Two major situations are commonly considered (if the bond defaults prior to or on the maturity date then): (a) the recovery payment is received by the holder of the defaultable bond at the default time of the bond, or (b) the recovery payment is received by the holder of the defaultable bond at the maturity time of the bond. Of course, if the defaultable bond does not default prior to or on the maturity date, then it pays \$1 at maturity.

In this chapter we present a survey of recent research efforts aimed at pricing and hedging of default-prone debt instruments. We concentrate on intensity and ratings based approaches. In particular we review some results derived by Duffie, Schröder and Skiadas (1996), Duffie and Singleton (1998a, 1999), Jarrow and Turnbull (1995, 2000), Jarrow, Lando and Turnbull (1997), Lando (1998), Madan and Unal (1998a, 1998b), Jeanblanc and Rutkowski (2000a, 2000b), Bielecki and Rutkowski (1999, 2000), and Lotz and Schlögl (2000), among results obtained by other researchers. In addition we present a brief survey of some important types of credit derivatives, that is derivative products linked to either corporate or sovereign debt, and we describe how to price them within the Bielecki and Rutkowski approach. It should be emphasized that the need to rationally price and hedge credit derivatives, whose presence in financial markets has been continuously growing in the recent years, was one of the motivations, besides the need to manage credit risk, behind the explosion of research on quantitative aspects of the credit risk that has been observed in the 1990s.

Let us mention here that the firm-specific approach – that is, an approach based on observations of the value of debt's issuer – is not addressed in the present

chapter. This alternative approach was initiated in the 1970s by Merton (1974), Black and Cox (1976), and Geske (1977). It was subsequently developed in various directions by several authors; to mention a few: Brennan and Schwartz (1997, 1980), Pitts and Selby (1983), Rendleman (1992), Kim *et al.* (1993), Nielsen *et al.* (1993), Leland (1994), Longstaff and Schwartz (1995), Leland and Toft (1996), Mella-Barral and Tychon (1996), Briys and de Varenne (1997), Crouhy *et al.* (1998, 2000), Duffie and Lando (1998), and Anderson and Sundaresan (2000). Reviewing this approach would require a separate article (see, e.g., Ammann (1999)). The list of references is not representative of all important papers and books published in this area in recent years, but it includes works that are most related to this presentation.

2 Credit derivatives

Credit derivatives are privately negotiated derivatives securities that are linked to a credit-sensitive asset as the underlying asset. More specifically, the reference security of a credit derivative can be an actively-traded corporate or sovereign bond or a portfolio of these bonds. A credit derivative can also have a loan (or a portfolio of loans) as the underlying reference credit. Credit derivatives can be structured in a large variety of ways; they are typically complex agreements, customized to the precise needs of an investor. The common feature of all credit derivatives is the fact that they allow for the transference of the credit risk from one counterparty to another, so that they can be used to control the credit risk exposure. *Credit risk* refers to the possibility that a borrower will fail to service or repay a debt on time. The overall risk we are concerned with involves two components: market risk and asset-specific credit risk. In contrast to ‘standard’ interest-rate derivatives, credit derivatives allow us to isolate and handle not only the market risk, but also the firm-specific credit risk. They provide also a way to synthesize assets that are otherwise not available to a particular investor (in this application, an investor ‘buys’ – rather than ‘sells’ – a specific credit risk).

Similarly as in the case of derivative securities associated with the risk-free term structure, we may formally distinguish three main types of agreements: forward contracts, swaps, and options. A *forward contract* commits the buyer to purchasing a specified bond at a specified future date at a price predetermined at contract inception. In a forward contract, the default risk is normally borne by the buyer. If a credit event occurs, the transaction is marked to market and unwound. Forward contracts can also be transacted in spread form; that is, the agreement can be based on the specified bond’s spread over a benchmark asset. It should be stressed that the classification above does not corresponds to market terminological conventions, as described below.

In market practice, the most popular credit-sensitive swap contract is a *total rate of return swap*, explained in some detail in Section 2.1 below. Credit options are typically embedded in complex credit-sensitive agreements, though the over-the-counter traded credit options – such as *default puts*, also described in Section 2.1 – are also available. Let us finally mention the so-called *vulnerable options*, or more generally, *vulnerable claims*. These are contingent agreements that are issued by credit-sensitive institutions, so that they are subject to default in much the same way as defaultable bonds.

2.1 Overview of instruments

We first review the most actively traded types of credit-sensitive agreements.¹ It should be stressed that we do not intend to examine here all aspects of credit derivatives as a tool in the risk management. The non-exhaustive list of examples given below makes it clear that a wide range of objectives can be achieved by trading in credit derivatives. For an extensive analysis of economical reasons which support the use of these products, we refer to Das (1998a, 1998b) or Tavakoli (1998).

Total rate of return swaps

Total rate of return swaps (*total return swaps*, for short) are agreements in which the total return of an underlying credit-sensitive asset (basket of assets, index, etc.) is exchanged for some other cash flow. More specifically, one party agrees to pay the total return (income plus or minus any change in the capital value) on a notional principal amount to another party in return for periodic fixed or floating-rate payments on the same notional amount. Let us enumerate the most important features of a total return swap: (a) no principal amounts are exchanged and no physical change of ownership occurs, (b) the maturity of the total return swap agreement need not match that of the underlying, (c) at the contract termination – i.e., at the contract maturity or upon default – according to Das (1998a), ‘a price settlement based on the change in the value of the bond or loan is made’. Total return swaps can incorporate put and call options (to establish caps and floors on the returns of the reference assets), as well as caps and floors on a floating interest rates.

Credit-spread swaps and options

With *credit-spread swaps* (that is, relative performance total return swaps), also known as *credit-spread forwards*, investors pay the total return of one asset while receiving the total return of another credit-sensitive asset. *Credit-spread options*

¹ Let us mention that the terminological conventions relative to credit derivatives are not yet fully standardized; we shall try to follow the most widely accepted terminology.

are option agreements whose payoff is associated with the yield differential of two credit-sensitive assets. For instance, the reference rate of the option can be a spread of a corporate bond over a benchmark asset of comparable maturity. The option can be settled either in cash or through physical delivery of the underlying bond, at a price whose yield spread over the benchmark asset equals the strike spread. Options on credit spreads allow one to isolate the firm-specific credit risk from the market risk.

Credit (default) swaps

These are agreements in which a periodic fixed payments (or upfront fee) from the protection buyer is exchanged for the promise of some specified payment from the protection seller to be made only if a particular, predetermined *credit event* occurs. If, during the term of the default swap, a credit event occurs, the seller pays the buyer an amount to cover the loss, and the swap then terminates. If no credit event has occurred by maturity of the swap, both sides end their obligations to each other. The most important covenants of a credit swap contract are: (a) the specification of the *credit event*, which is formally defined as a 'default' (in practice, it may include: bankruptcy, insolvency, payment default, a stipulated price decline for the reference asset, or a rating downgrade for the reference asset), (b) the contingent default payment, which may be structured in a number of ways; for instance, it may be linked to the price movement of the reference asset, or it can be set at a predetermined level (e.g., a fixed percentage of the notional amount of the transaction), (c) the specification of periodic payments which depend, in large part, on the credit quality of the reference asset. Credit swaps are usually settled in cash, but the agreement may also provide for physical delivery; for example, it may involve payment at par by the seller in exchange for the delivery of the defaulted reference asset. If the payment is triggered by the default and equals to the difference between the face value of a bond and its market price, the contract is named the *default swap*. Let us finally mention the so-called *first-to-default swaps*, which are examples of *basket default swaps* (i.e., default swaps linked to a portfolio of credit-sensitive securities).

Credit (default) options

A credit call (put, resp.) option gives the right to buy (to sell, resp.) an underlying credit-sensitive asset (index, credit spread, etc.) at a predetermined price. The most widely used type of a credit option is a *default put*. The buyer of the default put pays a premium (either an upfront fee or a periodic payment) to the seller who then assumes the default risk for the reference asset. If there is a credit (default) event during the term of the option, the seller pays the buyer a (fixed or variable) *default payment*.

Credit linked notes

Credit linked notes are debt instruments in which the coupon or price of the note is linked to the performance of a reference credit-sensitive asset (rate or index). For instance, a credit-linked note may stipulate that the principal repayment is reduced to a certain level below par if the external corporate or sovereign debt defaults before the maturity of the note. This means that the buyer of the note sells credit protection to the issuer of the note; in exchange the note pays a higher-than-normal yield.

2.2 Market pricing methods

Since a reliable benchmark model for credit derivatives is not yet available, it is common in market practice to value a credit derivative on a stand-alone basis, using a judiciously chosen *ad hoc* approach, rather than a sophisticated mathematical model. We shall review the most widely used of these approaches. For explanatory purposes, we focus on the valuation of a default swap, and we base our description of the pricing methods² on BeSaw (1997).

Same-cost as reference method

To estimate the price of a default swap, one assumes that there exists an insured bond which is otherwise identical to the reference bond of the swap. The spread between the yield of the insured bond and that of the reference bond can then be taken as the proxy of the default swap price. Notice that this method identifies a default swap with bond insurance, and disregards the credit difference between the bond insurer and the default swap counterparty.

Credit-spread-based method

This way of default swap valuation is based on a comparison of the yield of the reference bond and the yield of a risk-free bond with similar maturity. It is thus implicitly assumed that the spread over the risk-free asset is entirely due to the credit risk so that the impact of tax and/or liquidity effects are neglected. Another difficulty arises when one wishes to price a swap with maturity which does not correspond to the maturity of the reference corporate bond.

Replication of cost method

In this method, the price of a default swap is calculated through evaluation of the cost of a portfolio necessary to replicate the swap. The replication of cost method

² For an exhaustive analysis of practical aspects of credit swaps and a review of non-technical methods of their valuation (including the estimation of hazard rates), we refer to Duffie (1999).

thus mimics the standard approach to contingent claims valuation in an arbitrage-free setup. Unfortunately, it is typically not possible or too costly to establish a (static or dynamic) portfolio which fully hedges (i.e. replicates) a credit derivative.

Ratings-based default method

This approach, which will be analysed in more details in what follows, determines the price of a credit derivative (for instance, a default swap) as the expected loss resulting from default. To derive default probabilities, it is common to model the Markov chain representing ratings migration process using the estimated credit ratings transition matrix. If the valuation is made on a stand-alone basis, it would be more adequate to use the firm-specific transition matrix corresponding to the reference asset. It is clear that such a matrix is not easily available, however. Similarly, constant (or random) recovery rates, which are needed to evaluate the expected loss, are either inferred using the historical data, or assessed on a stand-alone basis. The *credit-spread-based default method* can be seen as a variant of a ratings-based default method. It uses an issuer-specific credit spread over default-free instruments of similar maturity to estimate the probability of default and the expected recovery rate in default.

3 Valuation of defaultable claims

The exposition in this section is mainly based on Duffie *et al.* (1996). In this section, our goal is to present the most fundamental results which can be obtained using the intensity-based approach. In Section 4, special attention will be paid to the various kinds of recovery rates, such as, for instance, zero recovery, fractional recovery of par, and fractional recovery of market value. On the other hand, in order to obtain as explicit valuation formulae as possible, we shall still assume that only two states are possible, namely, non-default and default. An analysis of the case of several credit rating classes is postponed to Sections 5–7. We make the following standing assumptions.

(A.1) We are given a probability space $(\Omega, \mathcal{G}, \mathbb{P}^*)$, endowed with the filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{R}_+}$ (of course, $\mathcal{F}_t \subset \mathcal{G}$ for every $t \in \mathbb{R}_+$). The probability measure $\mathbb{P}^*$ is interpreted as a martingale measure for our underlying securities market model (complete or not). Let τ be a non-negative random variable on the probability space $(\Omega, \mathcal{G}, \mathbb{P}^*)$. In what follows, we shall refer to τ as the *default time*.

For convenience, we assume that for every $t \in \mathbb{R}_+$, $\mathbb{P}^*\{\tau = 0\} = 0$ and $\mathbb{P}^*\{\tau > t\} > 0$. Given a default time τ , we introduce the associated (single) jump process H by setting $H_t = \mathbb{1}_{\{\tau \leq t\}}$ for $t \in \mathbb{R}_+$. It is obvious that H is a right-continuous process. Let $\mathbb{H}$ be the filtration generated by the process H ,

i.e., $\mathcal{H}_t = \sigma(H_u : u \leq t)$. We introduce the enlarged filtration $\mathbb{G}$ which satisfies $\mathbb{G} = \mathbb{H} \vee \mathbb{F}$ – that is, $\mathcal{G}_t = \mathcal{H}_t \vee \mathcal{F}_t = \sigma(\mathcal{H}_t, \mathcal{F}_t)$ for every t .

(A.2) For a given default-risky security, its *default process* is modelled through a jump process H with strictly positive *intensity* (or *hazard rate*) process³ λ under $\mathbb{P}^*$. The intensity λ is an $\mathbb{F}$ -progressively measurable process such that the *compensated process*

$$M_t := H_t - \int_0^{t \wedge \tau} \lambda_u du = H_t - \int_0^t h_u du, \quad \forall t \in [0, T^*], \quad (3.1)$$

follows an $\mathbb{G}$ -martingale under $\mathbb{P}^*$. Notice that the auxiliary $\mathbb{G}$ -adapted process h satisfies $h_t := \mathbb{1}_{\{t \leq \tau\}} \lambda_t$.

Remarks Let us stress that the stochastic intensity λ is assumed to follow an $\mathbb{F}$ -adapted adapted process, and the filtration of reference $\mathbb{F}$ can be strictly smaller than $\mathbb{G}$, in general. On the other hand, the case of an $\mathbb{F}$ -stopping time is also covered (in this case, $\mathbb{F} = \mathbb{G}$).

(A.3) Given a maturity date $T > 0$, an $\mathcal{F}_T$ -measurable random variable X represents the *promised claim*, that is, the amount of cash which the owner of a defaultable claim is entitled to receive at time T , provided that the default has not occurred before the maturity date T .

(A.4) An $\mathbb{F}$ -predictable process Z models the payoff which is actually received by the owner of a defaultable claim, if default occurs before maturity T . We shall refer to Z as the *recovery process* of X .

(A.5) An $\mathbb{F}$ -adapted process r stands for the short-term interest rate, and $B_t := \exp(\int_0^t r_u du)$, $t \in \mathbb{R}_+$, is the associated savings account process.

The main result in the intensity-based approach states that a defaultable security can be priced as if it were a default-risk free security, provided that the credit spread is already incorporated in the risk premium. In other words, the risk premium process of a defaultable security differs from that associated with a risk-free bond, both in the real-world and in the risk-neutral world. In particular, in a risk-neutral world the risk premium associated with a risk-free bond vanishes, but the risk premium associated with a defaultable security is still present.

³ We refer to Artzner and Delbaen (1995), Kusuoka (1999), Rutkowski (1999), Elliott *et al.* (2000) or Jeanblanc and Rutkowski (2000a, 2000b) for more details on stochastic intensities.

Example 3.1 If the intensity process $\lambda_t = \lambda > 0$ is constant, the process H can be seen as a continuous-time Markov chain with the state space $\{0, 1\}$, and with constant intensity matrix $\Lambda = [\lambda_{ij}]_{0 \leq i, j \leq 1}$, where $\lambda_{00} = -\lambda$, $\lambda_{01} = \lambda$, and $\lambda_{1i} = 0$ for $i = 0, 1$ (so that the state 1 is absorbing). In this case, τ can be seen as the first jump time of a standard Poisson process N with constant intensity λ . This simple example can be generalized in two directions. First, in some circumstances it might be natural to assume that $\lambda_t = \lambda(Y_t)$, where Y is a given k -dimensional $\mathbb{F}$ -adapted stochastic process, and $\lambda : \mathbb{R}^k \rightarrow \mathbb{R}_+$ is a positive deterministic function. Second, the basic model can be extended to accommodate for different credit rating classes, $\Lambda_t = [\lambda_{ij}(Y_t)]_{0 \leq i, j \leq K}$, with K being an absorbing state (see, e.g., Jarrow *et al.* (1997) or Section 6).

We need first to formally define the value process S of a (European) *defaultable claim*, represented by a triplet (X, Z, τ) and maturity date T . Since we assume throughout that $\mathbb{P}^*$ is a spot martingale measure, it is natural to postulate that the value S_0 at time 0 of a defaultable claim (X, Z, τ) equals

$$S_0 := B_0 \mathbb{E}_{\mathbb{P}^*} \left(\int_{[0, T]} B_u^{-1} dD_u \right), \quad (3.2)$$

where B stands for the savings account process, and D is the ‘dividend process’ (cf. (A.3)–(A.4))

$$D_t = \int_{[0, t]} Z_u dH_u + X(1 - H_T) \mathbb{1}_{\{t=T\}}. \quad (3.3)$$

Formula (3.2) can be easily generalized to give the price of a defaultable claim at any date t , namely

$$S_t := B_t \mathbb{E}_{\mathbb{P}^*} \left(\int_{[t, T]} B_u^{-1} dD_u \mid \mathcal{G}_t \right), \quad (3.4)$$

or equivalently,

$$S_t := B_t \mathbb{E}_{\mathbb{P}^*} \left(\int_{[t, T]} B_u^{-1} Z_u dH_u + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right). \quad (3.5)$$

In particular, at maturity of the contract we have $S_T = X \mathbb{1}_{\{T < \tau\}}$, as expected. Notice that (3.5) can be also rewritten as follows:

$$S_t = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_\tau^{-1} Z_\tau \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right), \quad (3.6)$$

or finally,

$$S_t = \mathbb{E}_{\mathbb{P}^*} \left(e^{-\int_t^{\tau \wedge T} r_u du} (Z_\tau \mathbb{1}_{\{t < \tau \leq T\}} + X \mathbb{1}_{\{T < \tau\}}) \mid \mathcal{G}_t \right). \quad (3.7)$$

Definition 3.2 By a *defaultable claim* we mean a triplet (X, Z, τ) , where X is the *promised payoff*, Z represents the *recovery process* of X , and τ is the *default time*. The *price* (or *value*) *process* S of a defaultable claim (X, Z, τ) is given by any of the formulae (3.4)–(3.7).

Remarks Notice that Definition 3.2 specifies the price of a defaultable security on the ex-dividend basis. In particular, for any t we have $S_t = 0$ on the event $\{\tau \leq t\}$. Intuitively, this means that the payoff at the event of default is received in cash (and invested, e.g., in the risk-free savings account), and the defaultable security becomes worthless forever. This convention agrees, of course, with our current set of Assumptions (A.1)–(A.5), but does not necessarily reflect the actual bankruptcy procedures. Once again, it should be generalized to fit more adequately the real-world behaviour of defaultable securities.

The following lemma provides still another representation for the price process S of a defaultable claim. It appears that, due to Assumption (A.2), the integration with respect to the process H_t can be substituted with the integration with respect to the associated *intensity measure* $h_t dt$.

Lemma 3.3 *The price process S admits the following representations*

$$S_t = B_t \mathbb{E} \mathbb{P}^* \left(\int_t^T B_u^{-1} Z_u h_u du + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right) \quad (3.8)$$

and

$$S_t = \mathbb{E} \mathbb{P}^* \left(\int_t^T (Z_u h_u - r_u S_u) du + X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right). \quad (3.9)$$

Proof The first formula follows from (3.5), combined with the equality

$$\mathbb{E} \mathbb{P}^* \left(\int_{[t, T]} B_u^{-1} Z_u dH_u \mid \mathcal{G}_t \right) = \mathbb{E} \mathbb{P}^* \left(\int_{[t, T]} B_u^{-1} Z_u (dM_u + h_u du) \mid \mathcal{G}_t \right),$$

which in turn is an immediate consequence of (3.1). For the second, it is enough to rewrite (3.8) as follows:

$$S_t = B_t \left(\tilde{M}_t - \int_0^t B_u^{-1} Z_u h_u du \right), \quad (3.10)$$

where we have put

$$\tilde{M}_t = \mathbb{E} \mathbb{P}^* \left(\int_0^T B_u^{-1} Z_u h_u du + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right).$$

By applying Itô's formula to (3.10), we obtain

$$dS_t = (r_t S_t - Z_t h_t) dt + B_t d\tilde{M}_t,$$

and thus

$$\mathbb{E}_{\mathbb{P}^*}(S_T | \mathcal{G}_t) = S_t + \mathbb{E}_{\mathbb{P}^*}\left(\int_t^T (r_u S_u - Z_u h_u) du \mid \mathcal{G}_t\right).$$

Since obviously $S_T = X \mathbb{1}_{\{T < \tau\}}$, the last equality yields (3.9). $\square$

Notice that for Lemma 3.3 to hold, it is enough to assume that processes B and Z are $\mathbb{G}$ -predictable, and X is $\mathcal{G}_T$ -measurable. The following result – due to Duffie *et al.* (1996) – plays a crucial role in what follows.

Theorem 3.4 *For a given $\mathbb{F}$ -predictable process Z and $\mathcal{F}_T$ -measurable random variable X , we define the process V by setting*

$$V_t = \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}\left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{G}_t\right), \quad (3.11)$$

where $\tilde{B}$ is the ‘savings account’ corresponding to the default-adjusted short-term rate $R_t = r_t + \lambda_t$, that is,

$$\tilde{B}_t = \exp\left(\int_0^t (r_u + \lambda_u) du\right). \quad (3.12)$$

Then

$$\mathbb{1}_{\{t < \tau\}} V_t = B_t \mathbb{E}_{\mathbb{P}^*}\left(B_t^{-1} (Z_\tau + \Delta V_\tau) \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t\right). \quad (3.13)$$

Proof In view of (3.11), we have

$$V_t = \tilde{B}_t \left(N_t - \int_0^t \tilde{B}_u^{-1} Z_u \lambda_u du \right), \quad (3.14)$$

where N is a $\mathbb{G}$ -martingale given by the formula

$$N_t = \mathbb{E}_{\mathbb{P}^*}\left(\int_0^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{G}_t\right). \quad (3.15)$$

Using Itô's product rule, we obtain

$$dV_t = r_t V_t dt - (Z_t - V_{t-}) \lambda_t dt + \tilde{B}_t dN_t. \quad (3.16)$$

Define $U_t = \tilde{H}_t V_t$, where $\tilde{H}_t = 1 - H_t = \mathbb{1}_{\{t < \tau\}}$, so that $U_t = \mathbb{1}_{\{t < \tau\}} V_t$. It is useful to observe that (3.13) may be rewritten as follows

$$U_t = B_t \mathbb{E}_{\mathbb{P}^*}\left(\int_{[t, T]} B_u^{-1} (Z_u + \Delta V_u) dH_u + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t\right). \quad (3.17)$$

On the other hand, an application of Itô's product rule yields (obviously the process $\tilde{H}$ is of finite variation)

$$dU_t = d(V_t \tilde{H}_t) = \tilde{H}_{t-} dV_t + V_{t-} d\tilde{H}_t + \Delta V_t \Delta \tilde{H}_t.$$

In view of (3.16) and the equality $h_t = \lambda_t \mathbb{1}_{\{t \leq \tau\}}$, this yields

$$dU_t = d(V_t \tilde{H}_t) = \tilde{H}_{t-} (r_t V_t dt - (Z_t - V_{t-}) h_t dt + \tilde{B}_t dN_t) + V_{t-} d\tilde{H}_t + \Delta V_t \Delta \tilde{H}_t.$$

After rearranging and noticing that $\Delta \tilde{H}_t = -\Delta H_t$, we obtain

$$dU_t = r_t U_t dt - (Z_t + \Delta V_t) dH_t + d\tilde{N}_t, \quad (3.18)$$

where $\tilde{N}$ stands for the local $\mathbb{G}$ -martingale, more precisely,

$$d\tilde{N}_t = \tilde{H}_{t-} \tilde{B}_t dN_t + (Z_t - V_{t-}) dM_t.$$

Since $U_T = X \mathbb{1}_{\{T < \tau\}}$, formula (3.18) gives expression (3.17) (if the local martingale $\tilde{N}$ is in fact a 'true' martingale). $\square$

Corollary 3.5 *Let the processes S and V be defined by (3.5) and (3.11), respectively. Then (i)*

$$S_t = \mathbb{1}_{\{t < \tau\}} \left(V_t - B_t \mathbb{E}_{\mathbb{P}^*} (B_\tau^{-1} \mathbb{1}_{\{\tau \leq T\}} \Delta V_\tau \mid \mathcal{G}_t) \right), \quad (3.19)$$

(ii) *if $\Delta V_\tau = 0$, then $S_t = \mathbb{1}_{\{t < \tau\}} V_t$ for every $t \in [0, T]$.*

Proof A comparison of expressions (3.6) and (3.13) yields

$$S_t = U_t - B_t \mathbb{E}_{\mathbb{P}^*} (B_\tau^{-1} \mathbb{1}_{\{t < \tau \leq T\}} \Delta V_\tau \mid \mathcal{G}_t).$$

Formula (3.19) now easily follows. $\square$

For easy further reference, we shall write down the particular case of (3.19) when $\Delta V_\tau = 0$. In this case, we have simply $S_t = U_t$, that is,

$$S_t = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{G}_t \right). \quad (3.20)$$

In view of the relationship established in part (ii) of Corollary 3.5, the process V given by formula (3.11) is commonly referred to as the *pre-default value* of a defaultable claim X . A more general version of (3.20) is proved in Proposition 5 in Wong (1998). The formula there is called the *price representation theorem*.

Remarks To examine the continuity condition $\Delta V_\tau = 0$, we find it convenient to introduce additional restrictions on the underlying filtrations.⁴ It will soon become clear that we need to restrict our attention to the case of $\mathbb{F}$ -predictable processes B and Z , and to an $\mathcal{F}_T$ -measurable random variable X .

3.1 Hypotheses (H)

We shall now examine some specific assumptions related to the underlying filtrations. Let us first formulate the following hypothesis (recall that $\mathcal{F}_t \subseteq \mathcal{G}_t$ so that $\mathcal{G}_t \vee \mathcal{F}_t = \mathcal{G}_t$).

Assumption (H.1) For any t , the σ -fields $\mathcal{F}_\infty$ and $\mathcal{G}_t$ are conditionally independent given $\mathcal{F}_t$. Equivalently, for any t , and any bounded $\mathcal{F}_\infty$ -measurable r.v. ξ we have $\mathbb{E}_{\mathbb{P}^*}(\xi \mid \mathcal{G}_t) = \mathbb{E}_{\mathbb{P}^*}(\xi \mid \mathcal{F}_t)$.

Definition 3.6 We say that a filtration $\mathbb{F}$ has the *martingale invariance property* with respect to a filtration $\mathbb{G}$ if every $\mathbb{F}$ -martingale is also a $\mathbb{G}$ -martingale.

Lemma 3.7 A filtration $\mathbb{F}$ has the martingale invariance property with respect to a filtration $\mathbb{G}$ if and only if condition (H.1) is satisfied.

Proof Assume first that (H.1) holds. Let M be an arbitrary $\mathbb{F}$ -martingale. Then for any $t \leq s$ we have

$$\mathbb{E}_{\mathbb{P}^*}(M_s \mid \mathcal{G}_t) = \mathbb{E}_{\mathbb{P}^*}(M_s \mid \mathcal{F}_t) = M_t,$$

so that M is a $\mathbb{G}$ -martingale. Conversely, let us assume that every $\mathbb{F}$ -martingale is a $\mathbb{G}$ -martingale. We shall check that this implies (H.1). To this end, for any fixed $t \leq s$ we consider an arbitrary set $A \in \mathcal{F}_\infty$. We introduce the $\mathbb{F}$ -martingale $M_u := \mathbb{E}_{\mathbb{P}^*}(\mathbb{1}_A \mid \mathcal{F}_u)$, $u \in \mathbb{R}_+$. Since M is also a $\mathbb{G}$ -martingale, we obtain

$$\mathbb{E}_{\mathbb{P}^*}(\mathbb{1}_A \mid \mathcal{G}_t) = M_t = \mathbb{E}_{\mathbb{P}^*}(\mathbb{1}_A \mid \mathcal{F}_t).$$

By standard arguments this shows that (H.1) is satisfied. □

Recall that in the present setup we have $\mathbb{G} = \mathbb{H} \vee \mathbb{F}$ for a certain filtration $\mathbb{H}$. Let us introduce the following condition.

Assumption (H.2) For any t , the σ -fields $\mathcal{F}_\infty$ and $\mathcal{H}_t$ are conditionally independent given $\mathcal{F}_t$.

⁴ Notice that these hypotheses are satisfied in the widely used case of Cox processes.

Since $\mathcal{H}_t \subset \mathcal{G}_t$, it is easily seen that (H.1) is stronger than (H.2). It appears that Assumptions (H.1) and (H.2) are in fact equivalent.

Lemma 3.8 *Conditions (H.1) and (H.2) are equivalent.*

Proof It is enough to check that (H.2) implies (H.1). Condition (H.2) is equivalent to the following one: for any bounded $\mathcal{F}_\infty$ -measurable random variable ξ , we have $\mathbb{E}_{\mathbb{P}^*}(\xi \mid \mathcal{H}_t \vee \mathcal{F}_t) = \mathbb{E}_{\mathbb{P}^*}(\xi \mid \mathcal{F}_t)$. Since $\mathcal{G}_t = \mathcal{H}_t \vee \mathcal{F}_t$, this immediately gives (H.1). $\square$

Under Assumption (H.1) the conditioning with respect to $\mathcal{G}_t$ in (3.11) may be replaced by conditioning with respect to $\mathcal{F}_t$, that is, we may set

$$V_t = \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right). \quad (3.21)$$

This follows from the fact that the process N given by (see formula (3.15) in the proof of Theorem 3.4)

$$N_t = \mathbb{E}_{\mathbb{P}^*} \left(\int_0^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right) \quad (3.22)$$

is not only an $\mathbb{F}$ -martingale but also a $\mathbb{G}$ -martingale. Therefore, (3.16) gives the semimartingale decomposition of V with respect to both filtrations, $\mathbb{F}$ and $\mathbb{G}$. The remaining part of the proof of Theorem 3.4 is thus still valid. If, in addition, $\Delta V_\tau = 0$ then we have

$$S_t = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right). \quad (3.23)$$

In some particular cases – for instance when the filtration $\mathbb{F}$ is generated by a Brownian motion (under $\mathbb{P}^*$) – the continuity of the process N given by (3.22), and thus also the continuity of V is obvious. In many other important practical cases, the validity of (3.23) can be verified directly (see, e.g., Proposition 6.1 below).

In the general case, it seems more convenient to derive formula (3.23) using the standard results on intensities of random times (see, e.g., Kusuoka (1999), Rutkowski (1999), Elliott *et al.* (2000), Jeanblanc and Rutkowski (2000a, 2000b)), rather than Theorem 3.4. To this end, notice that since obviously $\mathcal{F}_t \subset \mathcal{F}_\infty$, we may restate condition (H.2) as follows:

Condition (H.3) For any $t \in \mathbb{R}_+$ and every $u \leq t$, we have $\mathbb{P}(\tau \leq u \mid \mathcal{F}_t) = \mathbb{P}(\tau \leq u \mid \mathcal{F}_\infty)$.

It is thus clear that in the present setup, the process $F_t := \mathbb{P}^*(\tau \leq t \mid \mathcal{F}_t)$ admits a modification with increasing sample paths. Assume that $F_t < 1$ for every $t \in$

$\mathbb{R}_+$. The $\mathbb{F}$ -hazard process of τ , denoted by Γ , is defined through the formula $1 - F_t = e^{-\Gamma_t}$, or, equivalently, $\Gamma_t = -\ln(1 - F_t)$ for every $t \in \mathbb{R}_+$. If F follows an absolutely continuous process, then it can be shown (see the abovementioned papers for details) that $\Gamma_t = \int_0^t \lambda_u du$, and

$$\begin{aligned} S_t &= B_t \mathbb{E}_{\mathbb{P}^*} \left(B_\tau^{-1} Z_\tau \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right) \\ &= \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u Z_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right). \end{aligned}$$

This means that under the above set of assumptions, for the process V given by (3.21) we have

$$\mathbb{E}_{\mathbb{P}^*} (B_\tau^{-1} \mathbb{1}_{\{\tau \leq T\}} \Delta V_\tau \mid \mathcal{G}_t) = 0.$$

4 Alternative recovery schemes

In this section, we shall further specify the model presented in the previous section, by introducing various kinds of recovery processes. The recent work by Wong (1998) provides an interesting study of various recovery schemes in the framework of a fairly general model. We do not present Wong's results here, however, and we refer an interested reader to the original paper. We assume throughout that (H.1) (or equivalently (H.2)) holds.

4.1 Exogenous recovery rates

Assume, as before, that Z is an exogenously given $\mathbb{F}$ -predictable process. The price process S of a defaultable claim is uniquely specified through expressions (3.5)–(3.6). It is thus clear that only the values of the process Z at default time τ are essential. Therefore, instead of specifying the $\mathbb{F}$ -predictable process Z , it is enough to consider a random variable Z_τ .

We postulate that we are given a bounded random variable, denoted by W , which models the recovery value at default time. By assumption, W is an $\mathcal{F}_\tau$ -measurable random variable, meaning that⁵ $W = Z_\tau$ for some $\mathbb{F}$ -adapted process Z . A slightly stronger assumption would be to postulate that W is an $\mathcal{F}_{\tau-}$ -measurable random variable; this would mean in turn that $W = Z_\tau$ for some $\mathbb{F}$ -predictable process Z .

Following Duffie (1998b), we shall now consider both the case of discrete-time and continual recovery of a defaultable claim with an arbitrary recovery value W . In the case of *continual recovery*, the price process S of a defaultable claim X is

⁵ Notice that τ is not necessarily an $\mathbb{F}$ -stopping time, so that $\mathcal{F}_\tau$ cannot be introduced as the 'usual' σ -field generated by an $\mathbb{F}$ -stopping time. For the more general definition of $\mathcal{F}_\tau$ -measurability we use here, see page 202 in Dellacherie and Meyer (1975).

set to satisfy (as before, we assume that the claim is of European style and it settles at time T)

$$S_t := B_t \mathbb{E}_{\mathbb{P}^*} (B_t^{-1} W \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t). \quad (4.1)$$

It appears (see Duffie (1998a) in this regard) that the results of Section 3 remain valid in the case of continual recovery with the recovery value W , provided that the recovery process Z is substituted with an $\mathbb{F}$ -predictable process $\tilde{W}$ which satisfies $\tilde{W}_\tau = \mathbb{E}_{\mathbb{P}^*}(W \mid \mathcal{G}_{\tau-})$.

A *discrete-time recovery* assumes that the payoff at the event of default is received by the owner of a claim on the first date after default among a predetermined set of *admissible* dates $0 = T_0 < T_1 < \dots < T_n = T$. Under this convention, the value process $\tilde{S}$ of a defaultable claim equals

$$\tilde{S}_t := \sum_{T_i \geq t} B_t \mathbb{E}_{\mathbb{P}^*} (B_{T_i}^{-1} W \mathbb{1}_{\{T_{i-1} < \tau \leq T_i\}} \mid \mathcal{G}_t) + B_t \mathbb{E}_{\mathbb{P}^*} (B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t). \quad (4.2)$$

In practical terms, when default occurs, the associated payoff (if any) is postponed to the nearest date T_i after default. It should be stressed that it is now enough to assume that a random variable W is such that for every $i = 1, \dots, n$, the random variable $W_i = W \mathbb{1}_{\{T_{i-1} < \tau \leq T_i\}}$ is $\mathcal{G}_{T_i}$ -measurable. Put another way, the amount which is paid to the owner of the claim at the date T_i is based on the total information which is available at this time, including the default event $\{T_{i-1} < \tau \leq T_i\}$. For technical reasons, we shall postulate that for every i we have $W_i = \hat{W}_i \mathbb{1}_{\{T_{i-1} < \tau \leq T_i\}}$, where for each i the random variable $\hat{W}_i$ is $\mathcal{F}_{T_i}$ -measurable.

It is worthwhile to observe that the valuation formula (4.2) has slightly different practical features than the basic valuation formula (3.5). Indeed, formula (3.5) implicitly assumes that a defaultable claim becomes worthless as soon as a default occurs. On the other hand, when formula (4.2) is used to value a defaultable claim, a claim becomes worthless not at the time of default, but after the nearest date from the set of admissible dates.

Our next goal is to get a more explicit expression for (4.2). For a fixed $t \leq T$, we shall write $i_0 = i_0(t) = \inf\{i : T_i \geq t\}$. It is thus clear that

$$\tilde{S}_t = \sum_{i=i_0}^n (\hat{U}_t^i - \tilde{U}_t^i) + U_t^n,$$

where

$$\hat{U}_t^i = B_t \mathbb{E}_{\mathbb{P}^*} (B_{T_i}^{-1} \hat{W}_i \mathbb{1}_{\{T_{i-1} < \tau\}} \mid \mathcal{G}_t), \quad \tilde{U}_t^i = B_t \mathbb{E}_{\mathbb{P}^*} (B_{T_i}^{-1} \hat{W}_i \mathbb{1}_{\{T_i < \tau\}} \mid \mathcal{G}_t),$$

and

$$U_t^n = B_t \mathbb{E}_{\mathbb{P}^*} (B_{T_n}^{-1} X \mathbb{1}_{\{T_n < \tau\}} \mid \mathcal{G}_t).$$

Since for every $i = i_0, \dots, n$ we have: (a) $\mathcal{G}_t \subset \mathcal{G}_{T_i}$, and (b) the random variable W_i is $\mathcal{G}_{T_i}$ -measurable, the evaluation of $\tilde{U}_t^i$, $i = 1, \dots, n$ and U_t^n is standard. Indeed, we may apply previously established results, with $Z = 0$ and $T = T_i$. To get a more transparent expression for the valuation formula, we shall assume that $\Delta V_\tau = 0$, where V stands for the pre-default value process introduced in Theorem 3.4 (since in the present context V depends on i , so that the assumption that V doesn't jump at default time is made for every i). Using (3.23), we obtain

$$\tilde{U}_t^i = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}^{\mathbb{P}^*}(\tilde{B}_{T_i}^{-1} \hat{W}_i \mid \mathcal{F}_t)$$

for $i = 1, \dots, n$, and

$$U_t^n = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}^{\mathbb{P}^*}(\tilde{B}_{T_n}^{-1} X \mid \mathcal{F}_t).$$

We may proceed in a similar way when dealing with $\hat{U}_t^i$, provided that $i \geq i_0 + 1$ (this ensures that $\mathcal{G}_t \subset \mathcal{G}_{T_{i-1}}$). To this end, we find it convenient to represent $\hat{U}_t^i$ as follows

$$\hat{U}_t^i = B_t \mathbb{E}^{\mathbb{P}^*} \left(B_{T_{i-1}}^{-1} \mathbb{E}^{\mathbb{P}^*}(B_{T_i}^{-1} \hat{W}_i \mid \mathcal{G}_{T_{i-1}}) \mathbb{1}_{\{T_{i-1} < \tau\}} \mid \mathcal{G}_t \right).$$

This means that

$$\hat{U}_t^i = B_t \mathbb{E}^{\mathbb{P}^*}(B_{T_{i-1}}^{-1} Y_i \mathbb{1}_{\{T_{i-1} < \tau\}} \mid \mathcal{G}_t),$$

where Y_i is an $\mathcal{F}_{T_{i-1}}$ -measurable random variable (in the second equality below, we make use of Assumption (H.2))

$$Y_i = B_{T_{i-1}} \mathbb{E}^{\mathbb{P}^*}(B_{T_i}^{-1} \hat{W}_i \mid \mathcal{F}_{T_{i-1}} \vee \mathcal{H}_{T_{i-1}}) = B_{T_{i-1}} \mathbb{E}^{\mathbb{P}^*}(B_{T_i}^{-1} \hat{W}_i \mid \mathcal{F}_{T_{i-1}}). \quad (4.3)$$

Notice that Y_i represents the price at time T_{i-1} of a non-defaultable claim that pays $\hat{W}_i$ at time T_i . Arguing along the same lines as before, we get

$$\hat{U}_t^i = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}^{\mathbb{P}^*}(\tilde{B}_{T_{i-1}}^{-1} Y_i \mid \mathcal{F}_t).$$

It thus remains to analyse the following term:

$$\hat{U}_t^{i_0} = B_t \mathbb{E}^{\mathbb{P}^*} \left(\mathbb{E}^{\mathbb{P}^*}(B_{T_{i_0}}^{-1} \hat{W}_{i_0} \mid \mathcal{G}_{T_{i_0-1}}) \mathbb{1}_{\{T_{i_0-1} < \tau\}} \mid \mathcal{G}_t \right).$$

Since $\mathcal{G}_{T_{i_0}} \subset \mathcal{G}_t$ and the event $\{T_{i_0-1} < \tau\}$ belongs to $\mathcal{G}_{T_{i_0-1}}$, we obtain

$$\hat{U}_t^{i_0} = \mathbb{1}_{\{T_{i_0-1} < \tau\}} B_t \mathbb{E}^{\mathbb{P}^*}(B_{T_{i_0}}^{-1} \hat{W}_{i_0} \mid \mathcal{G}_t) = \mathbb{1}_{\{T_{i_0-1} < \tau\}} Y_{i_0},$$

where Y_{i_0} represents the price at time t of a non-defaultable claim that pays $\hat{W}_{i_0}$ at time T_{i_0} . We are in a position to state the following result. Let us stress that we assume that formula (3.23) may be applied to each term $\hat{U}_t^i$ and $\tilde{U}_t^i$.

Proposition 4.1 *Let the price $\tilde{S}_t$ at time $t \leq T$ of a defaultable claim X with discrete-time recovery be given by formula (4.2). Then*

$$\begin{aligned} \tilde{S}_t &= \mathbb{1}_{\{T_{i_0-1} < \tau\}} B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_{i_0}}^{-1} \hat{W}_{i_0} \mid \mathcal{F}_t) + \mathbb{1}_{\{t < \tau\}} \sum_{i=i_0+1}^n \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_{T_{i-1}}^{-1} Y_i \mid \mathcal{F}_t) \\ &\quad - \mathbb{1}_{\{t < \tau\}} \sum_{i=i_0}^n \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_{T_i}^{-1} \hat{W}_i \mid \mathcal{F}_t) + \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_{T_n}^{-1} X \mid \mathcal{F}_t), \end{aligned}$$

where $i_0 = i_0(t) = \inf\{i : T_i > t\}$, Y_i is given by (4.3), and $\tilde{B}$ by (3.12).

We shall now focus on the case of a defaultable term structure, that is, we set $X = 1$. The most tractable cases are: (i) the case of zero recovery: $W = 0$, (ii) the case of fractional recovery of par: $W = \delta$ with $0 < \delta < 1$ (in principle, δ can be any real number). For any adapted process γ , we find it convenient to denote

$$B^\gamma(t, T) = \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^T (r_u + \gamma_u) du \right) \mid \mathcal{F}_t \right\}. \quad (4.4)$$

Notice that $B^0(t, T) = B(t, T)$, and $B^\gamma(t, T) < B(t, T)$ if γ is strictly positive.

Zero recovery

In the case of zero recovery, formulae (4.1) and (4.2) yield, as expected, the same result for the price process $D^0(t, T)$ of the T -maturity defaultable bond. Specifically, we have

$$D^0(t, T) = B_t \mathbb{E}_{\mathbb{P}^*}(B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t). \quad (4.5)$$

As usual, we assume that we are in a position to use formula (3.23) (i.e. $\Delta V_\tau = 0$). Then

$$D^0(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_T^{-1} \mid \mathcal{F}_t) = \mathbb{1}_{\{t < \tau\}} B^\lambda(t, T).$$

This means that the price of a bond before default can be calculated in a ‘standard’ way, provided that the risk-free rate r is substituted with the default-adjusted rate $R = r + \lambda$. In particular, if λ is strictly positive then $D^0(t, T) < B(t, T)$ for $t < T$, and $D^0(T, T) \leq B(T, T) = 1$.

Fractional recovery of par

In the case of a non-zero recovery coefficient δ , for the price $D^\delta(t, T)$ of a defaultable bond with continual recovery we get

$$\begin{aligned} D^\delta(t, T) &:= B_t \mathbb{E}_{\mathbb{P}^*}(\delta B_\tau^{-1} \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t) \\ &= \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\delta \int_t^T \tilde{B}_u^{-1} \lambda_u du + \tilde{B}_T^{-1} \mid \mathcal{F}_t \right), \end{aligned}$$

where the second equality holds provided that $\Delta V_\tau = 0$. The price of a defaultable bond with discrete-time recovery equals (cf. (4.2))

$$\tilde{D}^\delta(t, T) := \sum_{T_i \geq t} B_t \mathbb{E}_{\mathbb{P}^*}(\delta B_{T_i}^{-1} \mathbb{1}_{\{T_{i-1} < \tau \leq T_i\}} \mid \mathcal{G}_t) + B_t \mathbb{E}_{\mathbb{P}^*}(B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t).$$

Let us analyse the latter case in more detail. Suppose that $T_{i_0-1} \leq t < T_{i_0}$. First, we have

$$\begin{aligned} \tilde{D}^\delta(t, T) &= \delta B_t \sum_{i=i_0}^n \left(\mathbb{E}_{\mathbb{P}^*}(B_{T_i}^{-1} \mathbb{1}_{\{T_{i-1} < \tau\}} \mid \mathcal{G}_t) - \mathbb{E}_{\mathbb{P}^*}(B_{T_i}^{-1} \mathbb{1}_{\{T_i < \tau\}} \mid \mathcal{G}_t) \right) \\ &\quad + B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_n}^{-1} \mathbb{1}_{\{T_n < \tau\}} \mid \mathcal{G}_t), \end{aligned}$$

or in an abbreviated form,

$$\tilde{D}^\delta(t, T) = \sum_{i=i_0}^n \delta \hat{U}(t, T_i) - \sum_{i=i_0}^n \delta \tilde{U}(t, T_i) + U(t, T_n). \quad (4.6)$$

Since $T_{i_0-1} \leq t$ and thus $\mathcal{G}_{T_{i_0-1}} \subset \mathcal{G}_t$, it is clear that

$$\hat{U}(t, T_{i_0}) = B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_{i_0}}^{-1} \mathbb{1}_{\{T_{i_0-1} < \tau\}} \mid \mathcal{G}_t) = \mathbb{1}_{\{T_{i_0-1} < \tau\}} B(t, T_{i_0}). \quad (4.7)$$

Furthermore, for any $i = i_0 + 1, \dots, n$ we have $\mathcal{G}_t \subset \mathcal{G}_{T_{i-1}}$, and thus

$$\begin{aligned} \hat{U}(t, T_i) &= B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_i}^{-1} \mathbb{1}_{\{T_{i-1} < \tau\}} \mid \mathcal{G}_t) \\ &= B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_{i-1}}^{-1} \mathbb{1}_{\{T_{i-1} < \tau\}} B(T_{i-1}, T_i) \mid \mathcal{G}_t). \end{aligned}$$

By applying (3.23), we get (as usual, we assume that V does not jump at τ)

$$\hat{U}(t, T_i) = \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_{i-1}} (r_u + \lambda_u) du \right) B(T_{i-1}, T_i) \mid \mathcal{F}_t \right\},$$

or equivalently (cf. (4.4))

$$\begin{aligned} \hat{U}(t, T_i) &= \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_i} (r_u + \lambda_u \mathbb{1}_{[0, T_{i-1}]}(u)) du \right) \mid \mathcal{F}_t \right\} \\ &= \mathbb{1}_{\{t < \tau\}} B^{\lambda^{i-1}}(t, T_i), \end{aligned} \quad (4.8)$$

where we set $\lambda_t^{i-1} = \lambda_t \mathbb{1}_{[0, T_{i-1}]}(t)$ for $t \in [0, T]$. Finally, once again using (3.23), we get for any $i = i_0, \dots, n$

$$\begin{aligned} \tilde{U}(t, T_i) &= B_t \mathbb{E}_{\mathbb{P}^*}(B_{T_i}^{-1} \mathbb{1}_{\{T_i < \tau\}} \mid \mathcal{G}_t) \\ &= \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_i} (r_u + \lambda_u) du \right) \mid \mathcal{F}_t \right\}, \end{aligned} \quad (4.9)$$

so that

$$\tilde{U}(t, T_i) = \mathbb{1}_{\{t < \tau\}} B^\lambda(t, T_i) = D^0(t, T_i).$$

By plugging (4.7)–(4.9) into (4.6), we arrive at the following representation of the price $\tilde{D}^\delta(t, T)$.

Proposition 4.2 *Let $I_0 := \mathbb{1}_{\{T_{i_0-1} < \tau\}} \delta B(t, T_{i_0})$. For every $t \leq T$, the price $\tilde{D}^\delta(t, T)$ of a defaultable bond with discrete-time fractional recovery of par equals*

$$\begin{aligned} \tilde{D}^\delta(t, T) = & I_0 + \mathbb{1}_{\{t < \tau\}} \sum_{i=i_0+1}^n \delta \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_i} (r_u + \lambda_u^{i-1}) du \right) \middle| \mathcal{F}_t \right\} \\ & - \mathbb{1}_{\{t < \tau\}} \sum_{i=i_0}^n \delta \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_i} (r_u + \lambda_u) du \right) \middle| \mathcal{F}_t \right\} \\ & + \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^{T_n} (r_u + \lambda_u) du \right) \middle| \mathcal{F}_t \right\}, \end{aligned}$$

where $i_0 = i_0(t) = \inf\{i : T_i > t\}$ and $\lambda_t^{i-1} = \lambda_t \mathbb{1}_{[0, T_{i-1}]}(t)$. Put another way,

$$\tilde{D}^\delta(t, T) = I_0 + \mathbb{1}_{\{t < \tau\}} \left(\sum_{i=i_0+1}^n \delta B^{\lambda^{i-1}}(t, T_i) - \sum_{i=i_0}^n \delta B^\lambda(t, T_i) + B^\lambda(t, T_n) \right).$$

Example 4.3 Let us consider a very special case of a T -maturity defaultable bond with a discrete-time recovery, with only two admissible dates $T_0 = 0$ and $T_1 = T$. Since default at time 0 is excluded with probability 1, it is clear that the payment always occurs at time T , no matter whether a bond has defaulted before maturity or not. For any $t \leq T$ we have

$$\tilde{D}^\delta(t, T) = B_t \mathbb{E}_{\mathbb{P}^*} (\delta B_T^{-1} \mathbb{1}_{\{0 < \tau \leq T\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t).$$

On the other hand, since $i_0(t) = 1$ for any $t \leq T$, formula the established in Proposition 4.2 gives

$$\tilde{D}^\delta(t, T) = \delta B(t, T) + \mathbb{1}_{\{t < \tau\}} (1 - \delta) B^\lambda(t, T). \quad (4.10)$$

Under the present assumptions, since a defaulted bond pays the amount δ at time T , we get $\tilde{D}^\delta(t, T) = \delta B(t, T)$ on the random set $[\tau, T]$, that is, after default. Before default, its value is strictly greater than $\delta B(t, T)$, but we have always $\tilde{D}^\delta(t, T) < B(t, T)$. The last inequality is trivial, since the process λ is strictly positive, and thus $B^\lambda(t, T) < B(t, T)$ for every $t \leq T$. We conclude that under the present assumptions, the price of the defaultable bond never exceeds the price of the risk-free bond,⁶ which is a natural property to require from a model valuing risky debt. On the other hand, for the general model of the continual recovery we have only the following equivalence, which holds on the set $\{\tau > t\}$: the inequality $\tilde{D}^\delta(t, T) \leq B(t, T)$ holds if and only if $\delta \mathbb{E}_{\mathbb{P}^*} (B_\tau^{-1} \mathbb{1}_{\{t < \tau \leq T\}} \mid \mathcal{G}_t) \leq \mathbb{E}_{\mathbb{P}^*} (B_T^{-1} \mathbb{1}_{\{t < \tau \leq T\}} \mid \mathcal{G}_t)$. Of

⁶ This holds true also in the case of zero recovery.

course, $D^\delta(t, T) = 0 < B(t, T)$ on $\{\tau \leq t \leq T\}$. This shows that the valuation in the case of the continual fractional recovery appears to be rather delicate.

4.2 Endogenous recovery rules

If Z is not an exogenously given process (but, for instance, a deterministic function of the value process S), the problem of existence and uniqueness of a process S defined by (3.5) arises. We take the uniqueness of solution to (3.5) for granted, and we address the problem of pricing of defaultable claims of the form (X, Z, τ) , where Z is a specific ‘recovery rule’, rather than a given process.

Fractional recovery of market value

Following Duffie and Singleton (1999), we assume that $Z_t = (1 - L_t)S_{t-}$, where S is an unknown process, and L is a given $\mathbb{F}$ -predictable process. We start with the following lemma, which deals with the process V only. Notice that formula (4.11) represents a stochastic equation which needs to be solved for the unknown $\mathbb{F}$ -adapted process V .

Lemma 4.4 *Under (H.1), let V satisfy (3.11) with $Z_t = (1 - L_t)V_{t-}$ for some predictable process L , that is,*

$$V_t = \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} (1 - L_u) V_u \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right). \quad (4.11)$$

Then V is unique, and it is given by the formula

$$V_t = \hat{B}_t \mathbb{E}_{\mathbb{P}^*} (\hat{B}_T^{-1} X \mid \mathcal{F}_t), \quad (4.12)$$

where the $\mathbb{F}$ -adapted process $\hat{B}$ equals

$$\hat{B}_t = \exp \left(\int_0^t (r_u + \lambda_u L_u) du \right). \quad (4.13)$$

Proof In view of (3.16) with N is given by (3.22), we obtain

$$dV_t = V_t(r_t + \lambda_t) dt - (1 - L_t)V_t\lambda_t dt + \tilde{B}_t dN_t,$$

or equivalently,

$$dV_t = V_t(r_t + \lambda_t L_t) dt + \tilde{B}_t dN_t.$$

This immediately yields (4.12) (as usual, we assume that the last term follows a martingale). Of course, this proves also that equation (4.11) admits a unique solution. $\square$

The next step is to examine the relationship between the process V (or rather $U_t = \mathbb{1}_{\{t < \tau\}} V_t$) and the price process of a defaultable claim. In view of Theorem 3.4 (which we may apply since $Z_t = (1 - L_t)V_{t-}$ follows an $\mathbb{F}$ -predictable process), we find that U satisfies

$$U_t = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_{\tau}^{-1} ((1 - L_{\tau})V_{\tau-} + \Delta V_{\tau}) \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \middle| \mathcal{G}_t \right). \quad (4.14)$$

Corollary 4.5 *Let the process V be given by formula (4.11) for some predictable process L . Assume that $\Delta V_{\tau} = 0$. Then the process $U_t = \mathbb{1}_{\{t < \tau\}} V_t$ satisfies*

$$U_t = \mathbb{1}_{\{t < \tau\}} \hat{B}_t \mathbb{E}_{\mathbb{P}^*} (\hat{B}_T^{-1} X \mid \mathcal{F}_t) \quad (4.15)$$

and

$$U_t = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_{\tau}^{-1} (1 - L_{\tau}) U_{\tau-} \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \middle| \mathcal{G}_t \right). \quad (4.16)$$

Proof Equality (4.15) is an immediate consequence of (4.12). The second formula follows from (4.14) (we use the trivial equality $U_{\tau-} = V_{\tau-}$). $\square$

In view of Corollary 4.5, the process U satisfies equation (4.16), that is, the implicit definition of the price process S . Note that we have not proved that the uniqueness of solutions holds for the equation

$$S_t = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_{\tau}^{-1} (1 - L_{\tau}) S_{\tau-} \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \middle| \mathcal{G}_t \right). \quad (4.17)$$

We have merely shown that (4.17) admits a solution. The uniqueness of solutions to (4.17) can be deduced from standard results on backward SDEs, however. To this end, it might be convenient to use the equivalent representation of equation (4.17), i.e. (cf. (3.9))

$$S_t = \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T S_u ((1 - L_u) h_u - r_u) du + X \mathbb{1}_{\{T < \tau\}} \middle| \mathcal{G}_t \right). \quad (4.18)$$

For the existence and uniqueness of adapted solutions to backward SDEs like (4.18) see, for instance, Theorem 2.4 in Antonelli (1993).

General recovery rule

In principle, we may also deal with a ‘general recovery rule’, more precisely, we may assume that the payoff process Z satisfies $Z_t = p(t, S_{t-})$, where the function $p(t, s)$ is Lipschitz continuous with respect to s , and satisfies $p(t, 0) = 0$. In this case, however, we have merely the following result, which again is a consequence of Theorem 3.4 (once again, the problem of existence and uniqueness of solutions to (4.20) and (4.22) is not addressed here; this follows from standard results on backward SDEs).

Corollary 4.6 *Let S be the unique solution to the backward SDE*

$$S_t = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_{\tau}^{-1} p(\tau, S_{\tau-}) \mathbb{1}_{\{t < \tau \leq T\}} + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right), \quad (4.19)$$

or equivalently, to the equation (cf. (3.9))

$$S_t = \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T (p(u, S_u) h_u - r_u S_u) du + X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right). \quad (4.20)$$

Let V be the unique solution to the backward SDE

$$V_t = \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} p(u, V_u) \lambda_u du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right), \quad (4.21)$$

or equivalently, to the equation

$$V_t = \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T (p(u, V_u) \lambda_u - (r_u + \lambda_u) V_u) du + X \mid \mathcal{F}_t \right). \quad (4.22)$$

If $\Delta V_{\tau} = 0$, then $S_t = \mathbb{1}_{\{t < \tau\}} V_t$. Otherwise, S is given by formula (3.19).

For other applications of backward SDEs in mathematical finance, and further references, see the papers by Antonelli (1993), El Karoui and Quenez (1997a, 1997b) and El Karoui *et al.* (1997).

5 Credit-ratings-based Markov model

To produce a tractable model which accounts for the migration between rating grades, Jarrow *et al.* (1997) make the following, rather stringent, assumptions: (i) there exists a unique equivalent martingale measure $\mathbb{P}^*$ making all default-free and default-risky zero coupon bond prices martingales, after normalization by the savings account, (ii) the default time τ is independent of the risk-free rate r under the martingale measure $\mathbb{P}^*$, (iii) the recovery coefficient is a constant δ . They first develop a discrete-time model which takes into account the migration of a defaultable bond in the finite set of *credit rating classes*. Subsequently, a continuous-time counterpart is also examined. Methodology developed in Jarrow *et al.* (1997) is a direct extension of the approach in Jarrow and Turnbull (1995). They assume that a defaulted bond pays at maturity a fraction of its par value.⁷ Therefore, the price at time $t \leq T$ of a T -maturity defaultable bond equals

$$\tilde{D}^{\delta}(t, T) = B_t \mathbb{E}_{\mathbb{P}^*} \left(B_T^{-1} (\delta \mathbb{1}_{\{\tau \leq T\}} + \mathbb{1}_{\{T < \tau\}}) \mid \mathcal{G}_t \right), \quad (5.1)$$

where τ is the default time, and δ is the constant recovery rate. Suppose that we have chosen a model for the short-term rate r . It is clear from expression (5.1) that

⁷ This convention coincides with the concept of discrete-time fractional recovery of par introduced in Section 4, provided that we take $T_0 = 0$ and $T_1 = T$ (cf. Example 4.3).

we need only model a random time τ . In addition, under assumption (i), formula (5.1) can be substantially simplified, specifically,

$$\tilde{D}^\delta(t, T) = B(t, T) \mathbb{E} \mathbb{P}^* (\delta \mathbb{I}_{\{\tau \leq T\}} + \mathbb{I}_{\{T < \tau\}} \mid \mathcal{G}_t). \quad (5.2)$$

Consequently (it might be instructive to compare (5.3) with (4.10)),

$$\tilde{D}^\delta(t, T) = B(t, T) (\delta + (1 - \delta) \mathbb{P}^* \{T < \tau \mid \mathcal{G}_t\}). \quad (5.3)$$

As will soon become clear, the stopping time τ is explicitly dependent on the initial rating of a particular bond. Therefore, expressions (5.1)–(5.3) should be seen as generic valuation formulae for defaultable bonds. Given an initial rating of a defaultable bond, the future changes in its assessments by a rating agency are described by a stochastic process, referred to as the *migration process*. Formally, for a given bond, the value at time t of the associated migration process coincides with its current rating. There is no loss of generality, if we assume that the set of rating classes of is $\{1, \dots, K\}$, where the state K is assumed to correspond to the default event. It is assumed that the migration process, C say, follows a Markov chain (under both real-world probability $\mathbb{P}$ and the spot martingale measure $\mathbb{P}^*$), that is, the future evolution of ratings classes of a particular bond does not depend on the bond's history, but only on its current rating.

5.1 Discrete-time model

In a discrete-time setup, the migration process and the default time are assumed to satisfy: (iv) the migration process C follows, under the real-world probability $\mathbb{P}$, a time-homogeneous Markov chain with the transition matrix (by definition, $p_{ij} = \mathbb{P}\{C_{t+1} = j \mid C_t = i\}$)

$$P = [p_{ij}]_{1 \leq i, j \leq K}, \quad p_{i,j} \geq 0, \quad \sum_{j=1}^K p_{ij} = 1,$$

with $p_{Kj} = 0$ for every $j < K$ (so that $p_{KK} = 1$; that is, the state K is absorbing), and (v) C follows a (time-inhomogeneous) Markov chain under $\mathbb{P}^*$, with time-dependent transition matrix

$$Q(t) = [q_{ij}(t, t+1)]_{1 \leq i, j \leq K}$$

where

$$q_{ij}(t, t+1) \geq 0, \quad \sum_{j=1}^K q_{ij}(t, t+1) = 1,$$

and finally $q_{Kj}(t, t+1) = 0$ for every $j < K$ and t (so that once again the state K is absorbing).

The default time τ is the first moment the rating process hits the state K (the horizon date T^* is assumed to be a natural number). Formally,

$$\tau := \inf \{ t \in \{0, 1, \dots, T^*\} : C_t = K \} \quad (5.4)$$

where, by convention, the infimum over an empty set equals $+\infty$.

To ensure analytical tractability of the model, an additional ‘technical’ assumption is made. Specifically, it is postulated that the following relationship holds

$$q_{ij}(t, t+1) = \pi_i(t)p_{ij}, \quad \forall i \neq j, \quad (5.5)$$

where time-dependent coefficients $\pi_i(t)$ are interpreted as discrete-time *risk premia*. The last assumption implies, in particular, that

$$q_{ii}(t, t+1) = 1 + \pi_i(t)(p_{ii} - 1), \quad \forall i.$$

In other words, for any state i , the probability under the martingale measure $\mathbb{P}^*$ of jumping to the state $j \neq i$ is assumed to be proportional to the corresponding probability under the real-world probability $\mathbb{P}$, with the proportionality factor which may depend on i and t , but not on j .

Assume that we are given the initial term structures of default-free and defaultable bonds, and the real-world transition matrix P (in principle, all these quantities can be ‘observed’). Then, under the above set of assumptions, Jarrow *et al.* (1997) offer a recursive procedure which leads to the unique determination of the ‘risk premium’ process $\pi(t)$, $t = 0, \dots, T^* - 1$. Consequently, the time-dependent transition matrix $Q(t)$ under $\mathbb{P}^*$ is also uniquely specified.

5.2 Continuous-time model

A similar approach is developed in the continuous-time setup. It is postulated that: (iv') under the real-world probability $\mathbb{P}$, the migration process C follows a time-homogeneous Markov chain, with intensity matrix $\tilde{\Lambda}$ satisfying mild ‘technical’ conditions (which guarantee that the state K is absorbing, and a suitable monotonicity of default probabilities holds), (v') under the martingale measure $\mathbb{P}^*$, the migration process also follows a Markov chain, but with a possibly time-dependent intensity matrix Λ_t . As before, the default time τ is the first time the rating process hits the absorbing state K . Tractability condition (5.5) now takes the following form: there exists a diagonal matrix U , whose first $K - 1$ entries, $U_{ii}(t)$, $i = 1, \dots, K - 1$, are strictly positive deterministic functions, and the last entry, $U_{KK}(t) = 1$ for every t , such that the risk-neutral and real-world intensity matrices satisfy

$$\Lambda_t = U(t)\tilde{\Lambda}, \quad \forall t \in [0, T^*]. \quad (5.6)$$

Suppose that the initial term structures of default-free and default-risky zero coupon bonds are known. Then for any choice of the ‘historical’ intensity matrix $\tilde{\Lambda}$, one can produce a model for defaultable term structure in two steps. In the first step, we construct the migration process C under the real-world probability $\mathbb{P}$, using the intensity matrix $\tilde{\Lambda}$ (by assumption, the migration process is independent of the underlying risk-free short-term rate r). Subsequently, we search for an equivalent probability measure $\mathbb{P}^*$, which would reproduce the observed prices of all defaultable bonds through the risk-neutral valuation formula (5.3). If we denote by $\tilde{D}_i^\delta(0, T)$ the initial price of the defaultable bond which belongs to the i^{th} rating class at time 0, then we have

$$\tilde{D}_i^\delta(0, T) = B(0, T) (\delta + (1 - \delta)\mathbb{P}^*\{T < \tau \mid C_0 = i\}). \quad (5.7)$$

Since τ is the hitting time of K , and the state K is absorbing, it is also clear that

$$\mathbb{P}^*\{T < \tau \mid C_0 = i\} = \mathbb{P}^*\{C_T = K \mid C_0 = i\} = q_{iK}(0, T),$$

where $Q(0, T) = [q_{ij}(0, T)]_{1 \leq i, j \leq K}$ is the transition matrix corresponding to the time interval $[0, T]$.

6 Modelling with state variables

In this section – in which we follow Duffie and Singleton (1999) and Lando (1998) – we place ourselves again within the general framework, as presented in Section 3. In order to make the model of Section 3 analytically more tractable, we impose additional conditions on the default time τ – more specifically, on the intensity process λ of the default process H . It should be stressed that additional conditions of this kind are complementary to those considered in Section 5. For instance, it seems natural to examine a model of defaultable debt which combines the presence of the migration process C with the presence of the state variables process Y (as, for instance, in Lando (1998)).

We assume that we are given a k -dimensional stochastic process Y defined on the underlying filtered probability space $(\Omega, \mathbb{F}, \mathbb{P}^*)$. The $\mathbb{F}$ -adapted process Y , which typically is assumed to be Markovian under the spot martingale measure $\mathbb{P}^*$, is assumed to model the dynamics of ‘state variables’ which underpin the evolution of all other variables in our model of the economy. As far as the default time is concerned, we postulate that τ is the first jump time of a Cox process, N say, with the stochastic intensity of the form $\lambda_t = \lambda(Y_t)$, for some function $\lambda : \mathbb{R}^k \rightarrow \mathbb{R}_+$. It is thus clear that the intensity of a default time is an $\mathbb{F}$ -adapted stochastic process.

Let us mention that at this stage no explicit distinction between defaultable bonds with different rating assessments is made. In other words, we focus on a bond

which currently belongs to a particular class, and we exclude the possibility of the bond's migration to any other class but to the 'default class'.

The construction of the default time τ with these properties can be achieved as follows. Let $\mathbb{F}$ be the filtration with respect to which the process Y is adapted, and let η be a random variable independent of $\mathbb{F}$. Of course, η and Y are assumed to be defined on a common probability space $(\Omega, \mathcal{G}, \mathbb{P}^*)$, so that a suitable enlargement of the underlying probability space might be required. More specifically, we assume that η has a unit exponential probability law under $\mathbb{P}^*$. To define default time τ (that is, the first jump of the Cox process), we set

$$\tau = \inf \left\{ t \in \mathbb{R}_+ : \int_0^t \lambda(Y_u) du \geq \eta \right\}. \quad (6.1)$$

It should be stressed that the above construction implies validity of the hypothesis (H.1).

To get a neat valuation formula for this specification of the default time τ , we need to assume, in addition, that the promised claim X is an $\mathcal{F}_T$ -measurable random variable, that the recovery process Z is $\mathbb{F}$ -predictable, and, for instance, that $r_t = r(Y_t)$ (this agrees with our interpretation of Y as a state-variables process). Under this set of assumptions, in all previously established formulae in which the default time τ does not appear explicitly, that is, the presence of the default process N is manifested only through its intensity process $\lambda_t = \lambda(Y_t)$, we may replace the conditional expectation with respect to $\mathcal{G}_t$ by conditioning with respect to $\mathcal{F}_t$. For instance, using (3.23), we obtain

$$S_t = \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T e^{-\int_t^u R(Y_v) dv} Z_u \lambda(Y_u) du + e^{-\int_t^T R(Y_v) dv} X \mid \mathcal{F}_t \right), \quad (6.2)$$

where $R(Y_u) = r(Y_u) + h(Y_u)$. Let us notice that formula (6.2) is a direct consequence of equality (3.20), combined with the simple observation that $\mathcal{F}_t \subset \mathcal{G}_t \subset \mathcal{F}_t \vee \sigma(\eta)$, where, by assumption, the σ -fields $\mathcal{F}_T$ and $\sigma(\eta)$ are mutually independent. As shown by Lando (1998), formula (6.2) can be derived in a more straightforward way, without making explicit reference to the pre-default value process V (that is, using directly Lemma 3.3 rather than a suitable version of Corollary 3.5).

Proposition 6.1 *Let the default time τ be given by (6.1). Then we have*

$$S_t = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda(Y_u) du + \tilde{B}_T^{-1} X \mid \mathcal{F}_t \right), \quad (6.3)$$

where the process $\tilde{B}$ is given by (3.12) with $r_u = r(Y_u)$ and $\lambda_u = h(Y_u)$.

Proof Notice that for any $0 \leq t \leq u$ we have

$$\mathbb{P}^*\{\tau > u \mid \mathcal{F}_T \vee \mathcal{H}_t\} = \begin{cases} \exp\left(-\int_t^u \lambda(Y_v) dv\right), & \text{on the set } \{\tau > t\}, \\ 0, & \text{otherwise,} \end{cases}$$

where, as before, $\mathcal{H}_t = \sigma(H_u : u \leq t)$. Therefore (cf. (3.8)),

$$\begin{aligned} S_t &= B_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T B_u^{-1} Z_u \lambda(Y_u) \mathbb{1}_{\{u \leq \tau\}} du + B_T^{-1} X \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t \right) \\ &= B_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T B_u^{-1} Z_u \lambda(Y_u) \mathbb{P}^*\{\tau \geq u \mid \mathcal{F}_T \vee \mathcal{H}_t\} du \mid \mathcal{G}_t \right) \\ &\quad + B_t \mathbb{E}_{\mathbb{P}^*} \left(B_T^{-1} X \mathbb{P}^*\{\tau > T \mid \mathcal{F}_T \vee \mathcal{H}_t\} \mid \mathcal{G}_t \right) \\ &= \mathbb{1}_{\{t < \tau\}} B_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T B_u^{-1} Z_u \lambda(Y_u) \exp\left(-\int_t^u \lambda(Y_v) dv\right) du \mid \mathcal{G}_t \right) \\ &\quad + \mathbb{1}_{\{t < \tau\}} B_t \mathbb{E}_{\mathbb{P}^*} \left(B_T^{-1} X \exp\left(-\int_t^T \lambda(Y_v) dv\right) \mid \mathcal{G}_t \right) \\ &= \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*} \left(\int_t^T \tilde{B}_u^{-1} Z_u \lambda(Y_u) du + \tilde{B}_T^{-1} X \mid \mathcal{G}_t \right). \end{aligned}$$

We wish now to substitute $\mathcal{G}_t$ with $\mathcal{F}_t$ in the last expression. It is enough to observe that conditioning with respect to $\mathcal{G}_t$ coincides in our case with conditioning with respect to $\mathcal{F}_t \vee \mathcal{H}_t \subset \mathcal{F}_t \vee \sigma(\eta)$. Equality (6.3) now follows immediately from the fact that the random variable η is independent of $\mathcal{F}_T$, and thus σ -fields $\mathcal{F}_T$ and $\mathcal{H}_t$ are conditionally independent given $\mathcal{F}_t$ (cf. the hypothesis (H.2)). Since the random variable under the sign of the conditional expectation is measurable with respect to the σ -field $\mathcal{F}_T$, the result follows. $\square$

Proposition 6.1 combined with Corollary 3.5 suggest that the jump ΔV_τ , even if it does not vanish, plays no longer an important role in the present setup. Indeed, it shows that in the present setup we have $S_t = \mathbb{1}_{\{t < \tau\}} V_t$, where the process V is given by (3.11). Consequently, combining (3.6) with (3.13), we find that under the present assumptions the pre-default process associated with any defaultable claim (X, Z, τ) satisfies

$$\mathbb{E}_{\mathbb{P}^*} (B_\tau^{-1} \Delta V_\tau \mathbb{1}_{\{t < \tau \leq T\}} \mid \mathcal{G}_t) = 0, \quad \forall t \in [0, T].$$

Remarks Duffie and Singleton (1999) focus on the special case of fractional recovery of market value. They assume that: (i) there is a state-variables process Y that is Markovian under the spot martingale measure $\mathbb{P}^*$, (ii) the promised contingent claim is of the form $X = g(Y_T)$ for some function $g : \mathbb{R}^k \rightarrow \mathbb{R}$, (iii) the default-adjusted short-term rate $R_t = r_t + \lambda_t L_t = \rho(Y_t)$ for some function

$\rho : \mathbb{R}^k \rightarrow \mathbb{R}$. Under (i)–(iii), we have

$$V_t = \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^T \rho(Y_u) du \right) g(Y_T) \mid Y_t \right\}. \quad (6.4)$$

Moreover, if Y follows a non-degenerate diffusion process, then $\Delta V_\tau = 0$ and thus $S_t = \mathbb{1}_{\{t < \tau\}} V_t$. Indeed, in this case the martingale N given by formula (3.22) is continuous. Consequently, in view of (3.14), the process V is also continuous.

6.1 Conditionally Markov ratings process

We shall now describe an extension – due to Lando (1998) – of the credit ratings model elaborated by Jarrow *et al.* (1997). As usual, we assume that the spot martingale measure $\mathbb{P}^*$ and risk-free term structure $B(t, T)$ are given. Lando (1998) modifies the Jarrow–Lando–Turnbull approach by introducing a conditionally Markov migration process, which accounts for both the presence of different rating classes and the postulated existence of the underlying state variables, as modelled by a process Y . It appears that this can be achieved by a suitable modification of the migration process C introduced in Section 5 (whenever possible, we preserve the notation introduced in Section 5).

We place ourselves in a continuous-time setup. The migration process C is now assumed to follow, under the spot martingale measure, a conditional Markov chain with the stochastic intensity matrix $\Lambda(Y_t) = [\lambda_{ij}(Y_t)]_{1 \leq i, j \leq K}$ which is assumed to satisfy, for every $t \in [0, T^*]$ and $i = 1, \dots, K$,

$$\lambda_{ii}(Y_t) = - \sum_{j=1, j \neq i}^K \lambda_{ij}(Y_t), \quad \text{and} \quad \lambda_{K,i}(Y_t) = 0, \quad (6.5)$$

where $\lambda_{ij} : \mathbb{R}^k \rightarrow \mathbb{R}_+$ are non-negative functions. For any such a matrix, given the process Y and the initial rating i (at time 0, say), it is possible to construct a migration process C corresponding to the matrix $\Lambda(Y_t)$. More specifically, the migration process C is assumed to follow, conditionally on the path of the state-variables process Y , a Markov chain with finite state space $\{1, \dots, K\}$ and time-dependent (but deterministic) intensity matrix $\Lambda(Y_t)$. It follows from (6.5) that the K^{th} row of the matrix $\Lambda(Y_t)$ is assumed to vanish identically, so that K is an absorbing state. As in Section 5, the absorbing state K represents the default event, and the default time is the first time the migration process C hits K . The construction of a process C with these properties is a straightforward generalization of the construction of a default time provided by formula (6.1) (though we need to deal with an infinite family of mutually independent exponentially distributed random variables).

Remarks The migration process C can be seen as a generalization of the first jump process H introduced in Section 3. Recall that H was defined through the formula $H_t = \mathbb{1}_{\{t \geq \tau\}}$. If we put $C_t = 1 + H_t$ then the state space of C is $\{1, 2\}$ with 2 being the absorbing state. In a general framework, the process $C_t = 1 + H_t$ is not necessarily a (conditionally) Markov process, however.

Due to the nature of the default time τ , the valuation of defaultable claims becomes more cumbersome. It is essential to note that the default time τ and short-term rate r are no longer mutually independent (as was postulated in Jarrow *et al.* (1997)). Therefore, no explicit valuation results, such as formula (5.3), are available in the present setup. Consequently, one is bound to employ the basic definition (3.6) of the price process of a defaultable claim. This observation applies also to the case of a zero coupon bond, under the assumption that the recovery rate equals 0 (that is, when the recovery process Z vanishes identically). By definition, the price of such a bond equals (cf. (3.6) or (4.5))

$$D_i^0(t, T) = B_t \mathbb{E}_{\mathbb{P}^*} (B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{F}_t \vee \{C_t = i\}),$$

where we assume that at time t the bond belongs to the i^{th} rating class, for some $i < K$. Using a similar reasoning as in the proof of Proposition 6.1 (that is, conditioning first on the future evolution of the process Y), we find that

$$D_i^0(t, T) = B_t \mathbb{E}_{\mathbb{P}^*} (B_T^{-1} (1 - p_{iK}^Y(t, T)) \mid \mathcal{F}_t), \quad (6.6)$$

where

$$p_{iK}^Y(t, T) = \mathbb{P}^* \{C_T = K \mid \{C_t = i\} \vee \sigma(Y_u : u \in [t, T])\}. \quad (6.7)$$

Notice that $p_{iK}^Y(t, T)$ is simply the conditional transition probability of the migration process C , over the time interval $[t, T]$, with conditioning on the future behaviour of the state-variables process Y . Evaluation of the conditional probability $p_{iK}^Y(t, T)$, given a particular sample path of the process Y , would be thus a relatively simple task in the case of a diagonal intensity matrix $\Lambda(Y_t)$. Indeed, we would be then able to separate variables in the corresponding system of Kolmogorov differential equations. A similar – but slightly less explicit – result holds provided that

$$\Lambda(Y_t) = B \Gamma(Y_t) B^{-1},$$

where $\Gamma(Y_t)$ is a diagonal matrix, and B is a $K \times K$ matrix whose columns are the eigenvectors of $\Lambda(Y_t)$. Under this rather restrictive condition, Lando (1998) derived a quasi-explicit valuation formula for a defaultable bond, and indeed for any (promised) European claim of the form $X = g(Y_T, C_T)$.

To conclude, the problem of valuation of defaultable debt is reduced to that of finding a convenient representation of the right-hand side in (6.7), which would

subsequently allow us to evaluate the conditional expectation in (6.6). Generally speaking, this seems to be a rather difficult task, especially when restrictive regularity conditions are not imposed on the intensity matrix, or when we deal with a non-zero recovery rate. In any case, valuation of defaultable claims can be done through simulation techniques, though.

7 Credit-spreads-based HJM type model

Results presented in this section are mainly due to Bielecki and Rutkowski (1999, 2000) (for related results, see Schönbucher (1998)). In contrast to the previous sections, we shall no longer assume that the default time of a T -maturity defaultable bond is prespecified. We postulate instead that we start with a given default-free and defaultable term structure, represented by a finite family of defaultable instantaneous forward rates. Our aim is thus to support an exogenously given defaultable term structure through an associated family of default times, defined on a suitable enlargement of the underlying probability space.

It should thus be stressed that in this section we are no longer concerned with the valuation of defaultable bonds for a given risk-free term structure and a given recovery rate. On the contrary, we assume that the ‘pre-default’ values of defaultable bonds are given *a priori*, and we search for an arbitrage-free bond market model that supports these values.

7.1 Single credit rating case

In the first step, we focus on a defaultable bond from a given rating class and we assume that it cannot migrate to another class before default. We assume that the dynamics of defaultable instantaneous forward rates are given. Our goal is to explain these dynamics by introducing a judiciously chosen stopping time (on an enlarged probability space), which is interpreted as the bond’s default time. Throughout this section the focus is on the case of fractional recovery of treasury value (that is, a fixed fraction of the nominal value is received at the bond’s maturity, if default occurs before or at maturity). We make the following standing assumptions:

(B.1) We are given a d -dimensional standard Brownian motion W , defined on the underlying (real-world) filtered probability space $(\Omega, \mathbb{F}, \mathbb{P})$.

(B.2) For any fixed maturity $T \leq T^*$, the default-free instantaneous forward rate $f(t, T)$ satisfies⁸

$$df(t, T) = \alpha(t, T) dt + \sigma(t, T) \cdot dW_t, \quad (7.1)$$

where α and σ are adapted processes with values in $\mathbb{R}$ and $\mathbb{R}^d$, respectively.

(B.D) The defaultable instantaneous forward rate $g(t, T)$ satisfies

$$dg(t, T) = \tilde{\alpha}(t, T) dt + \tilde{\sigma}(t, T) \cdot dW_t, \quad (7.2)$$

for some processes $\tilde{\alpha}$ and $\tilde{\sigma}$.

Conditions (B.1)–(B.2) are the standard hypotheses of the Heath *et al.* (1992) approach to term structure modelling. By definition, the price at time t of a T -maturity default-free zero coupon bond thus equals

$$B(t, T) := \exp\left(-\int_t^T f(t, u) du\right). \quad (7.3)$$

The relevance of assumption (B.D) will be discussed later. For any $t \leq T$, we set

$$\tilde{D}(t, T) := \exp\left(-\int_t^T g(t, u) du\right), \quad (7.4)$$

and we interpret $\tilde{D}(t, T)$ as the pre-default value of a T -maturity defaultable zero coupon bond with fractional recovery of par. In other words, we interpret $\tilde{D}(t, T)$ as the value of a T -maturity defaultable zero coupon bond conditioned on the fact the bond had not defaulted by the time t . To justify this heuristic interpretation, we need first to develop an arbitrage-free model for default-free and defaultable term structures. Our main goal will be then to show that the pre-default value $\tilde{D}(t, T)$ can be seen as the price before default of a T -maturity defaultable zero coupon bond in this framework. We assume, in addition, that the *credit spread* $g(t, T) - f(t, T)$ is strictly positive, so that $\tilde{D}(t, T) < B(t, T)$ (the case of $\delta = 1$ is thus excluded as trivial).

Default-free term structure

For the reader's convenience, we quote the following well-known result (see Heath *et al.* (1992)).

Lemma 7.1 *The dynamics of the default free bond price $B(t, T)$ are*

$$dB(t, T) = B(t, T)(a(t, T) dt + b(t, T) \cdot dW_t), \quad (7.5)$$

⁸ For technical conditions under which formulae (7.1)–(7.2) make sense, see Heath *et al.* (1992) or Chapter 13 in Musiela and Rutkowski (1997).

where

$$a(t, T) = f(t, t) - \alpha^*(t, T) + \frac{1}{2} |\sigma^*(t, T)|^2, \quad b(t, T) = -\sigma^*(t, T),$$

with $\alpha^*(t, T) = \int_t^T \alpha(t, u) du$ and $\sigma^*(t, T) = \int_t^T \sigma(t, u) du$.

An analogous result holds for $\tilde{D}(t, T)$, with an obvious change of notation. That is,

$$d\tilde{D}(t, T) = \tilde{D}(t, T)(\tilde{a}(t, T) dt + \tilde{b}(t, T) \cdot dW_t) \quad (7.6)$$

with

$$\tilde{a}(t, T) = g(t, t) - \tilde{\alpha}^*(t, T) + \frac{1}{2} |\tilde{\sigma}^*(t, T)|^2, \quad \tilde{b}(t, T) = -\tilde{\sigma}^*(t, T). \quad (7.7)$$

We assume, as customary, that one may also invest in the risk-free savings account B , which corresponds to the short-term rate $r_t = f(t, t)$. In view of (7.5), the relative bond price $Z(t, T) = B_t^{-1} B(t, T)$ satisfies under $\mathbb{P}$

$$dZ(t, T) = Z(t, T) \left(\left(\frac{1}{2} |b(t, T)|^2 - \alpha^*(t, T) \right) dt + b(t, T) \cdot dW_t \right).$$

The following condition is known to exclude arbitrage across default-free bonds for all maturities $T \leq T^*$, as well as between default-free bonds and the savings account.

Condition (M.1) There exists an adapted $\mathbb{R}^d$ -valued process γ such that

$$\mathbb{E}_{\mathbb{P}} \left\{ \exp \left(\int_0^{T^*} \gamma_u \cdot dW_u - \frac{1}{2} \int_0^{T^*} |\gamma_u|^2 du \right) \right\} = 1$$

and, for any maturity $T \leq T^*$, we have

$$\alpha^*(t, T) = \frac{1}{2} |\sigma^*(t, T)|^2 - \sigma^*(t, T) \cdot \gamma_t.$$

Let γ be some process satisfying Condition (M.1). Then the probability measure $\mathbb{P}^*$, given by the formula

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = \exp \left(\int_0^{T^*} \gamma_u \cdot dW_u - \frac{1}{2} \int_0^{T^*} |\gamma_u|^2 du \right), \quad \mathbb{P}\text{-a.s.}, \quad (7.8)$$

is a spot martingale measure for the default-free term structure. Moreover, if we define a Brownian motion W^* under $\mathbb{P}^*$ by setting

$$W_t^* = W_t - \int_0^t \gamma_u du, \quad \forall t \in [0, T^*],$$

then, for any fixed maturity $T \leq T^*$, the discounted price of risk-free bond satisfies under $\mathbb{P}^*$

$$dZ(t, T) = Z(t, T) b(t, T) \cdot dW_t^*. \quad (7.9)$$

We shall assume from now on that the process γ is uniquely determined, so that the default-free bonds market is complete.⁹ Formally, this means that any default-free contingent claim can be priced through risk-neutral valuation formula. It should be stressed, however, that this remark does not apply to defaultable claims. After a recollection of the well-known facts about the Heath–Jarrow–Morton approach, we shall now focus on the dynamics of the relative pre-default value of a defaultable bond. First, under $\mathbb{P}$ the process $\tilde{Z}(t, T) = B_t^{-1} \tilde{D}(t, T)$ satisfies

$$d\tilde{Z}(t, T) = \tilde{Z}(t, T)(\tilde{a}(t, T) - r_t) dt + \tilde{b}(t, T) \cdot dW_t. \quad (7.10)$$

Consequently, under the unique spot martingale measure $\mathbb{P}^*$, we have

$$d\tilde{Z}(t, T) = \tilde{Z}(t, T)(\lambda_t dt + \tilde{b}(t, T) \cdot dW_t^*), \quad (7.11)$$

where we set

$$\lambda_t := \tilde{a}(t, T) - r_t + \tilde{b}(t, T) \cdot \gamma_t, \quad \forall t \in [0, T]. \quad (7.12)$$

Notice that the process λ may depend on maturity T , in general. We shall however assume that λ does not depend on T . This assumption is satisfied, for instance, when $\sigma(t, T) = \tilde{\sigma}(t, T)$ (see footnote ¹⁰ below).

The no-arbitrage condition between a defaultable bond and savings account reads:¹¹ $\lambda_t = 0$ for $t \leq T$. It is easily seen that this condition is never satisfied, under the present assumptions. Indeed, were it true, $\tilde{Z}(t, T)$ would follow a martingale under $\mathbb{P}^*$, and we would have

$$\tilde{D}(t, T) = \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^T r_u du \right) \middle| \mathcal{F}_t \right\} = B(t, T), \quad \forall t \in [0, T].$$

The last formula clearly contradicts our assumption that $\tilde{D}(t, T) < B(t, T)$. Therefore, we shall assume from now on that the process λ does not vanish identically, for any maturity in question. From the property that the credit spread $g(t, u) - f(t, u)$ is strictly positive, it is also possible to deduce that λ follows a strictly positive process.¹⁰ In fact, first let us observe that the process

$$\tilde{Z}(t, T) \exp \left(- \int_t^T \lambda_u du \right)$$

⁹ Strictly speaking, this assumption is not required for our further development.

¹⁰ This is obvious, if we assume, for instance, that $\sigma(t, T) = \tilde{\sigma}(t, T)$, since then $\lambda_t = g(t, t) - r_t$. Schönbucher (1998) derives the equality $\phi_t \lambda_t = g(t, t) - r_t$ for a strictly positive process ϕ , but he works in a somewhat different setup.

¹¹ More precisely, this would have been the no-arbitrage condition between defaultable bond and savings account, if we had have assumed that the process $\tilde{D}(t, T)$ represents the price (as opposed to the pre-default value) of a defaultable bond.

is a $\mathbb{P}^*$ -martingale. Put another way

$$\tilde{D}(t, T) = \mathbb{E}_{\mathbb{P}^*} \left\{ \exp \left(- \int_t^T (r_u + \lambda_u) du \right) \middle| \mathcal{F}_t \right\} \quad (7.13)$$

for every $t \in [0, T]$. Consequently, since we assume that $\tilde{D}(t, T) < B(t, T)$ for all $t \in [0, T]$ and for all maturities $T > 0$, it must hold that for every $s < t$

$$\int_s^t \lambda_u du > 0,$$

thereby implying that $\lambda_t > 0$ for almost all t and almost surely. Let us note that expression (7.13) jointly with the formula (7.20) below agree with the basic valuation formula (4.5) in the case of zero recovery.

Defaultable term structure

Let $\delta \in [0, 1)$ be a fixed, but otherwise arbitrary, number. We introduce an auxiliary process $\lambda_{1,2}$ by setting

$$(\tilde{Z}(t, T) - \delta Z(t, T))\lambda_{1,2}(t) = \tilde{Z}(t, T)\lambda_t, \quad \forall t \in [0, T]. \quad (7.14)$$

Notice that for $\delta = 0$ we simply have $\lambda_{1,2}(t) = \lambda_t$ for every $t \in [0, T]$. On the other hand, if we take $\delta > 0$ then the process $\lambda_{1,2}$ is strictly positive provided that $\tilde{D}(t, T) > \delta B(t, T)$ (recall that we have assumed that $\tilde{D}(t, T) < B(t, T)$).

Remarks If the assumption $\tilde{D}(t, T) > \delta B(t, T)$ is relaxed, the process $\lambda_{1,2}$ is strictly positive provided that

$$\lambda_t(\tilde{Z}(t, T) - \delta Z(t, T)) > 0, \quad \forall t \in [0, T].$$

Notice also that $\lambda_{1,2}$ will depend both on the recovery rate δ and on the maturity date T , in general. In what follows we shall be assuming that the process $\lambda_{1,2}$ is strictly positive.

We shall show that there exists a stopping time τ , such that the process (as before, $H_t = \mathbb{1}_{\{t \geq \tau\}}$)

$$M_t = H_t - \int_0^t \lambda_{1,2}(u) \mathbb{1}_{\{u < \tau\}} du, \quad \forall t \in [0, T], \quad (7.15)$$

follows a local martingale under $\mathbb{P}^*$ (or rather, under a suitable extension $\mathbb{Q}^*$ of $\mathbb{P}^*$, which we are now going to introduce). The existence of τ follows easily from standard results in the theory of stochastic processes, provided that we allow for a suitable enlargement of the underlying probability space. In fact, we cannot expect a stopping time τ with the desired properties to exist on the original

probability space $(\Omega, \mathbb{F}, \mathbb{P}^*)$, in general. For instance, if the underlying filtration is generated by a standard Brownian motion, which is the usual assumption imposed to ensure the uniqueness of the spot martingale measure $\mathbb{P}^*$, no stopping time with desired properties exists on the original space. Let us denote by $(\tilde{\Omega}, \mathbb{G}, \mathbb{Q}^*)$ the enlarged probability space, where $\mathbb{G} = (\mathcal{G}_t)_{t \in [0, T^*]}$. Our additional requirement is that W^* remains a standard Brownian motion when we switch from $\mathbb{P}^*$ to $\mathbb{Q}^*$. To satisfy all these requirements, it suffices to take a product space $(\Omega \times \tilde{\Omega}, (\mathcal{F}_t \otimes \hat{\mathcal{F}})_{t \in [0, T^*]}, \mathbb{P}^* \otimes \hat{\mathbb{P}})$ where the probability space $(\tilde{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ is large enough to support a unit exponential random variable, η say. Then we may put (cf. (6.1))

$$\tau = \inf \left\{ t \in \mathbb{R}_+ : \int_0^t \lambda_{1,2}(u) du \geq \eta \right\}. \quad (7.16)$$

As one might expect, we extend W^* (and all other previously introduced processes) to the enlarged space by setting $W_t^*(\omega, \hat{\omega}) = W_t^*(\omega)$, etc. Subsequently, we introduce the filtration $\mathbb{H} = (\mathcal{H}_t)_{t \in [0, T^*]}$ generated by the random time τ , more precisely, $\mathcal{H}_t = \sigma(H_u : u \leq t)$, where $H_t = \mathbb{1}_{\{\tau \leq t\}}$ is the jump process associated with τ . Finally, we set $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}_t = \sigma(\mathcal{F}_t, \mathcal{H}_t)$ for every t . Then, the desired properties are easily seen to hold under $\mathbb{Q}^* = \mathbb{P}^* \otimes \hat{\mathbb{P}}$. In particular, the process M given by (7.15) is a $\mathbb{G}$ -local martingale under $\mathbb{Q}^*$, and W^* is a $\mathbb{G}$ -Wiener process under $\mathbb{Q}^*$. It is worthwhile to notice that for obvious reasons we cannot require τ to be independent of W^* .

We are in a position to specify the price process of a T -maturity defaultable bond with fractional recovery of par. We first introduce an auxiliary process $\hat{Z}(t, T)$ by postulating that $\hat{Z}(t, T)$ solves the following SDE:

$$\begin{aligned} d\hat{Z}(t, T) &= \hat{Z}(t, T) (\tilde{b}(t, T) \mathbb{1}_{\{t < \tau\}} + b(t, T) \mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + (\delta Z(t, T) - \hat{Z}(t-, T)) dM_t \end{aligned} \quad (7.17)$$

with the initial condition $\hat{Z}(0, T) = \tilde{Z}(0, T)$. For obvious reasons, the process $\hat{Z}(t, T)$, if well defined, follows a local martingale under $\mathbb{Q}^*$. Combining (7.17) with (7.15), we obtain

$$\begin{aligned} d\hat{Z}(t, T) &= \hat{Z}(t, T) (\tilde{b}(t, T) \mathbb{1}_{\{t < \tau\}} + b(t, T) \mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + (\hat{Z}(t, T) - \delta Z(t, T)) \lambda_{1,2}(t) \mathbb{1}_{\{t < \tau\}} dt \\ &\quad + (\delta Z(t, T) - \hat{Z}(t-, T)) dH_t. \end{aligned}$$

On the other hand, inserting (7.11) into (7.14), we find that $\tilde{Z}(t, T)$ solves

$$d\tilde{Z}(t, T) = (\tilde{Z}(t, T) - \delta Z(t, T)) \lambda_{1,2}(t) dt + \tilde{Z}(t, T) \tilde{b}(t, T) \cdot dW_t^*. \quad (7.18)$$

It is thus easily seen that $\hat{Z}(t, T) = \tilde{Z}(t, T)$ on $[0, \tau[$, and thus $\hat{Z}(t, T)$ satisfies also the following SDE:

$$\begin{aligned} d\hat{Z}(t, T) &= \hat{Z}(t, T)(\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + \hat{Z}(t, T)\lambda_t \mathbb{1}_{\{t < \tau\}} dt + (\delta Z(t, T) - \hat{Z}(t-, T)) dH_t. \end{aligned}$$

Next, from (7.9) we obtain (to check (7.19), it is enough to solve the SDE above first on the interval $[0, \tau[$ and subsequently on $[\tau, T]$)

$$\hat{Z}(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{Z}(t, T) + \delta \mathbb{1}_{\{t \geq \tau\}} Z(t, T) \quad (7.19)$$

for any $t \in [0, T]$. In view of the last equality, we may represent the differential of $\hat{Z}(t, T)$ in a still another way, specifically,

$$\begin{aligned} d\hat{Z}(t, T) &= (\tilde{Z}(t, T)\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + \delta Z(t, T)b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + \tilde{Z}(t, T)\lambda_t \mathbb{1}_{\{t < \tau\}} dt + (\delta Z(t, T) - \tilde{Z}(t-, T)) dH_t. \end{aligned}$$

We are in a position to introduce the price process $D^\delta(t, T)$ of a T -maturity defaultable bond. For any $t \in [0, T]$, the process $D^\delta(t, T)$ is defined through the formula

$$D^\delta(t, T) := B_t \hat{Z}(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta \mathbb{1}_{\{t \geq \tau\}} B(t, T), \quad (7.20)$$

where the second equality is an immediate consequence of (7.19).

For $\delta = 0$, the process $\hat{Z}(t, T)$ vanishes on the stochastic interval $[\tau, T]$ and we have simply

$$d\hat{Z}(t, T) = \hat{Z}(t, T)(\lambda_t dt + \tilde{b}(t, T) \cdot dW_t^*) - \hat{Z}(t-, T) dH_t. \quad (7.21)$$

Remarks It is interesting to notice that $\hat{Z}(t, T)$ satisfies also

$$\begin{aligned} d\hat{Z}(t, T) &= (\tilde{Z}(t, T)\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + \delta Z(t, T)b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + (\tilde{Z}(t, T) - \delta Z(t, T))\lambda_{1,2}(t)\mathbb{1}_{\{t < \tau\}} dt \\ &\quad + (\delta Z(t, T) - \tilde{Z}(t, T)) dH_t. \end{aligned}$$

This means that the process $\hat{Z}(t, T)$ can alternatively be introduced through the expression

$$\begin{aligned} d\hat{Z}(t, T) &= (\tilde{Z}(t, T)\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + \delta Z(t, T)b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + (\delta Z(t, T) - \tilde{Z}(t, T)) dM_t \end{aligned} \quad (7.22)$$

with $\hat{Z}(0, T) = \tilde{Z}(0, T)$. We shall use an analogous approach in the next section.

To simplify the exposition, we shall make throughout the following technical assumption, which will also be in force in Section 7.3 (although the process $\hat{Z}(t, T)$ is defined differently in the next section).

Condition (M.D) The process $\hat{Z}(t, T)$, given by the stochastic differential equation (7.17) (or equivalently, by expression (7.22)), follows a $\mathbb{G}$ -martingale (as opposed to a local martingale) under $\mathbb{Q}^*$.

Remarks The necessity of enlarging the underlying probability space is closely related to the fact that it is not possible to replicate a defaultable bond using risk-free bonds. More exactly, the process $D^\delta(t, T)$ does not correspond to the wealth of a self-financing portfolio of risk-free bonds (i.e., it does not represent a redundant security in the risk-free bonds market). On the other hand, a defaultable bond $D^\delta(t, T)$ is redundant on the random set $[0, \tau]$, that is, before the default time. This is a rather weak statement, however, since the stopping time τ is not accessible.

Let us now focus on the migration process $C = (C^1, C^2)$. In the setting of this subsection, C lives on four states, since we have $K = 2$. We may and do assume that $C_0 = (C_0^1, C_0^2) = (1, 1)$. Also, we assume that $C_t^2 = 1$ for every t .¹² Therefore, the only relevant states for the process C are $(1, 1)$ and $(2, 1)$. The state $(1, 1)$ is the *pre-default state*, and the state $(2, 1)$ is the absorbing *default state*. Since the component C^2 is described by the history of C^1 , it is clear that it is enough to specify the dynamics of C^1 . We postulate that the conditional intensity matrix for C^1 is given by the formula

$$\Lambda_t = \begin{pmatrix} -\lambda_{1,2}(t) & \lambda_{1,2}(t) \\ 0 & 0 \end{pmatrix}. \quad (7.23)$$

In the special case of $\delta = 0$ the matrix Λ takes the following simple form

$$\Lambda_t = \begin{pmatrix} -\lambda_t & \lambda_t \\ 0 & 0 \end{pmatrix}. \quad (7.24)$$

The default time τ is given by the formula

$$\tau = \inf\{t \in \mathbb{R}_+ : C_t^1 = 2\} = \inf\{t \in \mathbb{R}_+ : C_t = (2, 1)\}. \quad (7.25)$$

Using (7.20), we obtain for $t \in [0, T]$

$$\begin{aligned} D_{C_t}(t, T) &:= \mathbb{1}_{\{C_t^1=1\}} \tilde{D}(t, T) + \delta \mathbb{1}_{\{C_t^1=2\}} B(t, T) \\ &= \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta \mathbb{1}_{\{t \geq \tau\}} B(t, T) = D^\delta(t, T) \end{aligned}$$

as expected. Notice that the component C^2 plays no essential role in the present setting. This will no longer be true in the case of multiple credit ratings.

¹² The rationale for this convention will appear clear in the multiple credit ratings setup.

Proposition 7.2 Assume that the recovery rate $\delta = 0$. Let $D^0(t, T)$ be given by (7.20), that is, $D^0(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T)$. Then

$$dD^0(t, T) = D^0(t, T) \left((\tilde{a}(t, T) + \tilde{b}(t, T) \cdot \gamma_t) dt + \tilde{b}(t, T) \cdot dW_t^* \right) - D^0(t-, T) dH_t$$

under the martingale measure $\mathbb{Q}^*$. The risk-neutral valuation formula holds under $\mathbb{Q}^*$

$$D^0(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} \mathbb{1}_{\{T < \tau\}} | \mathcal{G}_t). \quad (7.26)$$

Equivalently,

$$D^0(t, T) = B(t, T) \mathbb{E}_{\mathbb{Q}_T}\{T < \tau | \mathcal{G}_t\}, \quad (7.27)$$

where $\mathbb{Q}_T$ is the T -forward measure associated with $\mathbb{Q}^*$, that is,

$$\frac{d\mathbb{Q}_T}{d\mathbb{Q}^*} = \frac{1}{B(0, T)B_T}, \quad \mathbb{Q}^*\text{-a.s.} \quad (7.28)$$

Proof The first statement is an immediate consequence of definition (7.20), combined with (7.10) and (7.19)–(7.21). From (7.11), we get

$$d\tilde{D}(t, T) = \tilde{D}(t, T) \left((r_t + \lambda_t) dt + \tilde{b}(t, T) \cdot dW_t^* \right), \quad (7.29)$$

so that (recall that $\tilde{D}(T, T) = 1$)

$$\tilde{D}(t, T) = \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_T^{-1} | \mathcal{F}_t) = \tilde{B}_t \mathbb{E}_{\mathbb{Q}^*}(\tilde{B}_T^{-1} | \mathcal{G}_t) \quad (7.30)$$

with (cf. (3.12))

$$\tilde{B}_t = \exp \left(\int_0^t (r_u + \lambda_u) du \right). \quad (7.31)$$

This means that $\tilde{D}(t, T)$ corresponds to the process V introduced in Theorem 3.4 (with $Z = 0$ and $X = 1$). Since $\Delta V_\tau = 0$ (this holds since we know that the process $\tilde{D}(t, T)$ is continuous), using Corollary 3.5, we obtain

$$\mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} \mathbb{1}_{\{T < \tau\}} | \mathcal{G}_t).$$

In view of (7.20), this proves (7.26). $\square$

The next result deals with the case of a general recovery rate. Notice that Proposition 7.3 covers also the case of zero recovery, therefore equality (7.26) can be seen as a special case of (7.33).

Proposition 7.3 Assume that $\delta \in [0, 1)$. The price process $D^\delta(t, T)$ of a defaultable bond satisfies

$$D^\delta(t, T) = D_{C_t}(t, T) = \mathbb{1}_{\{C_t^1=1\}} \exp \left(- \int_t^T g(t, u) du \right)$$

$$+ \delta \mathbb{1}_{\{C_t^1=2\}} \exp \left(- \int_t^T f(t, u) du \right). \quad (7.32)$$

Moreover, the risk-neutral valuation formula holds:

$$D_{C_t}(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(\delta B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t). \quad (7.33)$$

Furthermore,

$$D_{C_t}(t, T) = B(t, T) \mathbb{E}_{\mathbb{Q}_T}(\delta \mathbb{1}_{\{T \geq \tau\}} + \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t), \quad (7.34)$$

where $\mathbb{Q}_T$ is the T -forward measure associated with $\mathbb{Q}^*$.

Proof Formula (7.32) is an immediate consequence of (7.3)–(7.4) combined with (7.20) and (7.25). In view of (7.20), it is also clear that $D^\delta(T, T) = \delta \mathbb{1}_{\{T \geq \tau\}} + \mathbb{1}_{\{T < \tau\}}$. It is thus enough to show that the discounted process $B_t^{-1} D^\delta(t, T)$ follows a martingale under $\mathbb{Q}^*$. This is obvious, however, since in view of equality (7.20) we have $B_t^{-1} D^\delta(t, T) = \hat{Z}(t, T)$. In view of (7.33), formula (7.34) is a consequence of the Bayes rule and the definition of the probability measure $\mathbb{Q}_T$. $\square$

Remarks The martingale property $B_t^{-1} D^\delta(t, T)$ can also be verified using the second equality in (7.20). Indeed, we may represent $D^\delta(t, T)$ as follows (recall that $H_t = \mathbb{1}_{\{t \geq \tau\}}$):

$$D^\delta(t, T) = (1 - H_t) \tilde{D}(t, T) + \delta H_t B(t, T).$$

Applying Itô's rule, we obtain

$$\begin{aligned} dD^\delta(t, T) &= (1 - H_t) d\tilde{D}(t, T) - \tilde{D}(t, T) dH_t + \delta H_t dB(t, T) + \delta B(t, T) dH_t \\ &= (1 - H_t) \tilde{D}(t, T) (r_t + \lambda_t) dt + \tilde{b}(t, T) \cdot dW_t^* \\ &\quad - \tilde{D}(t, T) (dM_t + \lambda_{1,2}(t)(1 - H_t) dt) \\ &\quad + \delta H_t B(t, T) (r_t dt + b(t, T) dW_t^*) \\ &\quad + \delta B(t, T) (dM_t + \lambda_{1,2}(t)(1 - H_t) dt) \\ &= (1 - H_t) \tilde{D}(t, T) (r_t + \lambda_t - \lambda_{1,2}(t)) dt \\ &\quad + \delta B(t, T) (r_t H_t + \lambda_{1,2}(t)(1 - H_t)) dt + dN_t, \end{aligned}$$

where N denotes a $\mathbb{Q}^*$ -martingale. Using (7.14), we get

$$dD^\delta(t, T) = r_t ((1 - H_t) \tilde{D}(t, T) + \delta H_t B(t, T)) dt + dN_t = r_t D^\delta(t, T) dt + dN_t,$$

and thus $d(B_t^{-1} D^\delta(t, T)) = B_t^{-1} dN_t$. Finally, one may check directly that $B_t^{-1} dN_t = d\hat{Z}(t, T)$.

Combining (7.30) with (7.20), we obtain

$$D^\delta(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_T^{-1} | \mathcal{F}_t) + \delta \mathbb{1}_{\{t \geq \tau\}} B_t \mathbb{E}_{\mathbb{P}^*}(B_T^{-1} | \mathcal{F}_t). \quad (7.35)$$

In view of (7.33), it is thus tempting to conjecture that

$$I_1(t) := B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} | \mathcal{G}_t) = \mathbb{1}_{\{t \geq \tau\}} B_t \mathbb{E}_{\mathbb{P}^*}(B_T^{-1} | \mathcal{F}_t)$$

and

$$I_2(t) := B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} \mathbb{1}_{\{T < \tau\}} | \mathcal{G}_t) = \mathbb{1}_{\{t < \tau\}} \tilde{B}_t \mathbb{E}_{\mathbb{P}^*}(\tilde{B}_T^{-1} | \mathcal{F}_t).$$

This conjecture is not true, however, as the following proposition shows.

Proposition 7.4 *For any $\delta \in [0, 1)$, we have*

$$I_1(t) = B(t, T) - \mathbb{1}_{\{t < \tau\}} \bar{B}_t \mathbb{E}_{\mathbb{P}^*}(\bar{B}_T^{-1} | \mathcal{F}_t), \quad (7.36)$$

and

$$I_2(t) = \mathbb{1}_{\{t < \tau\}} \bar{B}_t \mathbb{E}_{\mathbb{P}^*}(\bar{B}_T^{-1} | \mathcal{F}_t), \quad (7.37)$$

where

$$\bar{B}_t = \exp \left(\int_0^t (r_u + \lambda_{1,2}(u)) du \right).$$

Furthermore

$$D^\delta(t, T) = \delta B(t, T) + (1 - \delta) \mathbb{1}_{\{t < \tau\}} \bar{B}_t \mathbb{E}_{\mathbb{P}^*}(\bar{B}_T^{-1} | \mathcal{F}_t), \quad (7.38)$$

or equivalently,

$$D^\delta(t, T) = B(t, T) - (1 - \delta) \left(B(t, T) - \mathbb{1}_{\{t < \tau\}} \bar{B}_t \mathbb{E}_{\mathbb{P}^*}(\bar{B}_T^{-1} | \mathcal{F}_t) \right). \quad (7.39)$$

Finally, we have

$$D_{C_t}(t, T) = B(t, T) \left(\delta + (1 - \delta) \mathbb{1}_{\{t < \tau\}} \mathbb{E}_{\mathbb{P}_T} \left(e^{-\int_t^T \lambda_{1,2}(u) du} \mid \mathcal{F}_t \right) \right), \quad (7.40)$$

where $\mathbb{P}_T$ is the T -forward measure associated with $\mathbb{P}^*$.

Proof Let us rewrite $I_1(t)$ as follows:

$$I_1(t) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} H_T | \mathcal{G}_t) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} | \mathcal{G}_t) - B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} (1 - H_T) | \mathcal{G}_t).$$

Reasoning similarly as in Lando (1998) (see also Lemma 13 and Corollary 14 in Wong (1998)) or as in the proof of Proposition 5.1, we obtain

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}^*}(1 - H_T | \mathcal{F}_T \vee \mathcal{H}_t) &= \mathbb{Q}^*\{\tau > T | \mathcal{F}_T \vee \mathcal{H}_t\} = \mathbb{1}_{\{t < \tau\}} e^{-\int_t^T \lambda_{1,2}(u) du} \\ &= (1 - H_t) e^{-\int_t^T \lambda_{1,2}(u) du}, \end{aligned}$$

where $\mathcal{H}_t = \sigma(H_u : u \leq t)$. Combining the formulae above, we obtain

$$\begin{aligned} I_1(t) &= B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} | \mathcal{G}_t) - B_t \mathbb{E}_{\mathbb{Q}^*}\left(B_T^{-1}(1 - H_t) e^{-\int_t^T \lambda_{1,2}(u) du} \mid \mathcal{G}_t\right) \\ &= B_t \mathbb{E}_{\mathbb{P}^*}(B_T^{-1} | \mathcal{F}_t) - (1 - H_t) \bar{B}_t \mathbb{E}_{\mathbb{Q}^*}(\bar{B}_T^{-1} | \mathcal{G}_t) \\ &= B(t, T) - (1 - H_t) \bar{B}_t \mathbb{E}_{\mathbb{P}^*}(\bar{B}_T^{-1} | \mathcal{F}_t). \end{aligned}$$

Since for $I_2(t)$ we have

$$I_2(t) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1}(1 - H_T) | \mathcal{G}_t),$$

using the same arguments as for $I_1(t)$, we arrive at

$$I_2(t) = (1 - H_t) \bar{B}_t \mathbb{E}_{\mathbb{Q}^*}(\bar{B}_T^{-1} | \mathcal{G}_t).$$

Finally, $D^\delta(t, T) = \delta I_1(t) + I_2(t)$, and thus (7.38)–(7.39) are trivial consequences of (7.36)–(7.37). Formula (7.40) follows from (7.38) and the properties of the forward measure $\mathbb{P}_T$. $\square$

Notice that for $\delta = 0$, we have $\bar{B} = \tilde{B}$, and thus formula (7.38) reduces to $D^0(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T)$. On the other hand, for $\delta = 1$, we have, as expected, $D^1(t, T) = B(t, T)$. Finally, when $0 < \delta < 1$, expression (7.38) yields a decomposition of the price $D^\delta(t, T)$ of a defaultable bond into its predicted ‘post-default value’ $\delta B(t, T)$ and the ‘pre-default premium’ $D^\delta(t, T) - \delta B(t, T)$. Similarly, (7.39) represents $D^\delta(t, T)$ as the difference between its ‘potential value’ $B(t, T)$ and the ‘expected loss in value’ due to the credit risk. One might also look at (7.39) from the perspective of the buyer of a defaultable bond: the price $D^\delta(t, T)$ equals to the price of the default-free bond minus a compensation for credit risk.

Remarks Let us denote

$$J(t) = \mathbb{1}_{\{t < \tau\}} \bar{B}_t \mathbb{E}_{\mathbb{Q}^*}(\bar{B}_T^{-1} | \mathcal{G}_t) = B_t \mathbb{E}_{\mathbb{Q}^*}\left(B_T^{-1}(1 - H_t) e^{-\int_t^T \lambda_{1,2}(u) du} \mid \mathcal{G}_t\right).$$

From the proof of Proposition 7.4 we know that

$$(1 - H_t) e^{-\int_t^T \lambda_{1,2}(u) du} = \mathbb{Q}^*\{T < \tau \mid \mathcal{F}_T \vee \mathcal{H}_t\}$$

so that

$$J(t) = B_t \mathbb{E}_{\mathbb{Q}^*}(B_T^{-1} \mathbb{Q}^*\{T < \tau \mid \mathcal{F}_T \vee \mathcal{H}_t\} \mid \mathcal{G}_t).$$

As already mentioned, in the present setup the stopping time τ and the underlying Wiener process W^* (and consequently τ and B) usually are not mutually independent. Assume, on the contrary, that τ and B are mutually independent.¹³ Under

¹³ More precisely, we assume that the default time τ is independent of $\mathcal{F}_T$ and the process B is independent of $\mathcal{H}_t$.

this – rather unplauble – assumption, $J(t)$ would read

$$J(t) = B(t, T)\mathbb{Q}^*\{T < \tau \mid \mathcal{H}_t\}.$$

Consequently, we would be able to rewrite the valuation formula (7.38) on the set $\{t < \tau\} = \{C_t^1 = 1\}$ in the following way:

$$D^\delta(t, T) = \tilde{D}(t, T) = B(t, T)(\delta + (1 - \delta)\mathbb{Q}^*\{T < \tau \mid C_t^1 = 1\}). \quad (7.41)$$

The last formula corresponds to expression (5.7), obtained in a different setup by Jarrow *et al.* (1997). Let us recall that Jarrow *et al.* (1997) explicitly assume that the migrations process is independent of the underlying short-term rate process r . Needless to say that representation (7.38) is more general than (7.41) since it allows for the dependence between the migration process for defaultable bonds and the risk-free term structure.

7.2 Alternative specifications of recovery payment

We have assumed so far that the recovery payment is fixed, and takes place at the maturity T of a defaultable bond. In this section, we shall assume instead that the constant (or random) payment is done at the default time rather than at the bond's maturity date. It appears that our approach can be easily extended to cover this case as well.

In what follows, we shall focus on two important special cases. First, let us observe that the constant payoff δ at time $t < T$ corresponds to the payoff $\delta B^{-1}(t, T)$ at the terminal date T . Similarly, the payoff $\delta \tilde{D}(t, T)$, which corresponds to the *fractional recovery of market value*, can be represented by the payoff $\delta \tilde{D}(t, T)B^{-1}(t, T)$ at bond's maturity. We conclude that to cover typical cases when the recovery payment is done at time of default, it is enough to extend the construction above to the case of an $(\mathcal{F}_t)$ -adapted stochastic process δ_t .

Let δ_t be the given adapted process on the original probability space endowed with the filtration $(\mathcal{F}_t)$. Condition (7.14), which serves as a starting point in the specification of the default time τ now takes the following form:

$$(\tilde{Z}(t, T) - \delta_t Z(t, T))\lambda_{1,2}(t) = \tilde{Z}(t, T)\lambda_t, \quad \forall t \in [0, T]. \quad (7.42)$$

We assume, as before, that the condition above defines a strictly positive adapted process $\lambda_{1,2}(t)$. We shall now show how to modify the basic equations (7.17)–(7.20).

We now introduce an auxiliary process $\hat{Z}(t, T)$ about which we postulate that it solves the SDE

$$d\hat{Z}(t, T) = \hat{Z}(t, T)(\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^*$$

$$+ (\delta_t Z(t, T) - \hat{Z}(t-, T)) dM_t$$

with the initial condition $\hat{Z}(0, T) = \tilde{Z}(0, T)$. Notice that, as before, the process $\hat{Z}(t, T)$ follows a local martingale under $\mathbb{Q}^*$. Reasoning along the same lines as in the previous section, we find that $\hat{Z}(t, T)$ satisfies

$$\begin{aligned} d\hat{Z}(t, T) &= \hat{Z}(t, T)(\tilde{b}(t, T)\mathbb{1}_{\{t < \tau\}} + b(t, T)\mathbb{1}_{\{t \geq \tau\}}) \cdot dW_t^* \\ &\quad + \hat{Z}(t, T)\lambda_t \mathbb{1}_{\{t < \tau\}} dt + (\delta_t Z(t, T) - \hat{Z}(t-, T)) dH_t, \end{aligned}$$

and thus

$$\hat{Z}(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{Z}(t, T) + \delta_\tau \mathbb{1}_{\{t \geq \tau\}} Z(t, T)$$

for any $t \in [0, T]$. The price process $\hat{D}^\delta(t, T)$ of a T -maturity defaultable bond is now given by the following expression:

$$\hat{D}^\delta(t, T) := B_t \hat{Z}(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta_\tau \mathbb{1}_{\{t \geq \tau\}} B(t, T).$$

The payoff δ_τ at time τ corresponds to the random payoff $\delta^* = \delta_\tau B^{-1}(\tau, T)$ at time T . Therefore, arguing similarly as in the proof of Proposition 7.3, we may then show that

$$\hat{D}^\delta(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(\delta^* B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t).$$

Fractional recovery of par value

For $\delta_t = \delta B^{-1}(t, T)$, we obtain

$$\hat{D}^\delta(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta B^{-1}(\tau, T) \mathbb{1}_{\{t \geq \tau\}} B(t, T).$$

This corresponds to the random payoff $\delta^* = \delta B^{-1}(\tau, T)$ at time T . Consequently, we obtain the following expression for the price process of a T -maturity defaultable bond:

$$\hat{D}^\delta(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta^* \mathbb{1}_{\{t \geq \tau\}} B(t, T).$$

Arguing similarly as in the proof of Proposition 7.3, we may then show that

$$\hat{D}^\delta(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(\delta B^{-1}(\tau, T) B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t).$$

Fractional recovery of market value

Let us recall that this case was examined, in a slightly different setup, in Section 4.2. Let us assume that $\delta_t = \delta \tilde{D}(t, T) B^{-1}(t, T)$. Then

$$\hat{D}^\delta(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta \tilde{D}(\tau, T) B^{-1}(\tau, T) \mathbb{1}_{\{t \geq \tau\}} B(t, T).$$

Consequently,

$$\hat{D}^\delta(t, T) = \mathbb{1}_{\{t < \tau\}} \tilde{D}(t, T) + \delta^* \mathbb{1}_{\{t \geq \tau\}} B(t, T),$$

where $\delta^* = \delta \tilde{D}(\tau, T) B^{-1}(\tau, T)$, and thus

$$\hat{D}^\delta(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*}(\delta \tilde{D}(\tau, T) B^{-1}(\tau, T) B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} \mid \mathcal{G}_t).$$

7.3 Multiple credit ratings case

We assume now that the set of rating classes is $\mathcal{K} = \{1, \dots, K\}$, where the class K corresponds to the default event. For any $i = 1, \dots, K$, we write $\delta_i \in [0, 1)$ to denote the corresponding recovery rate. By assumption, δ_i is the fraction of par paid at bond's maturity, if the bond which is currently in the i^{th} rating class defaults. In this section, we will consider a risk-free term structure (see Section 7.1), as well as $K - 1$ different defaultable term structures (notice that the discussion in the previous section regarded the case where $K = 2$). We generalize condition (B.D) by making the following assumption.

(B.3) For any fixed maturity $T \leq T^*$, the instantaneous forward rate $g_i(t, T)$, corresponding to the rating class $i = 1, \dots, K$ satisfies under $\mathbb{P}$

$$dg_i(t, T) = \alpha_i(t, T) dt + \sigma_i(t, T) \cdot dW_t, \quad (7.43)$$

where $\alpha_i(\cdot, T)$ and $\sigma_i(\cdot, T)$ are adapted stochastic processes with values in $\mathbb{R}$ and $\mathbb{R}^d$, respectively. In addition, we assume that

$$g_{K-1}(t, T) > g_{K-2}(t, T) > \dots > g_1(t, T) > f(t, T). \quad (7.44)$$

As before, the price of a T -maturity default-free discount bond is denoted by $B(t, T)$ so that

$$B(t, T) = \exp\left(-\int_t^T f(t, u) du\right) \quad (7.45)$$

and we denote $Z(t, T) = B(t, T)/B_t$. We also set

$$D_i(t, T) := \exp\left(-\int_t^T g_i(t, u) du\right) \quad (7.46)$$

for $i = 1, \dots, K - 1$. Formulae analogous to (7.5)–(7.7) hold for processes $B(t, T)$ and $D_i(t, T)$, $i = 1, \dots, K - 1$, after a suitable change of notation. In particular, we now denote

$$a_i(t, T) = g_i(t, t) - \alpha_i^*(t, T) + \frac{1}{2} |\sigma_i^*(t, T)|^2, \quad b_i(t, T) = -\sigma_i^*(t, T), \quad (7.47)$$

where

$$\alpha_i^*(t, T) = \int_t^T \alpha_i(t, u) du, \quad \sigma_i^*(t, T) = \int_t^T \sigma_i(t, u) du.$$

As before, we assume that condition (M.1) is satisfied, with uniquely defined process γ .

Condition (M.2) For $i = 1, \dots, K - 1$, the process λ_i , which is given by the formula

$$\lambda_i(t) := a_i(t, T) - f(t, t) + b_i(t, T) \cdot \gamma_t, \quad \forall t \in [0, T], \quad (7.48)$$

does not depend on the maturity T .

Remarks If we assume, in addition, that¹⁴

$$a_i(t, T) + b_i(t, T) \cdot \gamma_t = g_i(t, T)$$

then $\lambda_i(t) = g_i(t, t) - f(t, t)$, so that obviously $\lambda_i(t) > 0$ for $i = 1, \dots, K$ (this is a consequence of (7.44)). It is worthwhile to stress, however, that neither the strict positivity of the λ_i nor their independence of maturity T are necessary requirements for our further developments.

From now on, we make standing assumptions (M.1)–(M.2). Proceeding as in Section 7.1, we construct a martingale measure $\mathbb{P}^*$ for the risk-free term structure. In particular, under $\mathbb{P}^*$ the process $Z(t, T) = B_t^{-1}B(t, T)$ satisfies

$$dZ(t, T) = Z(t, T)b(t, T) \cdot dW_t^*. \quad (7.49)$$

Similarly, if we define processes $Z_i(t, T) = B_t^{-1}D_i(t, T)$ for $i = 1, \dots, K - 1$, we obtain the following dynamics for $Z_i(t, T)$ under $\mathbb{P}^*$ (cf. (7.11))

$$dZ_i(t, T) = Z_i(t, T)(\lambda_i(t)dt + b_i(t, T) \cdot dW_t^*). \quad (7.50)$$

The next step is to introduce a conditionally Markov chain C^1 on the state space $\mathcal{K} = \{1, \dots, K\}$. To construct C^1 in a formal way, we shall typically need to enlarge the underlying probability space. Suitable extensions of $\mathcal{F}_t$ and $\mathbb{P}^*$ will be denoted by $\tilde{\mathcal{F}}_t$ and $\mathbb{Q}^*$, respectively, and they can be constructed in a way analogous to the one used in Section 7.1, although a countable number of independent unit exponential random variables will typically be needed for this construction (see Bielecki and Rutkowski (1999)). The infinitesimal generator of C^1 at time t , given the σ -field $\mathcal{F}_t$, is

$$\Lambda_t = \begin{pmatrix} \lambda_{1,1}(t) & \dots & \lambda_{1,K}(t) \\ \vdots & \ddots & \vdots \\ \lambda_{K-1,1}(t) & \dots & \lambda_{K-1,K}(t) \\ 0 & \dots & 0 \end{pmatrix}, \quad (7.51)$$

¹⁴ A sufficient condition for this is that $\sigma_i(t, T) = \sigma(t, T)$.

where $\lambda_{i,i}(t) = -\sum_{j \neq i} \lambda_{i,j}(t)$ for $i = 1, \dots, K-1$, and where $\lambda_{i,j}$ are adapted, strictly positive processes. To provide our pricing model with arbitrage free features, the processes $\lambda_{i,j}$ will be additionally assumed to satisfy the consistency condition (7.59) (or (7.56) if $K = 3$). We shall write $H_i(t) = \mathbb{1}_{\{C_t^1=i\}}$ for $i = 1, \dots, K$. Let us define

$$M_{i,j}(t) := H_{i,j}(t) - \int_0^t \lambda_{i,j}(s) H_i(s) ds, \quad \forall t \in [0, T], \quad (7.52)$$

for $i = 1, \dots, K-1$ and $j \neq i$, where $H_{i,j}(t)$ represents the number of transitions from i to j by C^1 over the time interval $(0, t]$. It can be shown (see Bielecki and Rutkowski (1999)) that $M_{i,j}(t)$ is a local martingale on the enlarged probability space $(\tilde{\Omega}, (\mathcal{G}_t)_{t \in [0, T^*]}, \mathbb{Q}^*)$. We set $C_t^2 = C_{u(t)-}^1$, where $u(t) = \sup\{u \leq t : C_u^1 \neq C_t^1\}$ (by convention, $\sup \emptyset = 0$, therefore $C_t^2 = C_t^1$ if $C_u^1 = C_0^1$ for every $u \in [0, t]$). In words, $u(t)$ is the time of the last jump of C^1 before (and including) time t , so that C_t^2 represents the last state of C^1 before the current state C_t^1 .

Case $K = 3$

For the reader's convenience, we shall first examine the case when $K = 3$. We assume that $(C_0^1, C_0^2) \in \{(1, 1), (2, 2)\}$, so that $H_1(0) + H_2(0) = \mathbb{1}_{\{C_0^1=1\}} + \mathbb{1}_{\{C_0^1=2\}} = 1$. We also observe that for $i, j = 1, 2, i \neq j$, and for all $t \in [0, T]$ we have

$$H_i(t) = H_i(0) + H_{j,i}(t) - H_{i,j}(t) - H_{i,3}(t) \quad (7.53)$$

and

$$H_{i,3}(t) = \mathbb{1}_{\{C_t^1=3, C_t^2=i\}}. \quad (7.54)$$

Next, we define an auxiliary process $\hat{Z}(t, T)$, which also follows a $\mathbb{G}$ -local martingale under $\mathbb{Q}^*$, by setting (the formula below is a straightforward generalization of (7.22))

$$\begin{aligned} d\hat{Z}(t, T) &:= (Z_2(t, T) - Z_1(t, T)) dM_{1,2}(t) + (Z_1(t, T) - Z_2(t, T)) dM_{2,1}(t) \\ &+ (\delta_1 Z(t, T) - Z_1(t, T)) dM_{1,3}(t) + (\delta_2 Z(t, T) - Z_2(t, T)) dM_{2,3}(t) \\ &+ (H_1(t) Z_1(t, T) b_1(t, T) + H_2(t) Z_2(t, T) b_2(t, T)) \cdot dW_t^* \\ &+ (\delta_1 H_{1,3}(t) + \delta_2 H_{2,3}(t)) Z(t, T) b(t, T) \cdot dW_t^* \end{aligned}$$

with the initial condition

$$\hat{Z}(0, T) = H_1(0) Z_1(0, T) + H_2(0) Z_2(0, T). \quad (7.55)$$

Using (7.52), we arrive at the following representation for the dynamics of $\hat{Z}(t, T)$:

$$\begin{aligned} d\hat{Z}(t, T) &= Z_1(t) (dH_{2,1}(t) - dH_{1,2}(t) - dH_{1,3}(t)) + H_1(t) dZ_1(t) \\ &+ Z_2(t) (dH_{1,2}(t) - dH_{2,1}(t) - dH_{2,3}(t)) + H_2(t) dZ_2(t) \end{aligned}$$

$$\begin{aligned}
& + Z(t) \left(\delta_1 dH_{1,3}(t) + \delta_2 dH_{2,3}(t) \right) + \left(\delta_1 H_{1,3}(t) + \delta_2 H_{2,3}(t) \right) dZ(t) \\
& - \left[\lambda_{1,2}(t) (Z_2(t) - Z_1(t)) + \lambda_{1,3}(t) (\delta_1 Z(t) - Z_1(t)) + \lambda_1(t) Z_1(t) \right] H_1(t) dt \\
& - \left[\lambda_{2,1}(t) (Z_1(t) - Z_2(t)) + \lambda_{2,3}(t) (\delta_2 Z(t) - Z_2(t)) + \lambda_2(t) Z_2(t) \right] H_2(t) dt,
\end{aligned}$$

where $Z_i(t) = Z_i(t, T)$ and $Z(t) = Z(t, T)$. To construct a consistent model of the term structure, it is indispensable to specify the matrix Λ in a judicious way. We postulate that the entries of Λ are chosen in such a way that the equalities

$$\begin{cases} \lambda_{1,2}(t) (Z_2(t) - Z_1(t)) + \lambda_{1,3}(t) (\delta_1 Z(t) - Z_1(t)) + \lambda_1(t) Z_1(t) = 0, \\ \lambda_{2,1}(t) (Z_1(t) - Z_2(t)) + \lambda_{2,3}(t) (\delta_2 Z(t) - Z_2(t)) + \lambda_2(t) Z_2(t) = 0 \end{cases} \quad (7.56)$$

are satisfied for all $t \in [0, T]$.

Remarks Suppose first that $\delta_1 = \delta_2 = 0$. In this case, we postulate that the entries of Λ satisfy

$$\begin{cases} \lambda_{1,2}(t) (1 - D_{21}(t)) + \lambda_{1,3}(t) = \lambda_1(t), \\ \lambda_{2,1}(t) (1 - D_{12}(t)) + \lambda_{2,3}(t) = \lambda_2(t), \end{cases}$$

where we set $D_{ij}(t) = Z_i(t, T)/Z_j(t, T) = D_i(t, T)/D_j(t, T)$. Notice that the coefficients $\lambda_{i,j}(t)$ are not uniquely determined. We may take, for instance, $\lambda_{1,2}(t) = \lambda_{2,1}(t) = 0$ (no migration between classes 1 and 2) to obtain $\lambda_{1,3}(t) = \lambda_1(t)$ and $\lambda_{2,3}(t) = \lambda_2(t)$, but other choices are also possible. Notice also that we cannot set $\lambda_{1,3}(t) = \lambda_{2,3}(t) = 0$ (no default possible) since we would then have either $\lambda_{1,2}(t) < 0$ or $\lambda_{2,1}(t) < 0$. Suppose, on the contrary, that $\delta_1 + \delta_2 > 0$. In this case, we have

$$\begin{cases} \lambda_{1,2}(t) (1 - D_{21}(t)) + \lambda_{1,3}(t) (1 - \delta_1 d_{31}(t)) = \lambda_1(t), \\ \lambda_{2,1}(t) (1 - D_{12}(t)) + \lambda_{2,3}(t) (1 - \delta_2 d_{32}(t)) = \lambda_2(t), \end{cases}$$

where $d_{ij}(t) = Z(t, T)/Z_j(t, T) = B(t, T)/D_j(t, T)$.

Let us return to the analysis of the process $\hat{Z}(t, T)$. Under (7.56), $\hat{Z}(t, T)$ satisfies

$$\begin{aligned}
d\hat{Z}(t, T) &:= (Z_2(t, T) - Z_1(t, T)) dH_{1,2}(t) + (Z_1(t, T) - Z_2(t, T)) dH_{2,1}(t) \\
&+ (\delta_1 Z(t, T) - Z_1(t, T)) dH_{1,3}(t) + (\delta_2 Z(t, T) - Z_2(t, T)) dH_{2,3}(t) \\
&+ H_1(t) dZ_1(t, T) + H_2(t) dZ_2(t, T) + (\delta_1 H_{1,3}(t) + \delta_2 H_{2,3}(t)) dZ(t, T)
\end{aligned}$$

with the initial condition (7.55). The above representation of the process $\hat{Z}(t, T)$, combined with (7.53) and (7.54), results in the following important formula:

$$\hat{Z}(t, T) = \mathbb{1}_{\{C_t^1=1\}} Z_1(t, T) + \mathbb{1}_{\{C_t^1=2\}} Z_2(t, T) + (\delta_1 H_{1,3}(t) + \delta_2 H_{2,3}(t)) Z(t, T).$$

Put another way:

$$\hat{Z}(t, T) = \mathbb{1}_{\{C_t^1 \neq 3\}} Z_{C_t^1}(t, T) + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = 3\}} Z(t, T). \quad (7.57)$$

Finally, we introduce the price process of a T -maturity defaultable bond by setting

$$D_{C_t}(t, T) := B_t \hat{Z}(t, T) = \mathbb{1}_{\{C_t^1 \neq 3\}} D_{C_t^1}(t, T) + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = 3\}} B(t, T). \quad (7.58)$$

Remarks Under the present assumptions the process $\hat{Z}(t) := \hat{Z}(t, T)$, given by (7.57), can also be defined as the unique solution of the following SDE (cf. (7.17)):

$$\begin{aligned} d\hat{Z}(t) = & (Z_2(t) - H_1(t)\hat{Z}(t-)) dM_{1,2}(t) + (Z_1(t) - H_2(t)\hat{Z}(t-)) dM_{2,1}(t) \\ & + (\delta_1 Z(t) - H_1(t)\hat{Z}(t-)) dM_{1,3}(t) + (\delta_2 Z(t) - H_2(t)\hat{Z}(t-)) dM_{2,3}(t) \\ & + (H_1(t)\hat{Z}(t)b_1(t, T) + H_2(t)\hat{Z}(t)b_2(t, T) + H_3(t)\hat{Z}(t)b(t, T)) \cdot dW_t^* \end{aligned}$$

with the initial condition (7.55). Indeed, since $H_3(t) = 1 - H_1(t) - H_2(t) = H_{13}(t) + H_{23}(t)$, we may rewrite this SDE as follows:

$$\begin{aligned} d\hat{Z}(t) = & (Z_2(t) - H_1(t)\hat{Z}(t-)) dH_{1,2}(t) + H_1(t)\hat{Z}(t)(\lambda_1(t) dt + b_1(t, T)) \cdot dW_t^* \\ & + (Z_1(t) - H_2(t)\hat{Z}(t-)) dH_{2,1}(t) + H_2(t)\hat{Z}(t)(\lambda_2(t) dt + b_2(t, T)) \cdot dW_t^* \\ & + (\delta_1 Z(t) - H_1(t)\hat{Z}(t-)) dH_{1,3}(t) + (\delta_2 Z(t) - H_2(t)\hat{Z}(t-)) dH_{2,3}(t) \\ & + (H_{1,3}(t) + H_{2,3}(t))\hat{Z}(t)b(t, T) \cdot dW_t^* \\ & - H_1(t)[\lambda_{1,2}(t)(Z_2(t) - \hat{Z}(t)) + \lambda_{1,3}(t)(\delta_1 Z(t) - \hat{Z}(t)) + \lambda_1(t)\hat{Z}(t)] dt \\ & - H_2(t)[\lambda_{2,1}(t)(Z_1(t) - \hat{Z}(t)) + \lambda_{2,3}(t)(\delta_2 Z(t) - \hat{Z}(t)) + \lambda_2(t)\hat{Z}(t)] dt. \end{aligned}$$

In view of (7.49)–(7.50) and (7.56), it is not difficult to check that the unique solution $\hat{Z}(t, T)$ to the SDE above coincides with the process given by the right-hand side of (7.57).

General case

We are in a position to examine the general case. For any $K \geq 3$, we define the process $\hat{Z}(t, T)$ by setting

$$\begin{aligned} d\hat{Z}(t, T) := & \sum_{i,j=1, i \neq j}^{K-1} (Z_j(t, T) - Z_i(t, T)) dM_{i,j}(t) \\ & + \sum_{i=1}^{K-1} (\delta_i Z(t, T) - Z_i(t, T)) dM_{i,K}(t) \\ & + \sum_{i=1}^{K-1} H_i(t) Z_i(t, T) b_i(t, T) \cdot dW_t^* \end{aligned}$$

$$+ \sum_{i=1}^{K-1} \delta_i H_{i,K}(t) Z(t, T) b(t, T) \cdot dW_t^*$$

with the initial condition

$$\hat{Z}(0, T) = \sum_{i=1}^{K-1} H_i(0) Z_i(0, T).$$

We shall now generalize the consistency condition (7.56). We write $Z_i(t) = Z_i(t, T)$.

Condition (M.3) The following equalities are satisfied for each $i = 1, \dots, K-1$, and for every $t \in [0, T]$,

$$\sum_{j=1, j \neq i}^{K-1} \lambda_{i,j}(t) (Z_j(t) - Z_i(t)) + \lambda_{i,K}(t) (\delta_i Z(t) - Z_i(t)) + \lambda_i(t) Z_i(t) = 0. \quad (7.59)$$

Under the assumption above, the process $\hat{Z}(t, T)$ is easily seen to satisfy

$$\begin{aligned} d\hat{Z}(t, T) &= \sum_{i,j=1, i \neq j}^{K-1} (Z_j(t, T) - Z_i(t, T)) dH_{i,j}(t) \\ &+ \sum_{i=1}^{K-1} (\delta_i Z(t, T) - Z_i(t, T)) dH_{i,K}(t) \\ &+ \sum_{i=1}^{K-1} H_i(t) dZ_i(t, T) + \sum_{i=1}^{K-1} \delta_i H_{i,K}(t) dZ(t, T). \end{aligned}$$

The following lemma can be proved along the similar lines as in the case of $K = 3$, therefore its proof is omitted.

Lemma 7.5 Under (7.59), the process $\hat{Z}(t, T)$ satisfies

$$\hat{Z}(t, T) = \sum_{i=1}^{K-1} (H_i(t) Z_i(t, T) + \delta_i H_{i,K}(t) Z(t, T)),$$

or equivalently

$$\hat{Z}(t, T) = \mathbb{1}_{\{C_t^1 \neq K\}} Z_{C_t^1}(t, T) + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = K\}} Z(t, T). \quad (7.60)$$

Moreover, the process $\hat{Z}(t, T)$ is the unique solution to the SDE

$$d\hat{Z}(t, T) = \sum_{i,j=1, i \neq j}^{K-1} (Z_j(t, T) - H_i(t) \hat{Z}(t-, T)) dM_{i,j}(t)$$

$$\begin{aligned}
& + \sum_{i=1}^{K-1} (\delta_i Z(t, T) - H_i(t) \hat{Z}(t-, T)) dM_{i,K}(t) \\
& + \sum_{i=1}^{K-1} H_i(t) \hat{Z}(t, T) b_i(t, T) \cdot dW_t^* + H_K(t) \hat{Z}(t, T) b(t, T) \cdot dW_t^*
\end{aligned}$$

with the initial condition $\hat{Z}(0, T) = \sum_{i=1}^{K-1} H_i(0) Z_i(0, T)$.

As expected, to define the price of a T -maturity defaultable bond we set

$$D_{C_t}(t, T) := B_t \hat{Z}(t, T) = \mathbb{1}_{\{C_t^1 \neq K\}} D_{C_t^1}(t, T) + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = K\}} B(t, T). \quad (7.61)$$

The following result is thus an immediate consequence of the properties of the auxiliary process $\hat{Z}(t, T)$.

Proposition 7.6 *The dynamics of the price process $D_{C_t}(t, T)$ under the risk-neutral probability $\mathbb{Q}^*$ are*

$$\begin{aligned}
dD_{C_t}(t, T) &= \sum_{i,j=1, i \neq j}^{K-1} (D_j(t, T) - D_i(t, T)) dH_{i,j}(t) \\
&+ \sum_{i=1}^{K-1} (\delta_i B(t, T) - D_i(t, T)) dH_{i,K}(t) + \sum_{i=1}^{K-1} H_i(t) dD_i(t, T) \\
&+ \sum_{i=1}^{K-1} \delta_i H_{i,K}(t) dB(t, T) + r_t D_{C_t}(t, T) dt,
\end{aligned}$$

where the differentials $dB(t, T)$ and $dD_i(t, T)$ are given by the formulae

$$dB(t, T) = B(t, T)(r_t dt + b(t, T) \cdot dW_t^*)$$

and

$$dD_i(t, T) = D_i(t, T)((r_t + \lambda_i(t)) dt + b_i(t, T) \cdot dW_t^*).$$

The next proposition shows that the process $D_{C_t}(t, T)$, formally introduced through (7.61), can be given an intuitive interpretation in terms of default time and recovery rate. To this end, we make the following technical assumption (cf. condition (M.D) of Section 7.1).

Condition (M.4) The process $\hat{Z}(t, T)$, given by formula (7.60), follows a $\mathbb{G}$ -martingale (as opposed to a local martingale) under $\mathbb{Q}^*$.

The main result of this section holds under assumptions (B.1)–(B.3) and (M.1)–(M.4).

Theorem 7.7 For any $i = 1, \dots, K - 1$, let $\delta_i \in [0, 1)$ be the recovery rate for a defaultable bond which belongs to the i^{th} rating class at time of default. The price process $D_{C_i}(t, T)$ of a T -maturity defaultable bond equals, for any $t \in [0, T]$,

$$D_{C_i}(t, T) = \mathbb{1}_{\{C_t^1 \neq K\}} e^{-\int_t^T g_{C_t^1}(t, u) du} + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = K\}} e^{-\int_t^T f(t, u) du}, \quad (7.62)$$

or equivalently,

$$D_{C_i}(t, T) = B(t, T) \left(\mathbb{1}_{\{C_t^1 \neq K\}} e^{-\int_t^T \gamma_{C_t^1}(t, u) du} + \delta_{C_t^2} \mathbb{1}_{\{C_t^1 = K\}} \right), \quad (7.63)$$

where $\gamma_i(t, u) = g_i(t, u) - f(t, u)$ is the i^{th} credit spread. Moreover, $D_{C_i}(t, T)$ satisfies the following version of the risk-neutral valuation formula:

$$D_{C_i}(t, T) = B_t \mathbb{E}_{\mathbb{Q}^*} (\delta_{C_T^2} B_T^{-1} \mathbb{1}_{\{T \geq \tau\}} + B_T^{-1} \mathbb{1}_{\{T < \tau\}} | \mathcal{G}_t), \quad (7.64)$$

where τ is the default time, i.e., $\tau = \inf\{t \in \mathbb{R}_+ : C_t^1 = K\}$. The last formula can also be rewritten as follows:

$$D_{C_i}(t, T) = B(t, T) \mathbb{E}_{\mathbb{Q}_T} (\delta_{C_T^2} \mathbb{1}_{\{T \geq \tau\}} + \mathbb{1}_{\{T < \tau\}} | \mathcal{G}_t), \quad (7.65)$$

where $\mathbb{Q}_T$ is the T -forward measure associated with $\mathbb{Q}^*$ through (7.28).

Proof The first formula is an immediate consequence of (7.61) combined with (7.45)–(7.46). For the second, notice first that in view of the second equality in (7.61) and the definition of τ , the process $D_{C_i}(t, T)$ satisfies the terminal condition

$$D_{C_i}(T, T) = \delta_{C_T^2} \mathbb{1}_{\{T \geq \tau\}} + \mathbb{1}_{\{T < \tau\}}.$$

Furthermore, using the first equality in (7.61), we deduce the discounted process $B_t^{-1} D_{C_i}(t, T)$ equals $\hat{Z}(t, T)$, so that it follows a $\mathbb{Q}^*$ -martingale. Equality (7.64) is thus obvious. $\square$

Defaultable coupon bonds

Consider a default-risky coupon bond with the face value F that matures at time T and promises to pay coupons c_i at times T_i ($T_i < T$), $i = 1, 2, \dots, n$. The coupon payments are only made prior to default. For simplicity we also assume that the recovery payment is made at maturity T , in case the bond defaults before or at the maturity. Arbitrage valuation of such a bond is a straightforward consequence of the results obtained earlier in this section. As we have noted before, the intensity matrix of the migration process C_t may depend on both the maturity T and the recovery rates δ_i , $i \in \mathcal{I} := \{1, 2, \dots, K - 1\}$. We shall emphasize this (possible) dependence by writing $C_t(T, \delta_{\mathcal{I}})$. In case of zero recovery we shall write $C_t(T, 0)$. Similarly, we find it convenient to emphasize the dependence of the defaultable

bond's value on the recovery rates by writing $D_{C_t(T, \delta_T)}^{\delta_T}(t, T)$ (or $D_{C_t(T, 0)}^0(t, T)$, in case of zero recovery).

We postulate that the arbitrage price $B_c(t, T)$ of the coupon bond considered here is given by

$$B_c(t, T) = \sum_{i=1}^n c_i D_{C_t(T_i, 0)}^0(t, T_i) + F D_{C_t(T, \delta_T)}^{\delta_T}(t, T), \quad (7.66)$$

with the usual convention that $D_{C_t(T_i, 0)}^0(t, T_i) = 0$ for $t > T_i$. Notice the defaultable bond covenants described above do not necessarily hold (unless a certain monotonicity of default times is imposed). Also, each zero coupon component of a defaultable coupon bond has its own ratings process.

7.4 Market prices of interest rate and credit risk

Let us fix a horizon date T^* . We shall now change, using a suitable generalization of Girsanov's theorem, the measure $\mathbb{Q}^*$ to the equivalent probability measure $\mathbb{Q}$. In financial interpretation, the probability measure $\mathbb{Q}$ plays the role of the real-world probability in our model. For this reason, we postulate that the restriction of $\mathbb{Q}$ to the original probability space Ω necessarily coincides with the underlying probability $\mathbb{P}$. To this end, we set

$$\frac{d\mathbb{Q}}{d\mathbb{Q}^*} \Big|_{\mathcal{G}_t} = L_t, \quad \mathbb{Q}^*\text{-a.s.},$$

where the $\mathbb{Q}^*$ -local positive martingale L is given by the formula (cf. (7.8))

$$dL_t = -L_t \gamma_t \cdot dW_t^* + L_{t-} dM_t, \quad L_0 = 1,$$

where in turn the $\mathbb{Q}^*$ -local martingale M equals

$$dM_t = \sum_{i \neq j} (\phi_{i,j}(t) - 1) dM_{i,j}(t) = \sum_{i \neq j} (\phi_{i,j}(t) - 1) (dH_{i,j}(t) - \lambda_{i,j}(t) H_i(t) dt),$$

and, for any $i \neq j$, we denote by $\phi_{i,j}$ an arbitrary non-negative $\mathbb{F}$ -predictable process such that

$$\int_0^{T^*} \phi_{i,j}(t) \lambda_{i,j}(t) dt < \infty, \quad \mathbb{Q}^*\text{-a.s.}$$

We assume that $\mathbb{E}_{\mathbb{Q}^*}(L_{T^*}) = 1$, so that the probability measure $\mathbb{Q}$ is well defined on $(\tilde{\Omega}, \mathcal{G}_{T^*})$. It can be verified that under the probability measure $\mathbb{Q}$ the migration process C^1 is still a conditionally Markov process, and it has under $\mathbb{Q}$ the infinitesimal generator $\bar{\Lambda}_t$ with the entries $\bar{\lambda}_{i,j}(t) = \phi_{i,j}(t) \lambda_{i,j}(t)$ for every $i \neq j$ and every $t \in [0, T^*]$ (see Bielecki and Rutkowski (1999)). The process γ (the processes

$\phi_{i,j}$, resp.) is referred to as the market price of interest rate risk (market prices of credit risk, resp.)

Remarks In particular, if the market price for credit risk depends only on the current rating i (and not on the rating j after jump) so that $\phi_{i,j} = \phi_{i,i} =: \phi_i$ for every j , the relationship between the intensity matrices under $\mathbb{Q}$ and $\mathbb{Q}^*$ is the following: $\tilde{\Lambda}_t = \Phi \Lambda_t$, where $\Phi = \text{diag}[\phi_i]$ is the diagonal matrix. Such a relationship has been postulated, for instance, in Jarrow *et al.* (1997).

7.5 Model parameters

For several reasons, the parameter specification is the most difficult task in any attempt to measure and to value the credit risk. First, a credit risk model usually involves a relatively large number of parameters, when compared with any standard model of market risk. Second, frequently the volume of available empirical data related to credit-sensitive assets is insufficient for statistical studies (the scarcity of data makes problematic even the possibility of reliable estimation of the credit-spread curve). Before discussing the question of specifying model parameters, let us emphasize that the notion of a credit rating should not be understood literally, but rather in a wider sense. Indeed, by a credit rating we mean here any ‘reasonable’ grouping of credit-sensitive assets, as opposed to ‘official’ credit ratings provided by any of the widely accepted ratings agencies.

Default probabilities

The notion of a credit event involves a number of various situations related to the credit quality of the reference asset. It is thus worthwhile to mention that, in most empirical studies undertaken before 1990, by a *default probability* researchers have meant a probability of defaulting on either interest or principal payment. In more recent studies, it is common to adopt a less stringent definition of default, which can be more adequately referred to as *credit distress*. In this context, let us observe that though the different debts of the same firm encounter credit distress at the same time, it may well happen that senior debt obligations are satisfied in full during bankruptcy procedures, while subordinated debt is paid of only partially. This feature is accounted for in the specification of differing recovery rates to different debts of the same firm, according to the debt seniority. Let us stress that observed default frequencies correspond to the actual probabilities of default, as opposed to the risk-neutral probabilities which are used to value derivative securities. In an arbitrage-free setup, the risk-neutral default probabilities should be seen as by-products obtained within the model, rather than the model inputs.

Recovery rates

It is commonly known that, in the case of default, the likely residual value net of recoveries heavily depends on the seniority class of the debt. To accommodate for this feature, we may assume that the value of a recovery rate reflects not only on the bond credit quality, but also on the seniority classification of the bond (from senior secured to junior unsecured). It is debatable whether it should be represented as a constant or as a random variable. For simplicity, a random recovery rate can be assumed to be independent of other random quantities involved in a model's construction.

Credit spreads

The knowledge of credit spreads represents a salient ingredient of the approach presented in Section 7. To be more specific, we need to examine beforehand not only the *credit-spread curves*, but also *credit-spread volatilities*, and, if several distinct assets are modelled simultaneously, the *credit-spread correlations*. Due to the relative scarcity of data, the estimation of the credit-spread curve is more problematic than the estimation of the risk-free yield curve. This is especially difficult to overcome when one deals with the debt issued by a particular firm. In such a case, one might use the rating-specific credit-spread curve as a proxy for the unobservable firm-specific credit-spread curve (see Fridson and Jónsson (1995)).

On the positive side, there is a good chance that the difficulty in collecting sufficient empirical data will lessen in the future, with the further development of the sector of credit derivatives. The same remarks apply to the estimation of credit-spread volatilities, which in principle can be statistically inferred from the observed variations of the credit-spread yield curve (see, e.g., Fons (1987, 1994) or Foss (1995)). An alternative, and perhaps more promising, approach would be to focus instead on volatilities implicit in market prices of the most actively traded option-like credit derivatives.

Let us finally mention that the valuation of complex credit derivatives requires us also to take into account correlations between the behaviour of several credit-sensitive assets (cf. Zhou (1997a, 1997b) or Duffie and Singleton (1998b)).

In view of the discussion above, it is apparent that our model relies on the strong belief that credit risk inherent in credit-sensitive securities is fully explained by the credit-spread curve and its volatility. Such an approach parallels the common belief that the market risk of interest-rate securities is entirely determined through the behaviour of the default-free yield curve and its volatility. This statement should not be misunderstood; it does not mean that several relevant quantities which are typically present in credit-risk considerations should be totally neglected in our setup. On the contrary, all other quantities commonly used in most econometric models of credit risk (that is: default probabilities, migration matrix, recovery rates,

as well as correlations) are also used. Since econometric models of credit risk are not discussed here, we refer the interested reader to Altman and Bencivenga (1995), Altman and Kishore (1996), Duffie and Singleton (1997), Monkkonen (1997), Wilson (1997), Duffee (1998) or Kiesel *et al.* (1999a, 1999b).

7.6 Valuation of credit derivatives

We shall only discuss here valuation issues for the two most common credit derivatives: a basic default swap and a total rate of return swap.

Default swaps

Consider first a *basic default swap*, as described, for instance, in Duffie (1999). The contingent payment X is triggered by the default event $\{C_t^1 = K\}$. It is settled at time τ , and equals

$$X = (1 - \delta_{C_T^2} B(\tau, T)) \mathbb{1}_{\{\tau \leq T\}}.$$

Notice the dependence of the payment X on the initial rating C_0^1 through default time τ and recovery rate $\delta_{C_T^2}$. We consider two cases. Either (i) the buyer pays a lump sum at the contract's inception (such a contract is referred to as the *default option*), or (ii) the buyer pays an annuity at the fixed time instants t_i , $i = 1, 2, \dots, m$ (*default swap*). In case (i), the value at time 0 of a default option is given by the risk-neutral valuation formula

$$\pi_0(X) = \mathbb{E}_{\mathbb{Q}^*} \left(B_\tau^{-1} (1 - \delta_{C_T^2} B(\tau, T)) \mathbb{1}_{\{\tau \leq T\}} \right).$$

In case (ii), the annuity κ satisfies

$$\pi_0(X) = \kappa \mathbb{E}_{\mathbb{Q}^*} \left(\sum_{i=1}^m B_{t_i}^{-1} \mathbb{1}_{\{t_i < \tau\}} \right).$$

Both the price $\pi_0(X)$ and the annuity κ depend on the initial rating C_0^1 of the underlying bond.

Total rate of return swaps

Next consider a *total rate of return swap* as described, for instance, in Das (1998a). We take as a *reference asset* the coupon bond described with the promised cash flows c_i at times T_i . We assume that its price process is described by equality (7.66). We assume that the contract maturity is $\tilde{T} \leq T$, where T is the maturity date of the underlying coupon-bond. In addition, suppose that the *reference rate* payments (the annuity payments) are made by the investor at fixed scheduled times $t_i \leq \tilde{T}$, $i = 1, 2, \dots, m$. As explained in Section 2.1, the owner of a total rate of return swap is entitled not only to all coupon payments during the life of the

contract, but also to the change in the value of the underlying bond paid as a lump sum at the contract's termination. Then, the reference rate ρ to be paid by the investor should be computed from

$$\begin{aligned} \rho \mathbb{E}_{\mathbb{Q}^*} \left(\sum_{i=1}^m B_{t_i}^{-1} \mathbb{1}_{\{C_i^1(T, \delta_{\mathcal{I}}) \neq K\}} \right) &= \sum_{i=1}^n c_i D_{C_0(T_i, 0)}^0(0, T_i) \mathbb{1}_{\{T_i \leq \tilde{T}\}} \\ &+ \mathbb{E}_{\mathbb{Q}^*} \left(B_{\tilde{\tau}}^{-1} (B_c(\tilde{\tau}, T) - B_c(0, T)) \right), \end{aligned}$$

where $\tilde{\tau} = \tau \wedge \tilde{T}$, and $\tau = \inf \{t \geq 0 : C_t^1(T, \delta_{\mathcal{I}}) = K\}$. For simplicity, in the left-hand side of the valuation formula above, as well as in the second term in the right-hand side, the default time of the underlying coupon bond was assumed to be represented by the default time of its face value component.

In view of the incompleteness of the model, the important issue of hedging strategies for credit derivatives should be dealt with caution; typically, only an approximate hedge is possible (see Arvanitis and Laurent (1999) and Lotz (1998, 1999) in this regard).

References

- Altman, E.I. and Bencivenga, J.C. (1995), A yield premium model for the high-yield debt market, *Financial Analysts Journal* **51**(5), 49–56.
- Altman, E.I. and Kishore, V.M. (1996), Almost everything you wanted to know about recoveries on defaulted bonds, *Financial Analysts Journal* **52**(6), 57–64.
- Ammann, M. (1999) *Pricing Derivative Credit Risk*. Lecture Notes in Economics and Mathematical Systems 470, Springer-Verlag, Berlin.
- Anderson, R. and Sundaresan, S. (2000), A comparative study of structural models of corporate bond yields: an exploratory investigation, *Journal of Banking and Finance* **24**, 255–69.
- Antonelli, F. (1993), Backward–forward stochastic differential equations, *Annals of Applied Probability* **3**, 777–93.
- Artzner, P. and Delbaen, F. (1995), Default risk insurance and incomplete markets, *Mathematical Finance* **5**, 187–95.
- Arvanitis, A. and Laurent, J.-P. (1999), On the edge of completeness, *Risk*, **12**(10).
- Arvanitis, A., Gregory, J. and Laurent, J.-P. (1999), Building models for credit spreads, *Journal of Derivatives* **6**(3), 27–43.
- BeSaw, J. (1997), Pricing credit derivatives, *Derivatives Week*, September 8, 6–7.
- Bielecki, T.R. and Rutkowski, M. (1999), Modelling of the defaultable term structure: conditionally Markov approach, working paper, Northeastern Illinois University and Warsaw University of Technology.
- Bielecki, T.R. and Rutkowski, M. (2000), Multiple ratings model of defaultable term structure, *Mathematical Finance* **10**, 125–39.
- Black, F. and Cox, J.C. (1976), Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–67.
- Brémaud, P. (1981) *Point Processes and Queues. Martingale Dynamics*, Springer-Verlag, Berlin.

- Brennan, M. and Schwartz, E. (1977), Convertible bonds: valuation and optimal strategies for call and conversion, *Journal of Finance* **32**, 1699–715.
- Brennan, M. and Schwartz, E. (1980), Analyzing convertible bonds, *Journal of Financial and Quantitative Analysis* **15**, 907–29.
- Briys, E. and de Varenne, F. (1997), Valuing risky fixed rate debt: an extension, *Journal of CreditMetrics: Technical Document*, J.P. Morgan, New York, 1997.
- CreditRisk⁺: Technical Document*, Credit Suisse Financial Products, 1997.
- Crouhy, M., Galai, D. and Mark, R. (1998), Credit risk revisited, *Risk – Credit Risk Supplement*, March, 40–4.
- Crouhy, M., Galai, D. and Mark, R. (2000), A comparative analysis of current credit risk models, *Journal of Banking and Finance* **24**, 59–117.
- Das, S. (1998a), Credit derivatives – instruments, in: *Credit Derivatives: Trading and Management of Credit and Default Risk*, S. Das, ed., J. Wiley, Singapore, pp. 7–77.
- Das, S. (1998b), Valuation and pricing of credit derivatives, in: *Credit Derivatives: Trading and Management of Credit and Default Risk*, S. Das, ed., J. Wiley, Singapore, pp. 173–231.
- Dellacherie, C. and Meyer, P.A. (1975) *Probabilités et potentiel*, Hermann, Paris.
- Duffee, G. (1998), The relation between Treasury yields and corporate bond yield spreads, forthcoming in *Journal of Finance*.
- Duffie, D. (1998a), First-to-default valuation, working paper, Stanford University.
- Duffie, D. (1998b), Defaultable term structure models with fractional recovery of par, working paper, Stanford University.
- Duffie, D. (1999), Credit swap valuation, *Financial Analysts Journal* **55**(1), 73–87.
- Duffie, D. and Lando, D. (1998), The term structure of credit spreads with incomplete accounting data, working paper, Stanford University and University of Copenhagen.
- Duffie, D. and Singleton, K. (1997), An econometric model of the term structure of interest rate swap yields, *Journal of Finance* **52**, 1287–321.
- Duffie, D. and Singleton, K. (1998a), Ratings-based term structures of credit spreads, working paper, Stanford University.
- Duffie, D. and Singleton, K. (1998b), Simulating correlated defaults, working paper, Stanford University.
- Duffie, D. and Singleton, K. (1999), Modelling term structures of defaultable bonds, *Review of Financial Studies* **12**, 687–720.
- Duffie, D., Schroder, M. and Skiadas, C. (1996), Recursive valuation of defaultable securities and the timing of resolution of uncertainty, *Annals of Applied Probability* **6**, 1075–90.
- El Karoui, N. and Quenez, M.C. (1997a), Nonlinear pricing theory and backward stochastic differential equations, in: *Financial Mathematics, Bressanone, 1996*, W. Runggaldier, ed. *Lecture Notes in Math.* 1656, Springer-Verlag, Berlin, pp. 191–246.
- El Karoui, N. and Quenez, M.C. (1997b), Imperfect markets and backward stochastic differential equations, in: *Numerical Methods in Finance*, L.C.G. Rogers, D. Talay, eds. Cambridge University Press, Cambridge, pp. 181–214.
- El Karoui, N., Peng, S. and Quenez, M.C. (1997), Backward stochastic differential equations in finance, *Mathematical Finance* **7**, 1–72.
- Elliott, R.J., Jeanblanc, M. and Yor, M. (2000), On models of default risk, *Mathematical Finance* **10**, 179–95.
- Fons, J.S. (1987), The default premium and corporate bond experience, *Journal of*

Finance **42**, 81–97.

- Fons, J.S. (1994), Using default rates to model the term structure of credit risk, *Financial Analysts Journal* **50**(5), 25–32.
- Foss, G.W. (1995), Quantifying risk in the corporate bond markets, *Financial Analysts Journal* **51**(2), 29–34.
- Fridson, M.S. and Jónsson, J.G. (1995), Spread versus Treasuries and the riskiness of high-yield bonds, *Journal of Fixed Income* **5**(3), 79–88.
- Geske, R. (1977), The valuation of corporate liabilities as compound options, *Journal of Financial and Quantitative Analysis* **12**, 541–52.
- Geske, R. (1979), The valuation of compound options, *Journal of Financial Economics* **7**, 63–81.
- Heath, D., Jarrow, R. and Morton, A. (1992), Bond pricing and the term structure of interest rates: a new methodology for contingent claim valuation, *Econometrica* **60**, 77–105.
- Huge, B. and Lando, D. (1998), Swap pricing with two-sided default risk in a rating-based model, working paper, University of Copenhagen.
- Hull, J.C. and White, A. (1995), The impact of default risk on the prices of options and other derivative securities, *Journal of Banking and Finance* **19**, 299–322.
- Jarrow, R.A. and Turnbull, S.M. (1995), Pricing derivatives on financial securities subject to credit risk, *Journal of Finance* **50**, 53–85.
- Jarrow, R.A. and Turnbull, S.M. (2000), The intersection of market and credit risk, *Journal of Banking and Finance* **24**, 271–99.
- Jarrow, R.A., Lando, D. and Turnbull, S.M. (1997), A Markov model for the term structure of credit risk spreads, *Review of Financial Studies* **10**, 481–523.
- Jeanblanc, M. and Rutkowski, M. (2000a), Modelling of default risk: an overview, in: *Mathematical Finance: Theory and Practice*, Higher Education Press, Beijing, pp. 171–269.
- Jeanblanc, M. and Rutkowski, M. (2000b), Modelling of default risk: mathematical tools, working paper, Université d'Evry and Warsaw University of Technology.
- Kiesel, R., Perraudin, W. and Taylor, A. (1999a), Credit and interest rate risk, working paper, Birbeck College.
- Kiesel, R., Perraudin, W. and Taylor, A. (1999b), The structure of credit risk, working paper, Birbeck College.
- Kijima, M. (1998), Monotonicity in a Markov chain model for valuing coupon bond subject to credit risk, *Mathematical Finance* **8**, 229–47.
- Kim, I.J., Ramaswamy, K. and Sundaresan, S. (1993), Does default risk in coupons affect the valuation of corporate bonds? *Financial Management* **22**, 117–31.
- Kusuoka, S. (1999), A remark on default risk models, *Advances in Mathematical Economics* **1**, 69–82.
- Lando, D. (1997), Modelling bonds and derivatives with credit risk, in: *Mathematics of Derivative Securities*, M. Dempster, S. Pliska, eds., Cambridge University Press, Cambridge, pp. 369–93.
- Lando, D. (1998), On Cox processes and credit-risky securities, *Review of Derivatives Research* **2**, 99–120.
- Leland, H.E. (1994), Corporate debt value, bond covenants, and optimal capital structure, *Journal of Finance* **49**, 1213–52.
- Leland, H.E. and Toft, K. (1996), Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads, *Journal of Finance* **51**, 987–1019.
- Litterman, R. and Iben, T. (1991), Corporate bond valuation and the term structure of credit spreads, *Journal of Portfolio Management* **17**(3), 52–64.

- Longstaff, F.A. and Schwartz, E.S. (1995), A simple approach to valuing risky fixed and floating rate debt, *Journal of Finance* **50**, 789–819.
- Lotz, C. (1998), Locally risk minimizing the credit risk, working paper, London School of Economics.
- Lotz, C. (1999), Optimal shortfall hedging of credit risk, working paper, University of Bonn.
- Lotz, C. and Schlögl, L. (2000), Default risk in a market model, *Journal of Banking and Finance* **24**, 301–27.
- Madan, D.B. and Unal, H. (1998a), Pricing the risk of default, *Review of Derivatives Research* **2**, 121–60.
- Madan, D.B. and Unal, H. (1998b), A two-factor hazard-rate model for pricing risky debt and the term structure of credit spreads, working paper, University of Maryland.
- Mella-Barral, P. and Tychon, P. (1996), Default risk in asset pricing, working paper, London School of Economics and Université Catholique de Louvain.
- Merton, R.C. (1974), On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–70.
- Monkkonen, H. (1997), Modelling default risk: theory and empirical evidence, Ph.D. thesis, Queen's University.
- Musiela, M. and Rutkowski, M. (1997) *Martingale Methods in Financial Modelling*, Springer-Verlag, Berlin.
- Nielsen, T.N., Saá-Requejo, J. and Santa-Clara, P. (1993), Default risk and interest rate risk: the term structure of default spreads, working paper, INSEAD.
- Pitts, C. and Selby, M. (1983), The pricing of corporate debt: a further note, *Journal of Finance* **38**, 1311–13.
- Rendleman, R.J. (1992), How risks are shared in interest rate swaps?, *Journal of Financial Services Research* 5–34.
- Rutkowski, M. (1999), On models of default risk: by R. Elliott, M. Jeanblanc and M. Yor, working paper, Warsaw University of Technology.
- Schönbucher, P.J. (1998), Term structure modelling of defaultable bonds, *Review of Derivatives Research* **2**, 161–92.
- Schönbucher, P.J. (2000), Credit risk modelling and credit derivatives, Ph.D. dissertation, University of Bonn.
- Tavakoli, J.M. (1998) *Credit Derivatives: A Guide to Instruments and Applications*, J. Wiley, New York.
- Thomas, L.C., Allen, D.E. and Morkel-Kingsbury, N. (1998), A hidden Markov chain model for the term structure of bond credit risk spreads, working paper, Edith Cowan University.
- Wilson, T. (1997), Portfolio credit risk, *Risk* **10(9,10)**, 111–17, 56–61.
- Wong, D. (1998), A unifying credit model, working paper, Scotia Capital Markets.
- Zhou, C. (1997a), A jump diffusion approach to modelling credit risk and valuing defaultable securities, working paper, Federal Reserve Board.
- Zhou, C. (1997b), Default correlation: an analytical result, working paper, Federal Reserve Board.

Towards a Theory of Volatility Trading*

Peter Carr and Dilip Madan

1 Introduction

Much research has been directed towards forecasting the volatility¹ of various macroeconomic variables such as stock indices, interest rates and exchange rates. However, comparatively little research has been directed towards the optimal way to invest given a view on volatility. This absence is probably due to the belief that volatility is difficult to trade. For this reason, a small literature has emerged which advocates the development of volatility indices and the listing of financial products whose payoff is tied to these indices. For example, Gastineau (1977) and Galai (1979) propose the development of option indices similar in concept to stock indices. Brenner and Galai (1989) propose the development of realized volatility indices and the development of futures and options contracts on these indices. Similarly, Fleming, Ostdiek and Whaley (1993) describe the construction of an implied volatility index (the VIX), while Whaley (1993) proposes derivative contracts written on this index. Brenner and Galai (1993, 1996) develop a valuation model for options on volatility using a binomial process, while Grunbichler and Longstaff (1993) instead assume a mean reverting process in continuous time.

In response to this hue and cry, some volatility contracts have been listed. For example, the OMLX, which is the London based subsidiary of the Swedish exchange OM, launched volatility futures at the beginning of 1997. At the time of this writing, the Deutsche Terminbörse (DTB) recently launched its own futures based on its already established implied volatility index. Thus far, the volume in these contracts has been disappointing.

One possible explanation for this outcome is that volatility can already be traded by combining static positions in options on price with dynamic trading in the underlying. Neuberger (1990) showed that by delta-hedging a contract paying the log

* Originally published as Chapter 29 of *Volatility: New Estimation Techniques for Pricing Derivatives*, R. Jarrow (ed.), Risk Books, 1998. Reprinted with permission of Risk Books.

¹ In this chapter, the term “volatility” refers to either the variance or the standard deviation of the return on an investment.

of the price, the hedging error accumulates to the difference between the realized variance and the fixed variance used in the delta-hedge. The contract paying the log of the price can be created with a static position in options, as shown in Breeden and Litzenberger (1978). Independently of Neuberger, Dupire (1993) showed that a calendar spread of two such log contracts pays the variance between the two maturities, and developed the notion of forward variance. Following Heath, Jarrow, and Morton (1992) (HJM), Dupire modeled the evolution of the term structure of this forward variance, thereby developing the first stochastic volatility model in which the market price of volatility risk does not require specification, even though volatility is imperfectly correlated with the price of the underlying.

The primary purpose of this chapter is to review three methods which have emerged for trading realized volatility. The first method reviewed involves taking static positions in options. The classic example is that of a long position in a straddle, since the value usually² increases with a rise in volatility. The second method reviewed involves delta-hedging an option position. If the investor is successful in hedging away the price risk, then a prime determinant of the profit or loss from this strategy is the difference between the realized volatility and the anticipated volatility used in pricing and hedging the option. The final method reviewed for trading realized volatility involves buying or selling an over-the-counter contract whose payoff is an explicit function of volatility. The simplest example of such a volatility contract is a vol swap. This contract pays the buyer the difference between the realized volatility³ and the fixed swap rate determined at the outset of the contract.⁴

A secondary purpose of this chapter is to uncover the link between volatility contracts and some recent path-breaking work by Dupire (1996) and by Derman, Kani, and Kamal (1997) (henceforth DKK). By restricting the set of times and price levels for which returns are used in the volatility calculation, one can synthesize a contract which pays off the “local volatility”, i.e. the volatility which will be experienced should the underlying be at a specified price level at a specified future date. These authors develop the notion of forward local volatility, which is the fixed rate the buyer of the local vol swap pays at maturity in the event that the specified price level is reached. Given a complete term and strike structure of options, the entire forward local volatility surface can be backed out from the prices of options. This surface is the two dimensional analog of the forward rate curve central to the HJM analysis. Following HJM, these authors impose a stochastic process on the forward local volatility surface and derive the risk-neutral dynamics of this surface.

² Jagannathan (1984) shows that in general options need not be increasing in volatility.

³ For marketing reasons, these contracts are usually written on the standard deviation, despite the focus of the literature on spanning contracts on variance.

⁴ This contract is actually a forward contract on realized volatility, but is nonetheless termed a swap.

The outline of this paper is as follows. The next section looks at trading realized volatility via static positions in options. The theory of static replication using options is reviewed in order to develop some new positions for profiting from a correct view on volatility. The subsequent section shows how dynamic trading in the underlying can alternatively be used to create or hedge a volatility exposure. The fourth section looks at over-the-counter volatility contracts as a further alternative for trading volatility. The section shows how such contracts can be synthesized by combining static replication using options with dynamic trading in the underlying asset. A fifth section draws a link between these volatility contracts and the work on forward local volatility pioneered by Dupire and DKK. The final section summarizes and suggests some avenues for future research.

2 Trading realized volatility via static positions in options

The classic position for gaining exposure to volatility is to buy an at-the-money⁵ straddle. Since at-the-money options are frequently used to trade volatility, the implied volatility from these options is widely used as a forecast of subsequent realized volatility. The widespread use of this measure is surprising since the approach relies on a model which itself assumes that volatility is constant.

This section derives an alternative forecast, which is also calculated from market prices of options. In contrast to implied volatility, the forecast does not assume constant volatility, or even that the underlying price process is continuous. In contrast to the implied volatility forecast, our forecast uses the market prices of options of all strikes. In order to develop the alternative forecast, the next subsection reviews the theory of static replication using options developed in Ross (1976) and Breeden and Litzenberger (1978). The following subsection applies this theory to determine a model-free forecast of subsequent realized volatility.

2.1 Static replication with options

Consider a single period setting in which investments are made at time 0 with all payoffs being received at time T . In contrast to the standard intertemporal model, we assume that there are no trading opportunities other than at times 0 and T . We assume there exists a futures market in a risky asset (e.g. a stock index) for delivery at some date $T' \geq T$. We also assume that markets exist for European-style futures options⁶ of all strikes. While the assumption of a continuum of strikes is far

⁵ Note that in the Black model, the sensitivity to volatility of a straddle is actually maximized at slightly below the forward price.

⁶ Note that listed futures options are generally American-style. However, by setting $T' = T$, the underlying futures will converge to the spot at T and so the assumption is that there exists European-style spot options in this special case.

from standard, it is essentially the analog of the standard assumption of continuous trading. Just as the latter assumption is frequently made as a reasonable approximation to an environment where investors can trade frequently, our assumption is a reasonable approximation when there are a large but finite number of option strikes (e.g. for S&P500 futures options).

It is widely recognized that this market structure allows investors to create any smooth function $f(F_T)$ of the terminal futures price by taking a static position at time 0 in options.⁷ Appendix 1 shows that any twice differentiable payoff can be re-written as:

$$f(F_T) = f(\kappa) + f'(\kappa)[(F_T - \kappa)^+ - (\kappa - F_T)^+] + \int_0^\kappa f''(K)(K - F_T)^+ dK + \int_\kappa^\infty f''(K)(F_T - K)^+ dK. \quad (1)$$

The first term can be interpreted as the payoff from a static position in $f(\kappa)$ pure discount bonds, each paying one dollar at T . The second term can be interpreted as the payoff from $f'(\kappa)$ calls struck at κ less $f'(\kappa)$ puts, also struck at κ . The third term arises from a static position in $f''(K)dK$ puts at all strikes less than κ . Similarly, the fourth term arises from a static position in $f''(K)dK$ calls at all strikes greater than κ .

In the absence of arbitrage, a decomposition similar to (1) must prevail among the initial values. Let V_0^f and B_0 denote the initial values of the payoff and the pure discount bond respectively. Similarly, let $P_0(K)$ and $C_0(K)$ denote the initial prices of the put and the call struck at K respectively. Then the no arbitrage condition requires that:

$$V_0^f = f(\kappa)B_0 + f'(\kappa)[C_0(\kappa) - P_0(\kappa)] + \int_0^\kappa f''(K)P_0(K)dK + \int_\kappa^\infty f''(K)C_0(K)dK. \quad (2)$$

Thus, the value of an arbitrary payoff can be obtained from bond and option prices. Note that no assumption was made regarding the stochastic process governing the futures price.

2.2 An alternative forecast of variance

Consider the problem of forecasting the variance of the log futures price relative, $\ln(F_T/F_0)$. For simplicity, we refer to the log futures price relative as a return, even though no investment is required in a futures contract. The variance of the

⁷ This observation was first noted in Breeden and Litzenberger (1978) and established formally in Green and Jarrow (1987) and Nachman (1988).

return over some interval $[0, T]$ is of course given by the expectation of the squared deviation of the return from its mean:

$$\text{Var}_0 \left\{ \ln \left(\frac{F_T}{F_0} \right) \right\} = E_0 \left\{ \ln \left(\frac{F_T}{F_0} \right) - E_0 \left[\ln \left(\frac{F_T}{F_0} \right) \right] \right\}^2. \quad (3)$$

It is well known that futures prices are martingales under the appropriate risk-neutral measure. When the futures contract marks to market continuously, then futures prices are martingales under the measure induced by taking the money market account as numeraire. When the futures contract marks to market daily, then futures prices are martingales under the measure induced by taking a daily rollover strategy as numeraire, where this strategy involves rolling over pure discount bonds with maturities of one day. Thus, given a mark-to-market frequency, futures prices are martingales under the measure induced by the rollover strategy with the same rollover frequency.

If the variance in (3) is calculated using this measure, then $E_0 [\ln (F_T / F_0)]$ can be interpreted as the futures⁸ price of a portfolio of options which pays off $f_m(F) \equiv \ln (F_T / F_0)$ at T . The spot value of this payoff is given by (2) with κ arbitrary and $f_m''(K) = -1/K^2$. Setting $\kappa = F_0$, the futures price of the payoff is given by:

$$\mathcal{F} \equiv E_0 \left[\ln \left(\frac{F_T}{F_0} \right) \right] = - \int_0^{F_0} \frac{1}{K^2} \hat{P}_0(K, T) dK - \int_{F_0}^{\infty} \frac{1}{K^2} \hat{C}_0(K, T) dK,$$

where $\hat{P}_0(K, T)$ and $\hat{C}_0(K, T)$ denote the initial futures price of the put and the call respectively, both for delivery at T . This futures price is initially negative⁹ due to the concavity (negative time value) of the payoff.

Similarly, the variance of returns is just the futures price of the portfolio of options which pays off $f_v(F) = \{\ln (F_T / F_0) - \mathcal{F}\}^2$ at T (see Figure 1). The second derivative of this payoff is $f_v''(K) = 2/K^2 [1 - \ln (K / F_0) + \mathcal{F}]$. This payoff has zero value and slope at $F_0 e^{\mathcal{F}}$. Thus, setting $\kappa = F_0 e^{\mathcal{F}}$, the futures price of the payoff is given by:

$$\begin{aligned} \text{Var}_0 \left\{ \ln \left(\frac{F_T}{F_0} \right) \right\} &= \int_0^{F_0 e^{\mathcal{F}}} \frac{2}{K^2} \left[1 - \ln \left(\frac{K}{F_0} \right) + \mathcal{F} \right] \hat{P}_0(K, T) dK \\ &\quad + \int_{F_0 e^{\mathcal{F}}}^{\infty} \frac{2}{K^2} \left[1 - \ln \left(\frac{K}{F_0} \right) + \mathcal{F} \right] \hat{C}_0(K, T) dK. \end{aligned} \quad (4)$$

⁸ Options do trade futures-style in Hong Kong. However, when only spot option prices are available, one can set $T' = T$ and calculate the mean and variance of the terminal spot under the forward measure. The variance is then expressed in terms of the forward prices of options, which can be obtained from the spot price by dividing by the bond price.

⁹ If the futures price process is a continuous semi-martingale, then Itô's lemma implies that $E_0 [\ln (F_T / F_0)] = -E_0 \frac{1}{2} \int_0^T \sigma_t^2 dt$, where σ_t is the volatility at time t .

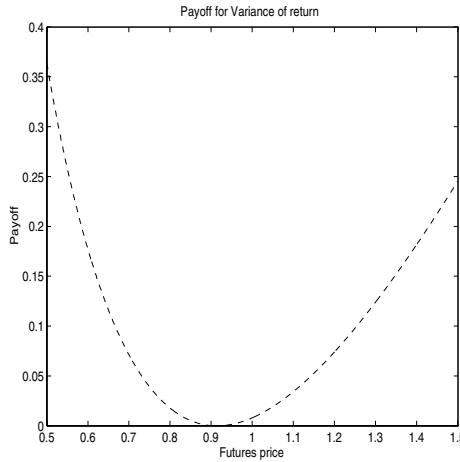


Fig. 1. Payoff for variance of return ($F_0 = 1$; $\mathcal{F} = -0.09$).

At time 0, this futures price is an interesting alternative to implied or historical volatility as a forecast of subsequent realized volatility. However, in common with any futures price, this forecast is a reflection of both statistical expected value and risk aversion. Consequently, by comparing this forecast with the ex-post outcome, the market price of variance risk can be inferred. We will derive a simpler forecast of variance in Section 4 under more restrictive assumptions, principally price continuity.

When compared to an at-the-money straddle, the static position in options used to create f_v has the advantage of maintaining sensitivity to volatility as the underlying moves away from its initial level. Unfortunately, like straddles, these contracts can take on significant price exposure once the underlying moves away from its initial level. An obvious solution to this problem is to delta-hedge with the underlying. The next section considers this alternative.

3 Trading realized volatility by delta-hedging options

The static replication results of the last section made no assumption whatsoever about the price process or volatility process. In order to apply delta-hedging with the underlying futures, we now assume that investors can trade continuously, that interest rates are constant, and that the underlying futures price process is a continuous semi-martingale. Note that we maintain our previous assumption that the volatility of the futures follows an arbitrary unknown stochastic process. While one could specify a stochastic process and develop the correct delta-hedge in such a model, such an approach is subject to significant model risk since one is unlikely

to guess the correct volatility process. Furthermore, such models generally require dynamic trading in options which is costly in practice. Consequently, in what follows we leave the volatility process unspecified and restrict dynamic strategies to the underlying alone. Specifically, we assume that an investor follows the classic replication strategy specified by the Black model, with the delta calculated using a constant volatility σ_h . Since the volatility is actually stochastic,¹⁰ the replication will be imperfect and the error results in either a profit or a loss realized at the expiration of the hedge.

To uncover the magnitude of this $P\&L$, let $V(F, t; \sigma)$ denote the Black model value of a European-style claim given that the current futures price is F and the current time is t . Note that the last argument of V is the volatility used in the calculation of the value. In what follows, it will be convenient to have the attempted replication occur over an arbitrary future period (T, T') rather than over $(0, T)$. Consequently, we assume that the underlying futures matures at some date $T'' \geq T'$.

We suppose that an investor sells a European-style claim at T for the Black model value $V(F_T, T; \sigma_h)$ and holds $\frac{\partial V}{\partial F}(F_t, t; \sigma_h)$ futures contracts over (T, T') . Applying Itô's lemma to $V(F, t; \sigma_h)e^{r(T'-t)}$ gives:

$$\begin{aligned} V(F_{T'}, T'; \sigma_h) &= V(F_T, T; \sigma_h)e^{r(T'-T)} + \int_T^{T'} e^{r(T'-t)} \frac{\partial V}{\partial F}(F_t, t; \sigma_h) dF_t \\ &\quad + \int_T^{T'} e^{r(T'-t)} \left[-rV(F_t, t; \sigma_h) + \frac{\partial V}{\partial t}(F_t, t; \sigma_h) \right] dt \\ &\quad + \int_T^{T'} e^{r(T'-t)} \frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) \frac{F_t^2}{2} \sigma_t^2 dt. \end{aligned} \quad (5)$$

Now, by definition, $V(F, t; \sigma_h)$ solves the Black partial differential equation subject to a terminal condition:

$$-rV(F, t; \sigma_h) + \frac{\partial V}{\partial t}(F, t; \sigma_h) = -\frac{\sigma_h^2 F^2}{2} \frac{\partial^2 V}{\partial F^2}(F, t; \sigma_h), \quad (6)$$

$$V(F, T'; \sigma_h) = f(F). \quad (7)$$

Substituting (6) and (7) in (5) and re-arranging gives:

$$\begin{aligned} f(F_{T'}) + \int_T^{T'} e^{r(T'-t)} \frac{F_t^2}{2} \frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) (\sigma_h^2 - \sigma_t^2) dt \\ = V(F_T, T; \sigma_h)e^{r(T'-T)} + \int_T^{T'} e^{r(T'-t)} \frac{\partial V}{\partial F}(F_t, t; \sigma_h) dF_t. \end{aligned} \quad (8)$$

¹⁰ In an interesting paper, Cherian and Jarrow (1997) show the existence of an equilibrium in an incomplete economy where investors believe the Black-Scholes formula is valid even though volatility is stochastic.

The right hand side is clearly the terminal value of a dynamic strategy comprising an investment at T of $V(F_T, T; \sigma_h)$ dollars in the riskless asset and a dynamic position in $\frac{\partial V}{\partial F}(F_t, t; \sigma_h)$ futures contracts over the time interval (T, T') . Thus, the left hand side must also be the terminal value of this strategy, indicating that the strategy misses its target $f(F_{T'})$ by:

$$P\&L \equiv \int_T^{T'} e^{r(T'-t)} \frac{F_t^2}{2} \frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) (\sigma_h^2 - \sigma_t^2) dt. \quad (9)$$

Thus, when a claim is sold for the implied volatility σ_h at T , the instantaneous $P\&L$ from delta-hedging it over (T, T') is $\frac{F_t^2}{2} \frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) (\sigma_h^2 - \sigma_t^2)$, which is the difference between the hedge variance rate and the realized variance rate, weighted by half the dollar gamma. Note that the $P\&L$ (hedging error) will be zero if the realized instantaneous volatility σ_t is constant at σ_h . It is well known that claims with convex payoffs have nonnegative gammas ($\frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) \geq 0$) in the Black model. For such claims (e.g. options), if the hedge volatility is always less than the true volatility ($\sigma_h < \sigma_t$ for all $t \in [T, T']$), then a loss results, regardless of the path. Conversely, if the claim with a convex payoff is sold for an implied volatility σ_h which dominates¹¹ the subsequent realized volatility at all times, then delta-hedging at σ_h using the Black model delta guarantees a positive $P\&L$.

When compared with static options positions, delta-hedging appears to have the advantage of being insensitive to the price of the underlying. However, (9) indicates that the $P\&L$ at T' does depend on the final price as well as on the price path. An investor with a view on volatility alone would like to immunize the exposure to this path. One solution is to use a stochastic volatility model to conduct the replication of the desired volatility dependent payoff. However, as mentioned previously, this requires specifying a volatility process and employing dynamic replication with options. A better solution is to choose the payoff function $f(\cdot)$, so that the path dependence can be removed or managed. For example, Neuberger (1990) recognized that if $f(F) = 2 \ln F$, then $\frac{\partial^2 V}{\partial F^2}(F_t, t; \sigma_h) = e^{-r(T'-t)} (-2/F_t^2)$ and thus from (9), the $P\&L$ at T' is the payoff of a variance swap $\int_T^{T'} (\sigma_t^2 - \sigma_h^2) dt$. This volatility contract and others related to it are explored in the next section.

4 Trading realized volatility by using volatility contracts

This section shows that several interesting volatility contracts can be manufactured by taking options positions and then delta-hedging them at zero volatility. Accord-

¹¹ See El Karoui, Jeanblanc-Picque, and Shreve (1996) for the extension of this result to the case when the hedger uses a delta-hedging strategy assuming that volatility is a function of stock price and time. Also see Avellaneda *et al.* (1995, 1996) and Lyons (1995) for similar results.

ingly, suppose we set $\sigma_h = 0$ in (8) and negate both sides:

$$\int_T^{T'} \frac{F_t^2}{2} f''(F_t) \sigma_t^2 dt = f(F_{T'}) - f(F_T) - \int_T^{T'} f'(F_t) dF_t. \quad (10)$$

The left hand side is a payoff at T' based on both the realized instantaneous volatility σ_t^2 and the price path. The dependence of this payoff on f arises only through f'' , and accordingly, we will henceforth only consider payoff functions f which have zero value and slope at a given point κ . The right hand side of (10) depends only on the price path and results from adding the following three payoffs:

1. The payoff from a static position in options maturing at T' paying $f(F_{T'})$ at T' .
2. The payoff from a static position in options maturing at T paying $-e^{-r(T'-T)} f(F_T)$ and future-valued to T' .
3. The payoff from maintaining a dynamic position in $-e^{-r(T'-t)} f'(F_t)$ futures contracts over the time interval (T, T') (assuming continuous marking-to-market and that the margin account balance earns interest at the risk-free rate).

Thus, the payoff on the left hand side can be achieved by combining a static position in options as discussed in Section 2, with a dynamic strategy in futures as discussed in Section 3. The dynamic strategy can be interpreted as an attempt to create the payoff $-f(F_{T'})$ at T' , conducted under the false assumption of zero volatility. Since realized volatility will be positive, an error arises, and the magnitude of this error is given by $\int_T^{T'} \frac{F_t^2}{2} f''(F_t) \sigma_t^2 dt$, which is the left side of (10). The payoff $f(\cdot)$ can be chosen so that when its second derivative is substituted into this expression, the dependence on the path is consistent with the investor's joint view on volatility and price. In this section, we consider the following three second derivatives of payoffs at T' and work out the $f(\cdot)$ which leads to them:

Description of payoff	$f''(F_t)$	Payoff at T'
Variance over future period	$\frac{2}{F_t^2}$	$\int_T^{T'} \sigma_t^2 dt$
Future corridor variance	$\frac{2}{F_t^2} 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)]$	$\int_T^{T'} 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)] \sigma_t^2 dt$
Future variance along strike	$\frac{2}{\kappa^2} \delta(F_t - \kappa)$	$\int_T^{T'} \delta(F_t - \kappa) \sigma_t^2 dt.$

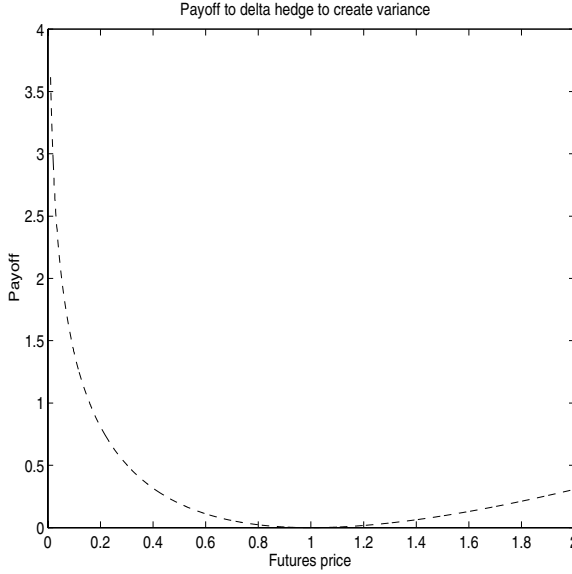


Fig. 2. Payoff to delta-hedge to create contract paying variance ($\kappa = 1$).

4.1 Contract paying future variance

Consider the following payoff function $\phi(F)$ (see Figure 2):

$$\phi(F) \equiv 2 \left[\ln \left(\frac{\kappa}{F} \right) + \frac{F}{\kappa} - 1 \right], \quad (11)$$

where κ is an arbitrary finite positive number. The first derivative is given by:

$$\phi'(F) = 2 \left[\frac{1}{\kappa} - \frac{1}{F} \right]. \quad (12)$$

Thus, the value and slope both vanish at $F = \kappa$. The second derivative of ϕ is simply:

$$\phi''(F) = \frac{2}{F^2}. \quad (13)$$

Substituting (11) to (13) into (10) results in a relationship between a contract paying the realized variance over the time interval (T, T') and three payoffs based on price:

$$\begin{aligned} \int_T^{T'} \sigma_t^2 dt &= 2 \left[\ln \left(\frac{\kappa}{F_{T'}} \right) + \frac{F_{T'}}{\kappa} - 1 \right] - 2 \left[\ln \left(\frac{\kappa}{F_T} \right) + \frac{F_T}{\kappa} - 1 \right] \\ &\quad - 2 \int_T^{T'} \left[\frac{1}{\kappa} - \frac{1}{F_t} \right] dF_t. \end{aligned} \quad (14)$$

The first two terms on the right hand side arise from static positions in options. Substituting (13) into (2) implies that for each term the required position is given by:

$$2 \left[\ln \left(\frac{\kappa}{F} \right) + \frac{F}{\kappa} - 1 \right] = \int_0^\kappa \frac{2}{K^2} (K - F)^+ dK + \int_\kappa^\infty \frac{2}{K^2} (F - K)^+ dK, \quad (15)$$

Thus, to create the contract paying $\int_T^{T'} \sigma_t^2 dt$ at T' , at $t = 0$, the investor should buy options at the longer maturity T' and sell options at the nearer maturity T . The initial cost of this position is given by:

$$\int_0^\kappa \frac{2}{K^2} P_0(K, T') dK + \int_\kappa^\infty \frac{2}{K^2} C_0(K, T') dK - e^{-r(T'-T)} \left[\int_0^\kappa \frac{2}{K^2} P_0(K, T) dK + \int_\kappa^\infty \frac{2}{K^2} C_0(K, T) dK \right]. \quad (16)$$

When the nearer maturity options expire, the investor should borrow to finance the payout of

$2e^{-r(T'-T)} [\ln(\kappa/F_T) + (F_T/\kappa) - 1]$. At this time, the investor should also start a dynamic strategy in futures, holding $-2e^{-r(T'-t)} [(1/\kappa) - (1/F_t)]$ futures contracts for each $t \in [T, T']$. The net payoff at T' is:

$$2 \left[\ln \left(\frac{\kappa}{F_{T'}} \right) + \frac{F_{T'}}{\kappa} - 1 \right] - 2 \left[\ln \left(\frac{\kappa}{F_T} \right) + \frac{F_T}{\kappa} - 1 \right] - 2 \int_T^{T'} \left[\frac{1}{\kappa} - \frac{1}{F_t} \right] dF_t = \int_T^{T'} \sigma_t^2 dt,$$

as required. Since the initial cost of achieving this payoff is given by (16), an interesting forecast $\hat{\sigma}_{T,T'}^2$ of the variance between T and T' is given by the future value of this cost:

$$\hat{\sigma}_{T,T'}^2 = e^{rT'} \int_0^\kappa \frac{2}{K^2} P_0(K, T') dK + \int_\kappa^\infty \frac{2}{K^2} C_0(K, T') dK - e^{rT} \left[\int_0^\kappa \frac{2}{K^2} P_0(K, T) dK + \int_\kappa^\infty \frac{2}{K^2} C_0(K, T) dK \right].$$

In contrast to implied volatility, this forecast does not use a model in which volatility is assumed to be constant. However, in common with any forward price, this forecast is a reflection of both statistical expected value and risk aversion. Consequently, by comparing this forecast with the ex-post outcome, the market price of volatility risk can be inferred.

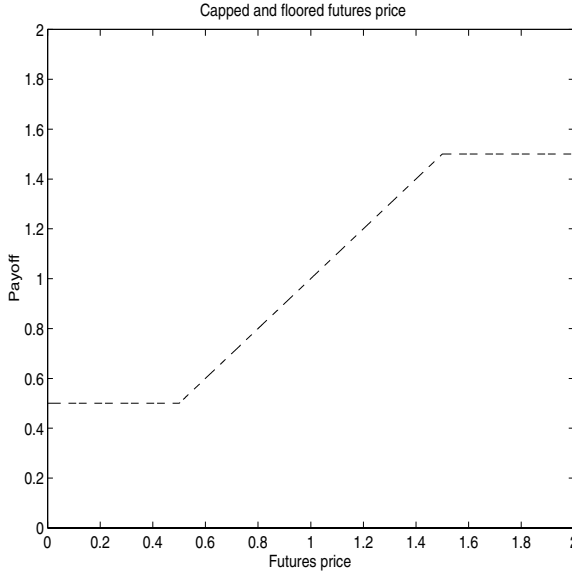


Fig. 3. Futures price capped and floored ($\kappa = 1$, $\Delta\kappa = 0.5$).

4.2 Contract paying future corridor variance

In this subsection, we generalize to a contract which pays the “corridor variance”, defined as the variance calculated using only the returns at times for which the futures price is within a specified corridor. In particular, consider a corridor $(\kappa - \Delta\kappa, \kappa + \Delta\kappa)$ centered at some arbitrary level κ and with width $2\Delta\kappa$. Suppose that we wish to generate a payoff at T' of $\int_T^{T'} 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)] \sigma_t^2 dt$. Thus, the variance calculation is based only on returns at times in which the futures price is inside the corridor.

Consider the following payoff $\phi_{\Delta\kappa}(\cdot)$:

$$\phi_{\Delta\kappa}(F) \equiv 2 \left[\ln \left(\frac{\kappa}{\bar{F}} \right) + F \left(\frac{1}{\kappa} - \frac{1}{\bar{F}} \right) \right], \quad (17)$$

where:

$$\bar{F}_t \equiv \max[\kappa - \Delta\kappa, \min(F_t, \kappa + \Delta\kappa)]$$

is the futures price floored at $\kappa - \Delta\kappa$ and capped at $\kappa + \Delta\kappa$ (see Figure 3).

From inspection, the payoff $\phi_{\Delta\kappa}(\cdot)$ is the same as ϕ defined in (11), but with F replaced by $\bar{F}$. The new payoff is graphed in Figure 4: this payoff is actually a generalization of (11) since $\lim_{\Delta\kappa \uparrow \infty} \bar{F} = F$. For a finite corridor width, the payoff $\phi_{\Delta\kappa}(F)$ matches $\phi(F)$ for futures prices within the corridor. Consequently, like $\phi(F)$, $\phi_{\Delta\kappa}(F)$ has zero value and slope at $F = \kappa$. However, in contrast to

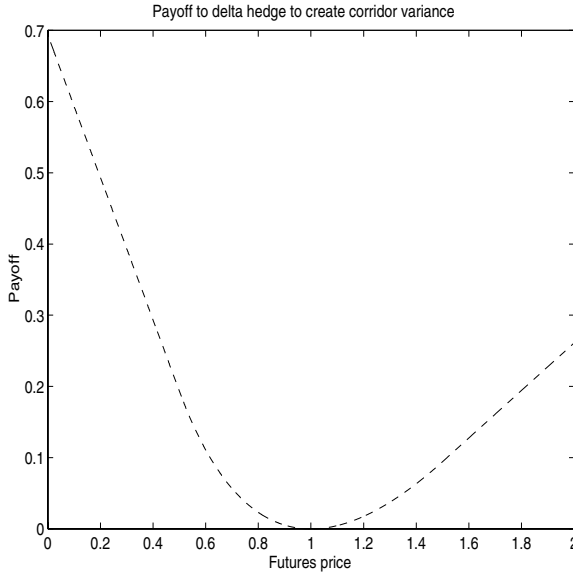


Fig. 4. Trimming the log payoff ($\kappa = 1$, $\Delta\kappa = 0.5$).

$\phi(F)$, $\phi_{\Delta\kappa}(F)$ is linear outside the corridor with the lines chosen so that the payoff is continuous and differentiable at $\kappa \pm \Delta\kappa$. The first derivative of (17) is given by:

$$\phi'_{\Delta\kappa}(F) = 2 \left[\frac{1}{\kappa} - \frac{1}{\underline{F}} \right], \quad (18)$$

while the second derivative is simply:

$$\phi''_{\Delta\kappa}(F) = \frac{2}{F^2} 1[F \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)]. \quad (19)$$

Substituting (17) to (19) into (10) implies that the volatility-based payoff decomposes as:

$$\begin{aligned} \int_T^{T'} \sigma_t^2 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)] dt &= 2 \left[\ln \left(\frac{\kappa}{\underline{F}_{T'}} \right) + F_{T'} \left(\frac{1}{\kappa} - \frac{1}{\underline{F}_{T'}} \right) \right] \\ &\quad - 2 \left[\ln \left(\frac{\kappa}{\underline{F}_T} \right) + F_T \left(\frac{1}{\kappa} - \frac{1}{\underline{F}_T} \right) \right] - 2 \int_T^{T'} \left[\frac{1}{\kappa} - \frac{1}{\underline{F}_t} \right] dF_t. \end{aligned}$$

The payoff function $\phi_{\Delta\kappa}(\cdot)$ has no curvature outside the corridor and consequently the static positions in options needed to create the first two terms will not require strikes set outside the corridor. Thus, to create the contract paying the future corridor variance, $\int_T^{T'} \sigma_t^2 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)] dt$ at T' , the investor should initially only buy and sell options struck within the corridor, for an initial

cost of:

$$\int_{\kappa-\Delta\kappa}^{\kappa} \frac{2}{K^2} P_0(K, T') dK + \int_{\kappa}^{\kappa+\Delta\kappa} \frac{2}{K^2} C_0(K, T') dK \\ - e^{-r(T'-T)} \left[\int_{\kappa-\Delta\kappa}^{\kappa} \frac{2}{K^2} P_0(K, T) dK + \int_{\kappa}^{\kappa+\Delta\kappa} \frac{2}{K^2} C_0(K, T) dK \right].$$

At $t = T$, the investor should borrow to finance the payout of $2e^{-r(T'-T)} [\ln(\kappa/\bar{F}_T) + F_T((1/\kappa) - (1/\bar{F}_T))]$ from having initially written the T maturity options. The investor should also start a dynamic strategy in futures, holding $-2e^{-r(T'-t)} [(1/\kappa) - (1/\bar{F}_t)]$ futures contracts for each $t \in [T, T']$. This strategy is semi-static in that no trading is required when the futures price is outside the corridor. The net payoff at T' is:

$$2 \left[\ln\left(\frac{\kappa}{\bar{F}_{T'}}\right) + F_{T'} \left(\frac{1}{\kappa} - \frac{1}{\bar{F}_{T'}} \right) \right] - 2 \left[\ln\left(\frac{\kappa}{\bar{F}_T}\right) + F_T \left(\frac{1}{\kappa} - \frac{1}{\bar{F}_T} \right) \right] \\ - 2 \int_T^{T'} \left[\frac{1}{\kappa} - \frac{1}{\bar{F}_t} \right] dF_t = \int_T^{T'} \sigma_t^2 1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)] dt,$$

as desired.

4.3 Contract paying future variance along a strike

In the last subsection, only options struck within the corridor were used in the static options position, and dynamic trading in the underlying futures was required only when the futures price was in the corridor. In this subsection, we shrink the width of the corridor of the last subsection down to a single point and examine the impact on the volatility based payoff and its replicating strategy. In order that this payoff have a non-negligible value, all asset positions in Subsection 4.2 must be re-scaled by $1/2\Delta\kappa$. Thus, the volatility-based payoff at T' would instead be $\int_T^{T'} \frac{1[F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa)]}{2\Delta\kappa} \sigma_t^2 dt$. By letting $\Delta\kappa \downarrow 0$, the variance received can be completely localized in the spatial dimension to $\int_T^{T'} \delta(F_t - \kappa) \sigma_t^2 dt$, where $\delta(\cdot)$ denotes a Dirac delta function.¹² Recalling that only options struck within the corridor are used to create the corridor variance, the initial cost of creating this localized cash flow is given by the following ratioed calendar spread of straddles:

$$\frac{1}{\kappa^2} [V_0(\kappa, T') - e^{-r(T'-T)} V_0(\kappa, T)],$$

¹² The Dirac delta function is a generalized function characterized by two properties:

- (i) $\delta(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ \infty & \text{if } x = 0 \end{cases}$
- (ii) $\int_{-\infty}^{\infty} \delta(x) dx = 1$.

See Richards and Youn (1990) for an accessible introduction to such generalized functions.

where $V_0(\kappa, T)$ is the initial cost of a straddle struck at κ and maturing at T :

$$V_0(\kappa, T) \equiv P_0(\kappa, T) + C_0(\kappa, T).$$

As usual, at $t = T$, the investor should borrow to finance the payout of $|F_T - \kappa|/\kappa^2$ from having initially written the T maturity straddle. Appendix 2 proves that the dynamic strategy in futures initiated at T involves holding $-\frac{e^{-r(T'-t)}}{\kappa^2} \text{sgn}(F_t - \kappa)$ futures contracts, where $\text{sgn}(x)$ is the sign function:

$$\text{sgn}(x) \equiv \begin{cases} -1 & \text{if } x < 0; \\ 0 & \text{if } x = 0; \\ 1 & \text{if } x > 0. \end{cases}$$

When $T = 0$, this strategy reduces to the initial purchase of a straddle maturing at T' , initially borrowing $e^{-rT'}|F_0 - \kappa|$ dollars and holding $-\frac{e^{-r(T'-t)}}{\kappa^2} \text{sgn}(F_t - \kappa)$ futures contracts for $t \in (0, T')$. The component of this strategy involving borrowing and futures is known as the stop-loss start-gain strategy, previously investigated by Carr and Jarrow (1990). By the Tanaka–Meyer formula,¹³ the difference between the payoff from the straddles and this dynamic strategy is known as the *local time* of the futures price process. Local time is a fundamental concept in the study of one dimensional stochastic processes. Fortunately, a straddle combined with a stop-loss start-gain strategy in the underlying provides a mechanism for synthesizing a contract paying off this fundamental concept. The initial time value of the straddle is the market's (risk-neutral) expectation of the local time. By comparing this time value with the ex-post outcome, the market price of local time risk can be inferred.

5 Connection to recent work on stochastic volatility

The last contract examined in the last section represents the limit of a localization in the futures price. When a continuum of option maturities is also available, we may additionally localize in the time dimension as has been done in some recent work by Dupire (1996) and DKK (1997). Accordingly, suppose we further re-scale all the asset positions described in Subsection 4.3 by $1/\Delta T$, where $\Delta T \equiv T' - T$. The payoff at T' would instead be:

$$\int_T^{T'} \frac{\delta(F_t - \kappa)}{\Delta T} \sigma_t^2 dt.$$

The cost of creating this position would be:

$$\frac{1}{\kappa^2} \left[\frac{V_0(\kappa, T') - e^{-r(T'-T)} V_0(\kappa, T)}{\Delta T} \right].$$

¹³ See Karatzas and Shreve (1988), p. 220.

By letting $\Delta T \downarrow 0$, one gets the beautiful result of Dupire (1996) that $\frac{1}{\kappa^2} \left[\frac{\partial V_0}{\partial T}(\kappa, T) + r V_0(\kappa, T) \right]$ is the cost of creating the payment $\delta(F_T - \kappa) \sigma_T^2$ at T . As shown in Dupire, the forward local variance can be defined as the number of butterfly spreads paying $\delta(F_T - \kappa)$ at T one must sell in order to finance the above option position initially. A discretized version of this result can be found in DKK (1997). One can go on to impose a stochastic process on the forward local variance, as in Dupire (1996) and in DKK (1997). These authors derive conditions on the risk-neutral drift of the forward local variance, allowing replication of price or volatility-based payoffs using dynamic trading in only the underlying asset and a single option.¹⁴ In contrast to earlier work on stochastic volatility, the form of the market price of volatility risk need not be specified.

Summary and suggestions for future research

We reviewed three approaches for trading volatility. While static positions in options do generate exposure to volatility, they also generate exposure to price. Similarly, a dynamic strategy in futures alone can yield a volatility exposure, but always has a price exposure as well. By combining static positions in options with dynamic trading in futures, payoffs related to realized volatility can be achieved which have either no exposure to price, or which have an exposure contingent on certain price levels being achieved in specified time intervals.

Under certain assumptions, we were able to price and hedge certain volatility contracts without specifying the process for volatility. The principle assumption made was that of price continuity. Under this assumption, a calendar spread of options emerges as a simple tool for trading the local volatility (or local time) between the two maturities. It would be interesting to see if this insight survives the relaxation of the critical assumption of price continuity. It would also be interesting to consider contracts which pay nonlinear functions of realized variance or local variance. Finally, it would be interesting to develop contracts on other statistics of the sample path such as the Sharpe ratio, skewness, covariance, correlation, etc. In the interests of brevity, such inquiries are best left for future research.

Appendix 1: Spanning with bonds and options

For any payoff $f(F)$, the sifting property of a Dirac delta function implies:

$$f(F) = \int_0^\infty f(K) \delta(F - K) dK$$

¹⁴ When two Brownian motions drive the price and the forward local volatility surface, any two assets whose payoffs are not co-linear can be used to span.

$$= \int_0^\kappa f(K)\delta(F-K)dK + \int_0^\kappa f(K)\delta(F-K)dK,$$

for any nonnegative κ . Integrating each integral by parts implies:

$$\begin{aligned} f(F) &= f(K)1(F < K) \Big|_0^\kappa - \int_0^\kappa f'(K)1(F < K)dK \\ &\quad + f(K)1(F \geq K) \Big|_\kappa^\infty + \int_\kappa^\infty f'(K)1(F \geq K)dK. \end{aligned}$$

Integrating each integral by parts once more implies:

$$\begin{aligned} f(F) &= f(\kappa)1(F < \kappa) - f'(K)(K-F)^+ \Big|_0^\kappa + \int_0^\kappa f''(K)(K-F)^+dK \\ &\quad + f(\kappa)1(F \geq \kappa) - f'(K)(F-K)^+ \Big|_\kappa^\infty + \int_\kappa^\infty f''(K)(F-K)^+dK \\ &= f(\kappa) + f'(\kappa)[(F-\kappa)^+ - (\kappa-F)^+] \\ &\quad + \int_0^\kappa f''(K)(K-F)^+dK + \int_\kappa^\infty f''(K)(F-K)^+dK. \end{aligned}$$

Appendix 2: Derivation of futures position when synthesizing contract paying future variance along a strike

Recall from Section 4.3, that all asset positions in Section 4.2 were normalized by multiplying by $1/2\Delta\kappa$. Thus in particular, the futures position of $-2e^{-r(T'-t)}[(1/\kappa) - (1/\bar{F}_t)]$ contracts in Subsection 4.2 is changed to $-e^{-r(T'-t)}/\Delta\kappa[(1/\kappa) - (1/\bar{F}_t)]$ contracts in Subsection 4.3. More explicitly, the number of contracts held is given by

$$\begin{cases} -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[\frac{1}{\kappa} - \frac{1}{\kappa - \Delta\kappa} \right] & \text{if } F_t \leq \kappa - \Delta\kappa; \\ -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[\frac{1}{\kappa} - \frac{1}{F_t} \right] & \text{if } F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa); \\ -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[\frac{1}{\kappa} - \frac{1}{\kappa + \Delta\kappa} \right] & \text{if } F_t \geq \kappa + \Delta\kappa. \end{cases}$$

Now, by Taylor's series:

$$\frac{1}{\kappa - \Delta\kappa} = \frac{1}{\kappa} + \frac{1}{\kappa^2}\Delta\kappa + O(\Delta\kappa^2)$$

and:

$$\frac{1}{\kappa + \Delta\kappa} = \frac{1}{\kappa} - \frac{1}{\kappa^2}\Delta\kappa + O(\Delta\kappa^2).$$

Substitution implies that the number of futures contracts held is given by:

$$\left\{ \begin{array}{ll} -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[-\frac{1}{\kappa^2} \Delta\kappa + O(\Delta\kappa^2) \right] & \text{if } F_t \leq \kappa - \Delta\kappa; \\ -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[\frac{1}{\kappa} - \frac{1}{F_t} \right] & \text{if } F_t \in (\kappa - \Delta\kappa, \kappa + \Delta\kappa); \\ -\frac{e^{-r(T'-t)}}{\Delta\kappa} \left[\frac{1}{\kappa^2} \Delta\kappa + O(\Delta\kappa^2) \right] & \text{if } F_t \geq \kappa + \Delta\kappa. \end{array} \right.$$

Thus, as $\Delta\kappa \downarrow 0$, the number of futures contracts held converges to $-\frac{e^{-r(T'-t)}}{\kappa^2} \text{sgn}(F_t - \kappa)$, where $\text{sgn}(x)$ is the sign function:

$$\text{sgn}(x) \equiv \begin{cases} -1 & \text{if } x < 0; \\ 0 & \text{if } x = 0; \\ 1 & \text{if } x > 0. \end{cases}$$

Acknowledgements

We thank the participants of presentations at Boston University, the NYU Courant Institute, M.I.T., Morgan Stanley, and the Risk 1997 Congress. We would also like to thank Marco Avellaneda, Joseph Cherian, Stephen Chung, Emanuel Derman, Raphael Douady, Bruno Dupire, Ognian Enchev, Chris Fernandes, Marvin Friedman, Iraj Kani, Keith Lewis, Harry Mendell, Lisa Polsky, John Ryan, Murad Taqqu, Alan White, and especially Robert Jarrow for useful discussions. They are not responsible for any errors.

References

- Avellaneda, M., Lévy, A. and Paras, A., 1995, Pricing and hedging derivative securities in markets with uncertain volatilities, *Applied Mathematical Finance*, **2**, 73–88.
- Avellaneda, M., Lévy, A. and Paras, A., 1996, Managing the volatility risk of portfolios of derivative securities: The Lagrangian uncertain volatility model, *Applied Mathematical Finance*, **3**, 21–52.
- Breeden, D. and Litzenberger, R., 1978, Prices of state contingent claims implicit in option prices, *Journal of Business*, **51**, 621–51.
- Brenner, M., and Galai, D., 1989, New financial instruments for hedging changes in volatility, *Financial Analyst's Journal*, July–August 1989, 61–5.
- Brenner, M., and Galai, D., 1993, Hedging volatility in foreign currencies, *The Journal of Derivatives*, Fall 1993, 53–9.
- Brenner, M., and Galai, D., 1996, Options on volatility, Chapter 13 of *Option Embedded Bonds*, I. Nelken, ed. 273–86.
- Carr P. and Jarrow, R., 1990, The stop-loss start-gain strategy and option valuation: a new decomposition into intrinsic and time value, *Review of Financial Studies*, **3**, 469–92.
- Carr P. and Madan, D., 1997, Optimal positioning in derivative securities, Morgan Stanley working paper.

- Cherian, J., and Jarrow, R., 1998, Options markets, self-fulfilling prophecies and implied volatilities, *Review of Derivatives Research* **2**, 5–37.
- Derman E., Kani, I. and Kamal, M., 1997, Trading and hedging local volatility, *Journal of Financial Engineering*, **6**, 3, 233–68.
- Dupire B., 1993, Model Art, *Risk*. Sept. 1993, p. 118 and 120.
- Dupire B., 1996, A unified theory of volatility, Paribas working paper.
- El Karoui, N., Jeanblanc-Picque, M. and Shreve, S., 1996, Robustness of the Black and Scholes formula, Carnegie Mellon University working paper.
- Fleming, J., Ostdiek, B. and Whaley, R., 1993, Predicting stock market volatility: a new measure, Duke University working paper.
- Galai, D., 1979, A proposal for indexes for traded call options, *Journal of Finance*, **XXXIV**, 5, 1157–72.
- Gastineau, G., 1977, An index of listed option premiums, *Financial Analyst's Journal*, May–June 1977.
- Green, R.C. and Jarrow, R.A., 1987, Spanning and completeness in markets with contingent claims, *Journal of Economic Theory*, **41**, 202–10.
- Grunbichler A., and Longstaff, F., 1993, Valuing options on volatility, UCLA working paper.
- Heath, D., Jarrow, R. and Morton, A., 1992, Bond pricing and the term structure of interest rates: a new methodology for contingent claim valuation, *Econometrica*, **66** 77–105.
- Jagannathan R., 1984, Call options and the risk of underlying securities, *Journal of Financial Economics*, **13**, 3, 425–34.
- Karatzas, I., and Shreve, S., 1988, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.
- Lyons, T., 1995, Uncertain volatility and the risk-free synthesis of derivatives, *Applied Mathematical Finance*, **2**, 117–33.
- Nachman, D., 1988, Spanning and completeness with options, *Review of Financial Studies*, **3**, 31, 311–28.
- Neuberger, A. 1990, Volatility trading, London Business School working paper.
- Richards, J.I., and Youn, H.K., 1990 *Theory of Distributions: A Non-technical Introduction*, Cambridge University Press, 1990.
- Ross, S., 1976, Options and efficiency, *Quarterly Journal of Economics*, **90** Feb., 75–89.
- Whaley, R., 1993, Derivatives on market volatility: hedging tools long overdue, *The Journal of Derivatives*, Fall 1993, 71–84.

Shortfall Risk in Long-Term Hedging with Short-Term Futures Contracts

Paul Glasserman

1 Introduction

Consider a firm with a commitment to deliver a fixed quantity of oil at a specified date T in the future. The commitment exposes the firm to the price of oil at time T . Suppose the firm buys futures contracts for an equal quantity of oil and for settlement at the same date T . In so doing, it has eliminated its exposure to the price of oil at T , but has it entirely eliminated its risk? If the futures contracts are marked-to-market – requiring, in particular, that the firm make payments should the futures price drop – but the forward commitment is not, then in eliminating its price exposure at time T the firm has potentially increased the risk of a cash shortfall before time T because of the funding requirements of the hedge. The possibility of an increased risk is even clearer if the original horizon T is long (say five years) but the futures contracts have a short maturity (say one month). The firm may seek to hedge the long-dated commitment through a sequence of short-term contracts, but this exposes the firm to price risk each time one contract is settled and the next is opened. In particular, should the price of oil decrease, funding the hedge will require infusions of additional cash.¹

The purpose of this chapter is to propose and illustrate a simple measure of the risk of a cash shortfall arising from the funding requirements of a futures hedge. We give particular attention to the probability of a large shortfall anytime *up to* a specified horizon as opposed to merely *at* that horizon. Rough approximations to such probabilities are available through the theory of Gaussian extremes (as in Adler (1990) and Piterbarg (1996)) and the theory of large deviations (as in Dembo and Zeitouni (1998) and Stroock (1984)); we compare the shortfall risk in alternative hedging strategies through these approximations.

Our analysis is motivated in part by the recent debate regarding the widely publicized derivatives losses of Metallgesellschaft Refining and Marketing (MGRM);

¹ See Appendix A for a brief review of futures and forward contracts.

see Benson (1994), Culp and Miller (1995), Edwards and Canter (1995), and Mello and Parsons (1995a) for accounts of this incident, and see Brennan and Crew (1995), Carverhill (1998), Hilliard (1996), Neuberger (1995), and Ross (1995) for related analyses. Briefly, MGRM had entered into long-term contracts to supply oil at fixed prices and was (ostensibly) hedging these commitments with one-month futures contracts. In 1993, as the price of oil dropped and the hedging strategy required increasingly large infusions of cash, MGRM's parent company found it necessary to abandon the strategy, resulting in derivatives losses reported in press accounts to exceed \$1 billion. In theory, as the price of oil dropped the value of the supply contracts increased, but in fact MGRM was forced to unwind its contracts on unfavorable terms.

Because of the complexities of this case and the many aspects that remain undisclosed, we do not attempt a direct application. We focus instead on an admittedly simple model of a central aspect of MGRM's strategy: the use of a *rolling stack* of short-dated futures contract to hedge long-term supply commitments. In this strategy, futures contracts are rolled into the next maturity as they expire, but the number of contracts is decreased over time to reflect the decrease in the remaining commitment in the supply contracts.

A primary objective of such a hedging strategy is to protect the firm from the effects of large price fluctuations. It is therefore reasonable to examine how effectively the rolling stack accomplishes this. In the simple single-factor model we study, the rolling stack eliminates the effect of spot price fluctuations completely – but only at the end of the hedging horizon. Early in the life of the hedge, the use of short-dated contracts increases the risk of a cash shortfall; we quantify this effect.

As a prelude to our analysis, consider the comparison in Figure 1. The solid lines plot the variance of the cash balance resulting from a long-term supply contract with and without hedging, based on a simple model of independent and identically distributed price changes. (The precise assumptions leading to these graphs are reviewed in Section 2.) Not surprisingly, the variance in the unhedged case increases over time. The variance of the hedged cumulative cashflow at the end of the horizon is zero, but (as noted by Mello and Parsons 1995b) early in the life of the contract the hedged variance is larger. This is certainly suggestive of an increased risk, but it is not immediately clear how to make this suggestion precise. At best, the curves give an indication of the relative probabilities of a cash shortfall at each *fixed* time t – what we will call the *spot* risk at time t – with and without hedging. They do not explicitly compare the more relevant probabilities of a cash shortfall any time *up to* time t , which we will call the *running* risk. We will argue that comparing spot risks understates the real shortfall risk resulting from the hedge. Indeed, one of our main conclusions, following from a result on Gaussian extremes, is that the unhedged variance should be compared with the running maximum of the hedged

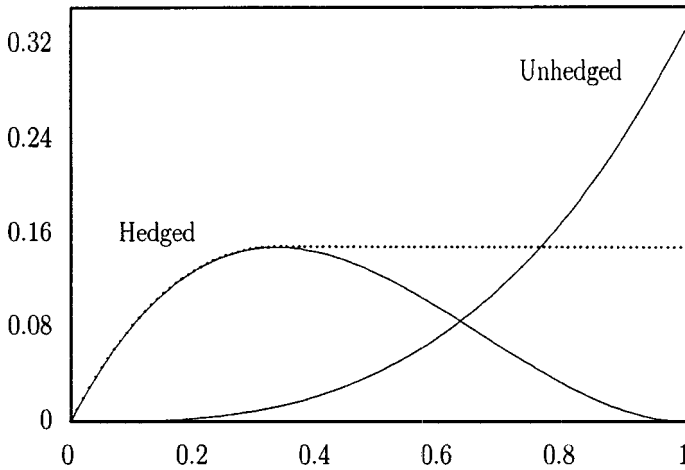


Fig. 1. Variance of unhedged and fully hedged cash balance over the life of the exposure. The dotted line indicates the running maximum of the hedged variance.

variance, indicated by the dotted line in Figure 1. Clearly, the dotted line assigns greater risk to the hedging strategy than does the corresponding solid line.

If the objective of a hedge is (at least in part) to reduce the chance of a cash shortfall, then the running risk is a relevant measure. Based on this premise and a measure of running risk, we make several observations. These will be detailed in later sections, but we highlight a few here. (a) A full rolling-stack hedge increases the risk of a cash shortfall for roughly 3/4 of the hedging horizon. (b) Under a full hedge, a cash shortfall is most likely to occur near 1/3 of the hedging horizon, and with no hedging it is most likely to occur near the end of the horizon. (c) Even under conditions that make the minimum-variance hedge ratio 1, a substantially smaller hedge ratio minimizes the running risk. (d) With a hedge ratio of 1, the optimal hedging horizon is substantially shorter than the full horizon.

We elaborate these conclusions in a model of spot prices that allows (but does not require) mean reversion. So, we have four basic cases: mean reverting or not, hedged or not. We will see that the degree of mean reversion has a major impact on both the appropriate extent and the effectiveness of hedging with short-dated futures. For each case, in addition to comparing risks of a cash shortfall, we identify *the most likely path* to a shortfall, in a sense to be made precise. Each such path solves a problem in the calculus of variations suggested by the theory of large deviations. These “optimal” paths give information about *how* risky events occur and not just their probability of occurrence. They may be thought of as “stress testing” scenarios of the type commonly formulated in practice on an *ad hoc* basis, here arrived at through a precise methodology.

A shortcoming of our analysis is that it rests on a single-factor model of spot and futures prices. As a consequence, we cannot fully model an unexpected shift from backwardation to contango of the type that seems to have precipitated MGRM's crisis. Indeed, as discussed by Benson (1994) and analyzed by Edwards and Canter (1995), the shape of the term structure of commodity prices is central to the rolling stack as a profit-generating strategy, as opposed to merely a hedge. (See Brennan and Crew (1997), Brennan (1991), Garbade (1993), Gibson and Schwartz (1990), Hilliard (1996), and Neuberger (1999) for some relevant multifactor models of commodity prices.) The tools we apply may, however, be extended to multifactor models.

Although we develop just one application here, it seems likely that the methods we use are relevant to other problems in risk management. There is, in particular, a close formal parallel between the model we consider and the exposure over time in an interest rate swap when interest rates follow the Vasicek (1977) model. The approach we follow in identifying price paths leading to shortfalls may be useful in constructing stress testing scenarios in other settings, or as a means of approximating value-at-risk. The evolution of exposures over time also plays a role in setting counterparty credit limits for swaps and other transactions. For background on these ideas, see Frye (1997), Jorion (1997), Picoult (1998), Wakeman (1999), and Wilson (1999).

The rest of this paper is organized as follows. Section 2 introduces the mechanics of the rolling stack and details our model of spot and futures prices, starting from a discrete-time formulation and then making a continuous-time approximation. Section 3 presents a measure of risk; Sections 4 and 5 develop the consequences of this measure with and without mean reversion, respectively. Section 6 presents the most likely paths to a cash shortfall. Section 7 compares our analysis (which is based on the continuous-time model) with simulations in discrete time. Some concluding remarks are collected in Section 8 and some technical issues are deferred to two appendices.

2 A model of exposure and hedging

Our point of departure is a simple model containing the essential features of examples discussed by Culp and Miller (1995) and Mello and Parsons (1995b) in their discussions of MGRM's hedging strategy. Consider a firm that commits to supplying a fixed quantity q of a commodity at a fixed price a at dates $n = 1, \dots, N$. The market price of the commodity at these dates is described by the sequence

$$S_n = c + \sum_{i=1}^n X_i, \quad n = 1, 2, \dots \quad (1)$$

At this point, we do not make any assumptions about the price increments X_i . If the firm's cost equals the market price, then at time n it earns $q(a - S_n)$, and its cumulative cashflow to time k is

$$C_k = q \sum_{n=1}^k (a - S_n) = q \left(k(a - c) - \sum_{n=1}^k \sum_{i=1}^n X_i \right). \quad (2)$$

Let $F_{n,n+1}$ be the time- n futures price for a contract on the underlying commodity maturing at $n + 1$, and set $b_{n,n+1} = F_{n,n+1} - S_n$. We use $b_{n,n+1}$ as a surrogate for an explicit model of the determinants of the cash-futures spread. Consider a *rolling stack* hedging strategy that buys $q(N - n)$ of these short-dated contracts at time n . Each contract bought at time n generates a profit or loss of $S_{n+1} - F_{n,n+1}$ at $n + 1$, so the cumulative cashflow to time k from the hedge is given by

$$\begin{aligned} H_k &= q \sum_{n=1}^k (N - n + 1) [S_n - F_{n-1,n}] \\ &= q \sum_{n=1}^k (N - n + 1) (X_n - b_{n-1,n}). \end{aligned} \quad (3)$$

Interchanging the order of summation in (2) yields

$$C_k = qk(a - c) - q \sum_{n=1}^k (k - n + 1) X_n. \quad (4)$$

Combining (3) and (4) and taking $k = N$, we see that the cash balance from the delivery contract and hedge combined, at the terminal date N , is

$$\bar{C}_N = C_N + H_N = qN(a - c) - q \sum_{n=1}^N (N - n + 1) b_{n-1,n}. \quad (5)$$

In particular, the hedging strategy exactly cancels the price increments X_n at time N , but – comparing the coefficients on X_n in (3) and (4) – only at time N .

In the Mello–Parsons example, the $b_{n-1,n}$ are all zero and the increments X_n are uncorrelated random variables with mean zero and variance σ^2 . As a result, $qN(a - c)$ is the expected profit from the delivery contract, and the rolling stack locks this in perfectly.² In the Culp–Miller example, the firm hedges to eliminate spot price risk and “play the basis”, meaning maintaining exposure to the $b_{n-1,n}$ (stochastic or not). Again the rolling stack accomplishes this perfectly – but only at the terminal date N . Under either interpretation, it is interesting to examine how far the hedging strategy deviates from its objective (be it locking in expected profits or isolating the basis) before the terminal date N .

² Note, however, that (2)–(4) show that this perfect-lock property of the rolling stack is the result of an algebraic identity that does not rely on stochastic assumptions.

Mello and Parsons (1995b) show that under their assumptions about the price increments the variance of the hedged cumulative cashflow is given by

$$\text{Var}[\bar{C}_k] = \text{Var}[C_k + H_k] = q^2 \sigma^2 (N - k)^2 k;$$

in particular, it is zero at $k = N$. The variance of the unhedged position at k is

$$\text{Var}[C_k] = q^2 \sigma^2 \sum_{i=1}^k i^2.$$

Mello and Parsons (1995b) point out that the hedged variance can therefore be greater than the unhedged one for small k . (Figure 1 graphs continuous versions of the two variances with units chosen so that $q = 1$ and $\sigma = 1$.) While this is certainly suggestive of an increased liquidity risk early in the life of the exposure as a result of the hedge, it is at best a comparison of risks at a fixed time k (if the distributions can reasonably be compared through their variances) but not, without further justification, a comparison of risks *up to* time k . We will argue that comparing *spot risks* as measured by variances at fixed times actually understates the *running risk* of a cash shortfall up to a fixed time.

The derivation leading to (5) relied solely on algebraic identities. A second interpretation of the rolling stack that is useful in more general settings is developed in Appendix B. We show there that any hedging strategy generating cumulative cashflows H_k satisfying

$$H_k - E[H_k] = E[C_k] - E_k[C_N] \quad (6)$$

locks in terminal value. (Here, E_k denotes conditional expectation given the price history to time k .) At intermediate dates, the exposure (actual cash balance minus expected) resulting from a hedge satisfying (6) is

$$\bar{C}_k - E[\bar{C}_k] = C_k - E_k[C_N]; \quad (7)$$

see Appendix B for details. Equation (7) sometimes provides a convenient shortcut.

We now give more detailed model assumptions, generalizing the setting considered so far. For simplicity, we take $q = 1$ from now on. We include mean reversion in the price dynamics to allow for more interesting behavior; specifically, we set

$$S_{n+1} = (1 - \alpha)S_n + \alpha c_n + \sigma Z_{n+1}. \quad (8)$$

Here, $0 \leq \alpha < 1$ measures the speed of mean reversion, c_n is the level toward which the price reverts at time n , and the Z_n are uncorrelated with mean 0 and variance 1. (When $\alpha = 0$ there is no mean reversion.) We express the futures price as

$$F_{n,n+1} = E_n[S_{n+1}] + B_{n,n+1}.$$

Notice that $b_{n,n+1} = B_{n,n+1} + E_n[S_{n+1}] - S_n$, so this change in representation does not by itself entail any assumptions. However, we do assume that the $B_{n,n+1}$ are deterministic.³ This is a shortcoming of our analysis, but one that can be suitably addressed only through a model of commodity prices with at least two factors. Culp and Miller (1995) present evidence that fluctuations in the oil basis are a small fraction of those in spot prices, so our approximation is not without some validity.⁴

By setting $V_n = E[S_n] - S_n$ we can express the unhedged exposure as

$$C_k - E[C_k] = \sum_{n=1}^k (E[S_n] - S_n) = \sum_{n=1}^k V_n, \quad (9)$$

with V_n satisfying

$$V_{n+1} = (1 - \alpha)V_n - \sigma Z_{n+1}.$$

Simple algebra verifies that

$$V_n = \sum_{i=1}^n (1 - \alpha)^{n-i} Z_i$$

and

$$\sum_{n=1}^k V_n = -\sigma \sum_{n=1}^k \frac{1 - (1 - \alpha)^{k-n+1}}{\alpha} Z_n,$$

so an application of (6) (or a derivation akin to that leading to (5)) shows that a perfect terminal hedge is achieved by buying

$$h_n^\alpha = \frac{1 - (1 - \alpha)^{N-n}}{\alpha} \quad (10)$$

one-period futures contracts at time n .⁵ The resulting cumulative hedge cashflows

³ Assuming $B_{n,n+1}$ deterministic can be interpreted as assuming a deterministic risk premium; see Section 6.4 of Duffie (1989) or 7.4.2 of Edwards and Ma (1992). Assuming $b_{n,n+1}$ deterministic rather than $B_{n,n+1}$ would change the number of contracts in a perfect terminal hedge but would not significantly affect our analysis.

⁴ Various notions of basis are commonly used: Culp and Miller (1995), Duffie (1989), and Stoll and Whaley (1993), for example, all give different definitions. The ambiguity in terminology is related to that in the use of the terms “contango” and “backwardation”. See Appendix A. To equate positive and negative basis with contango and backwardation, respectively, using the latter terms in the sense preferred by Duffie (1989) and by Stoll and Whaley (1993), one should take $B_{n,n+1}$ rather than $b_{n,n+1}$ as the basis.

⁵ When $\alpha = 0$, this and all similar expressions should be interpreted in the limit as $\alpha \downarrow 0$. Thus, $h_n^0 = N - n$. In fact, most discussions and assessments of the rolling stack equate the size of the futures position at time n to the remaining commitment, which corresponds to setting $h_n^\alpha = N - n$ in our setting. Our derivation shows that the size of the position should be adjusted to reflect the speed of mean reversion for the rolling stack to be most effective in hedging terminal value. Ross (1995) makes a related observation.

are

$$\begin{aligned} H_k &= \sum_{n=1}^k h_{n-1}^\alpha [S_n - F_{n-1,n}] \\ &= \sum_{n=1}^k h_{n-1}^\alpha \sigma Z_n - \sum_{n=1}^k h_{n-1}^\alpha B_{n-1,n}. \end{aligned}$$

If we set $\bar{C}_k = C_k + H_k$, then from the expressions above for C_k and H_k or more directly via (7), we find that the resulting exposure is

$$\bar{C}_k - E[\bar{C}_k] = -\frac{(1-\alpha) - (1-\alpha)^{N-k+1}}{\alpha} V_k. \quad (11)$$

Thus, we seek to compare the risks in (9) and (11).

We also consider other hedging strategies. A strategy is defined by $g = (g_1, \dots, g_N)$, where g_i denotes the number of futures contracts to buy at time i . The resulting cumulative hedge cashflows are

$$H_k(g) = \sum_{n=1}^k g_n \sigma Z_n - \sum_{n=1}^k g_n B_{n-1,n},$$

leaving an exposure of

$$(C_k + H_k(g)) - E[C_k + H_k(g)] = \sigma \sum_{n=1}^k \left(g_n - \frac{1 - (1-\alpha)^{k-n}}{\alpha} \right) Z_n. \quad (12)$$

For tractability, we work with continuous-time counterparts of the expressions above. Specifically, we replace (8) with

$$dS_t = -\alpha(S_t - c_t) dt + \sigma dW_t \quad (13)$$

with $\alpha \geq 0$, W a standard Wiener process, and c_t a deterministic function of time representing the level towards which the price reverts at time t .⁶ The firm contracts to deliver the commodity continuously at the rate of 1 unit of the commodity per unit of time throughout the interval $[0, T]$. The contracted price is a_t at time t . The cumulative cashflow process is now

$$C_t = \int_0^t (a_s - S_s) ds$$

with an exposure of

$$C_t - E[C_t] = \int_0^t (E[S_s] - S_s) ds = \int_0^t V_s ds,$$

⁶ The continuous-time and discrete-time speeds of mean reversion α_c and α_d are related via $\alpha_d = 1 - \exp(-\alpha_c)$. To lighten notation, we just use α and let context determine whether time is discrete or continuous.

where

$$dV_t = \alpha V_t dt - \sigma dW_s, \quad V_0 = 0.$$

The terminal unhedged exposure is

$$\int_0^T V_s ds = -\sigma \int_0^T \int_0^s e^{-\alpha(s-u)} dW_u ds.$$

Interchanging the order of integration and simplifying shows that this equals

$$-\sigma \int_0^T \frac{1}{\alpha} (1 - e^{-\alpha(T-u)}) dW_u.$$

In this continuous-time setting, we do not model futures explicitly, though it is convenient at times to think of contracts with maturities dt (as in Ross (1995)). We return to real maturities in Section 7. By analogy with (12),

$$\sigma \int_0^t g(s) - \frac{1}{\alpha} (1 - e^{-\alpha(t-s)}) dW_s$$

represents the exposure under the strategy of buying $g(s)$ contracts at time s . In particular, a rolling stack of $(1 - \exp[-\alpha(T-s)])/\alpha$ contracts at time s results in a terminal exposure of zero. We interpret this expression as $(T-s)$ when $\alpha = 0$.

We conclude this section with a remark on *tailing* the hedge – that is, locking in expected *present value*. Discounted at a continuously compounded rate r , the unhedged exposure becomes

$$\int_0^T e^{-rs} V_s ds = -\sigma \int_0^T \left(\frac{e^{-ru} - e^{-(\alpha+r)(T-u)}}{\alpha + r} \right) dW_u.$$

A tailed rolling stack holding

$$\frac{e^{-ru} - e^{-(\alpha+r)(T-u)}}{\alpha + r}$$

futures contracts at time u thus cancels the present value of the unhedged exposure and in so doing locks in the expected present value of the contract. An analogous modification applies in discrete time. Tailing the hedge complicates our analysis without fundamentally affecting it, so for the most part we exclude it from consideration.

3 Spot risk and running risk

For reasons discussed in Section 1, we presume that the firm seeks to hedge expected cashflows from its delivery contract throughout the life of the contract and not just at the terminal date. In particular, we suppose that the firm hedges to try

to prevent the actual cash balance from falling short of the expected cash balance by an amount x , which we take to be large. Write A_t for the actual cash balance at time t under an arbitrary hedging strategy, and say that a *shortfall* occurs when $A_t \leq E[A_t] - x$. Small shortfalls are unlikely to have a significant impact on the firm, so we are primarily interested in large x .

By the *spot risk* at time t we mean

$$P(A_t - E[A_t] < -x),$$

the probability of a shortfall at time t . If, as in our setting, the cash balance is Gaussian, the spot variance $\sigma_t^2 = \text{Var}[A_t]$ measures this risk perfectly. But a more relevant measure is

$$P(\min_{0 \leq s \leq t} (A_s - E[A_s]) < -x), \quad (14)$$

the probability of a shortfall any time up to t , which we call the *running risk* to t .

Calculating the running risk exactly is difficult,⁷ even in our simple model, so we compare risks based on an asymptotic measure that applies for large x . It follows from the Gaussian property of our model that the shortfall probability (hedged or not) can be written as

$$P(\min_{0 \leq s \leq t} (A_s - E[A_s]) < -x) = e^{-\gamma x^2 + o(x^2)}, \quad (15)$$

where

$$\gamma = - \lim_{x \rightarrow \infty} \frac{1}{x^2} \log P(\min_{0 \leq s \leq t} (A_s - E[A_s]) < -x)$$

depends on the hedging strategy and t , and $o(x^2)$ denotes a quantity converging to 0 as $x \rightarrow \infty$, when divided by x^2 . If one hedging strategy has a larger γ than another, it results in smaller probability of a shortfall of magnitude x , for all sufficiently large x . In this sense, a larger γ means less risk.

We use two tools for evaluating γ in particular and the running cashflow risk in general. The first is a remarkable result of Marcus and Shepp (1971)⁸ that, so long as A_t is Gaussian with sample paths that are bounded on bounded intervals (e.g., continuous)

$$\gamma = \frac{1}{2v_t^2}, \quad (16)$$

with

$$v_t = \sup_{0 \leq s \leq t} \sigma_t.$$

⁷ Adler (1990), p. 5, calls this “an almost impossible problem” for general Gaussian processes and notes that (14) is known for very few examples.

⁸ See Adler (1990) for a more extensive treatment and numerous references to related results.

Thus, the running risk is measured by the running maximum standard deviation. If, over some interval $[0, t]$, one hedging strategy has a larger maximum variance than another, then the shortfall probabilities are ordered the same way, for all sufficiently large x . (This is not true without the Gaussian assumption.) In fact, v_t is frequently an even better measure of risk than suggested by (15). If, for example, the supremum defining v_t is attained at a unique point and some additional smoothness conditions are satisfied, then

$$\frac{P(\min_{0 \leq s \leq t} (A_s - E[A_s]) < -x)}{\Phi(-x/v_t)} \rightarrow 1,$$

with Φ denoting the standard normal cumulative distribution. (See Adler (1990), p. 121, quoting a result of Talagrand (1988), and Piterbarg (1996), p. 19; we return to this point in Section 7.) This result states that the probability of a shortfall below level x in $[0, t]$ is well approximated by the probability that a normal random variable lands more than x/v_t standard deviations below its mean.

Our second tool for studying the running risk is the theory of large deviations, which is not restricted to the Gaussian case, and – more importantly in our context – gives more detailed information about when and how a shortfall is likely to occur. The “most likely paths” identified by a large deviations analysis illustrates the types of risks to which different strategies are exposed. In the next three sections, we compare hedged and unhedged positions using $1/\gamma$ as a measure of risk and most likely paths to $-x$ found via large deviations.

4 Without mean reversion

In this section, we specialize to $\alpha = 0$ and compare risks in the unhedged position with risks in a few hedging strategies, including the full hedge that locks in terminal value. We justify the following conclusions:

- (i) A full hedge has greater spot risk than no hedge for approximately 63% $(3(1 - \sqrt{1/3})/2)$ of the life of the exposure.
- (ii) A full hedge has greater running risk than no hedge for approximately 76% $((4/9)^{1/3})$ of the life of the exposure.
- (iii) The optimal fixed fraction to hedge for the full horizon is approximately 63%.
- (iv) The optimal fixed horizon for a full hedge is approximately 73% of the life of the exposure.

Before explaining how we arrive at these observations, we make a few remarks. The crossover point in (i) corresponds to the point at which the two solid curves in Figure 1 cross. In contrast, the point identified in (ii) is where the unhedged variance crosses the dotted line. In view of the discussion in Section 3, we arrive

at the rather surprising conclusion that for any $t < 0.76T$, the probability of a cash shortfall of magnitude x at some time in $[0, t]$ is greater for the hedged position than the unhedged position, for large x . To put (iii) in perspective, notice that in our single-factor model of commodity prices, the minimum-variance hedge ratio would be 1. (For discussions of minimum-variance hedging with futures see Chapter 7 of Duffie (1989) or Chapter 6 of Edwards and Ma (1992).) But the minimum-variance criterion considers the risk at a fixed date only; our measure, which reflects risk throughout the life of the exposure, results in a substantially smaller hedge ratio. Finally, (iv) shows that if one does use a hedge ratio of 1 (as in the standard rolling stack), then the hedging horizon should be shortened to minimize risk.

We now proceed with the verification of (i)–(iv), beginning with some preliminary results. If $\alpha = 0$, then $V_t = -\sigma W_t$. Standard calculations give

$$\sigma_t^2 = \text{Var} \left[\int_0^t V_s ds \right] = \frac{\sigma^2}{3} t^3$$

for the variance of the unhedged exposure. Under a full hedge, the exposure at time t is

$$\int_0^t V_s ds - E_t \left[\int_0^T V_s ds \right] = (T - t) V_t.$$

Thus, under a full hedge we have a spot variance of

$$\bar{\sigma}_t^2 = (T - t)^2 \sigma^2 t.$$

As discussed in Section 2, a deterministic hedging strategy is a function g on $[0, T]$, with $g(s)$ interpreted as the number of futures contracts to hold at time s . In the absence of mean reversion, full hedging corresponds to $g(s) = (T - s)$ and no hedging corresponds to $g(s) \equiv 0$. The exposure under any strategy g is (using integration by parts for the first integral)

$$\int_0^t V_s ds + \sigma \int_0^t g(s) dW_s = \sigma \int_0^t [s - t + g(s)] dW_s,$$

which has variance

$$\sigma_t^2(g) = \sigma^2 \int_0^t [s - t + g(s)]^2 ds. \quad (17)$$

We use this repeatedly to compare the risks in different strategies.⁹

For (i) we set $\sigma_t^2 = \bar{\sigma}_t^2$ and solve to get $t = (3T/2)(1 - \sqrt{1/3})$. For (ii), we first note that the spot variance of the full hedge is maximized at $T/3$, where it takes the value $4\sigma^2/27$. The running variance of the full hedge thus remains at this level in

⁹ The problem of minimising (over g) the maximum (over t) of (17) has been given a fascinating solution by Larcher and Leobacher (2000).

the interval $[T/3, T]$. For the unhedged position, the running and spot variance are equal (the spot variance increases monotonically); hence, the unhedged position becomes less risky than the full hedge when

$$\frac{\sigma^2}{3}t^3 = \frac{4\sigma^2}{27}T^3,$$

i.e., at $t = (4/9)^{1/3}T$.

We next consider (iv). Recall that a full hedge makes the spot risk at T zero. By hedging to a horizon $\tau \leq T$, we mean hedging to make the spot risk at τ zero (and remaining unhedged in $[\tau, T]$). This is achieved by holding $(\tau - s)$ futures contracts at time s , rather than $(T - s)$; i.e., by the strategy

$$g_\tau(s) = \begin{cases} (\tau - s), & 0 \leq s \leq \tau; \\ 0, & s > \tau. \end{cases}$$

The optimal fixed-horizon hedge is the one that minimizes the running risk over the entire interval $[0, T]$. For any τ , we can evaluate the spot variance under g_τ using (17). The maximal spot risk occurs either at $\tau/3$ (where the hedged portion is riskiest) or at T (where the unhedged portion is riskiest). Using (17), we find that the spot variances at these times are $4\sigma^2\tau^3/27$ and

$$\sigma^2 \int_0^\tau (T - \tau)^2 ds + \sigma^2 \int_\tau^T (T - s)^2 ds = \sigma^2 \left(\frac{2}{3}\tau^3 - T\tau^2 + \frac{1}{3}T^3 \right),$$

respectively. The optimal τ – the one that minimizes the running risk – makes the spot variances at these times equal. This is the root of a cubic equation which can, in principle, be given explicitly; numerically, we find $\tau \approx 0.733T$ as indicated in (iv). Figure 2 displays the resulting variance over the life of the exposure along with that for a full hedge – i.e., with a hedging horizon of T .

We now turn to (iii). Fully hedging a fixed fraction π throughout $[0, T]$ corresponds to the strategy $g_\pi(s) = \pi(T - s)$ and therefore results in a spot variance of

$$\sigma^2 \int_0^t (\pi T + (1 - \pi)s - t)^2 ds.$$

This is evidently a cubic function of t ; it achieves a local maximum at

$$t^* = \frac{\pi T(1 + \pi - \sqrt{\pi})}{\pi^2 + \pi + 1}.$$

The other possible location of the maximal variance is T , where the spot variance is $(1 - \pi)^2\sigma^2 T$. The optimal π sets the values of the spot variance at t^* and T equal. Numerically, we find that the optimal π is 0.62996, which appears to coincide with $(1/4)^{1/3}$. The resulting variance over time is graphed in Figure 2. Both the

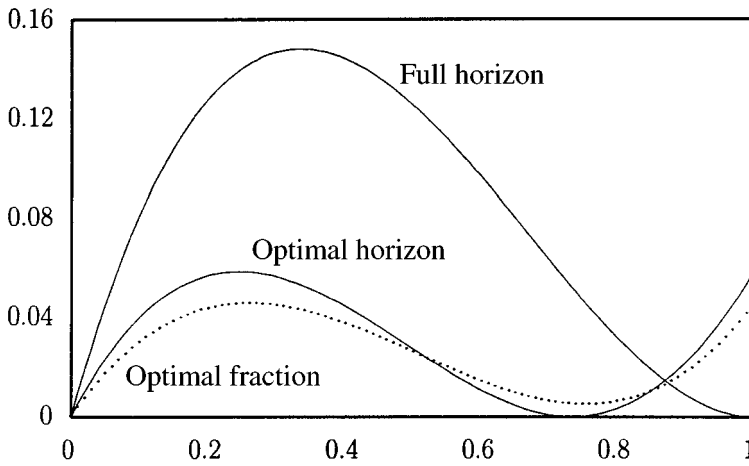


Fig. 2. Comparison of variances under different hedging strategies. The full hedge uses a hedge ratio of 1 for the full horizon T . The optimal fixed-horizon hedge uses a hedge ratio of 1 until time $\tau \approx 0.733T$ and thus balances the risk from the hedge early in the interval with the original risk later in the interval. The optimal fraction hedge uses a hedge ratio of $\pi \approx 0.63$ for the full interval $[0, T]$.

optimal hedge ratio and the optimal fixed horizon result in substantial reduction in the running risk, compared to a full stacked hedge. Hedging the optimal fixed fraction is slightly more effective than hedging fully for the optimal horizon.

We conclude this section with some observations on the impact of tailing the hedge, as described at the end of Section 2. Table 1 shows the location and value of the maximum variance with a full hedge and with no hedging, for various values of the discount rate r . The results indicate little change over a broad range of rates. Indeed, although maximum variances decrease with r (as they should), their ratio remains essentially unchanged.

5 With mean reversion

The possibility of mean reversion introduces more varied behavior in the dynamics of commodity prices and in the hedged and unhedged exposures. If we take $c_t \equiv c$ in (13), then expected future prices satisfy

$$E_t[S_{t+s}] = e^{-\alpha s} S_t + (1 - e^{-\alpha s})c.$$

A graph of expected future prices is thus upward sloping, flat, or downward sloping depending on whether S_t is below, at, or above c , and bears some resemblance to graphs in Figure 3 of Brennan and Crew (1997), Figure 8 of Edwards and Canter

Table 1. *The effect of tailing the hedge using a range of discount rates.*

Rate	Hedged		Unhedged		Ratio
	Location	Maximum	Location	Maximum	
0	0.333	0.148	1	0.333	44.4%
0.01	0.333	0.146	1	0.329	44.4%
0.05	0.330	0.139	1	0.313	44.3%
0.10	0.326	0.130	1	0.294	44.1%
0.15	0.322	0.121	1	0.277	43.9%
0.20	0.319	0.114	1	0.260	43.7%

The columns labeled “Location” and “Maximum” give the time at which the maximal variance is attained (as a fraction of T) and the magnitude of the maximal variance (as a fraction of $\sigma^2 T^3$). The last column gives the ratio of the maximal variances of the hedged and unhedged positions.

(1995), and Figure 1 of Neuberger (1999) showing the term structure of oil prices at various points in time.

The presence of mean reversion has important implications for hedging. If commodity prices are mean reverting, an exposure to them has a type of built-in hedge: unusually large price movements in the short term will be naturally offset over time. To lock in expected terminal profits, less hedging should be required with a greater speed α of mean reversion.

For the most part, our observations in this section depend on the magnitude of α . In thinking about what values of α are plausible, it is convenient to view $1/\alpha$ as the expected time for prices to revert about two-thirds of the way to their mean. (Data in Bessembinder *et al.* (1995) suggests $\alpha \approx 0.77$ for oil prices, with time measured in years.) In particular, α depends on the unit of time, so we state our conclusions in terms of the dimensionless quantity αT . This is equivalent to measuring time in multiples of the horizon T . The expressions we obtain for $\alpha > 0$ are more complicated than those we obtained for $\alpha = 0$ in the previous section; as a consequence, our results are somewhat less explicit. Through a combination of exact and numerical results, we make the following observations:

- (i') The spot risk of the fully hedged position is maximized at $T/3$, regardless of the rate of mean reversion.
- (ii') Unless αT is greater than about 2.375, a full hedge has greater running risk than no hedge for most of the life of the exposure. For the spot risk, the cut off is $\alpha T \approx 2.06$.
- (iii') The optimal fixed fraction to hedge for the full horizon is approximately 63–75%.

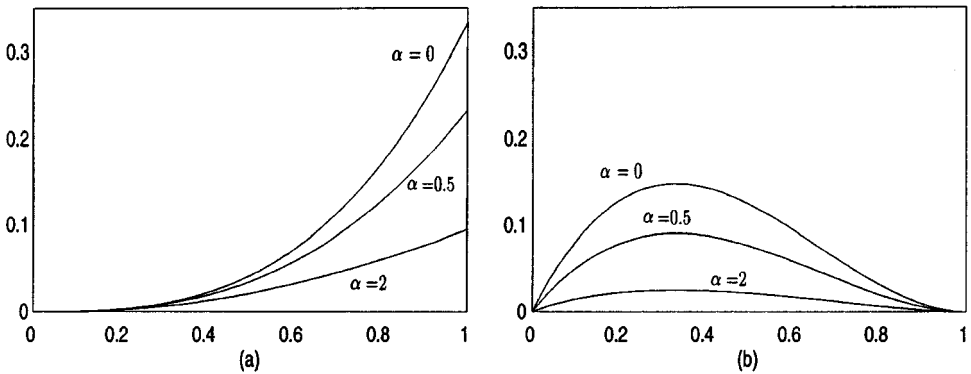


Fig. 3. Variance of (a) unhedged and (b) hedged cash balance over time for three values of the mean-reversion speed α .

(iv') The optimal fixed horizon for a full hedge is approximately 72–78% of the life of the exposure.

A useful result for the case $\alpha > 0$ is

$$\text{Cov}[V_s, V_t] = E[V_s V_t] = \frac{1}{2\alpha} e^{-\alpha(t+s)} (e^{2\alpha s} - 1), \quad s < t;$$

see, e.g., p. 358 of Karatzas and Shreve (1991). From this we can calculate the spot risk of the unhedged exposure to be

$$\begin{aligned} \sigma_t^2 = \text{Var} \left[\int_0^t V_s ds \right] &= 2 \int_0^t \int_0^s E[V_u V_s] du ds \\ &= \frac{\sigma^2}{\alpha^3} \left[\alpha t + 2(e^{-\alpha t} - 1) - \frac{1}{2}(e^{-2\alpha t} - 1) \right]. \end{aligned} \quad (18)$$

The fully hedged position has an exposure of (see (7))

$$\int_0^t V_s ds - E_t \left[\int_0^T V_s ds \right] = -\frac{1}{\alpha} V_t (1 - e^{-\alpha(T-t)}) \quad (19)$$

and a spot risk of

$$\bar{\sigma}_t^2 = \text{Var} \left[\frac{1}{\alpha} V_t (1 - e^{-\alpha(T-t)}) \right] = \frac{\sigma^2}{2\alpha^3} (1 - e^{-\alpha(T-t)})^2 (1 - e^{-2\alpha t}).$$

Some tedious but straightforward calculus shows that $\bar{\sigma}_t^2$ is maximized at $T/3$, as indicated in (i'); in particular, the location of the maximum is independent of α . For the unhedged position, σ_t^2 is, of course, always maximized at T . Figure 3 illustrates the dependence on α . With larger α there is less risk and the full hedge

Table 2. *Crossover points as a fraction of the life T of the exposure.*

Reversion rate αT	Spot risk crossover	Running risk crossover
0	0.63	0.76
0.10	0.63	0.75
0.5	0.60	0.71
1	0.57	0.65
2	0.50	0.53
5	0.31	0.31
10	0.16	0.16
100	0.02	0.02

is more effective in reducing what risk there is. Both properties reflect the natural hedge resulting from mean reversion.

To justify (ii'), we located the points $t > 0$ at which $\sigma_t^2 = \bar{\sigma}_t^2$ and $\max_{s \leq t} \sigma_s^2 = \max_{s \leq t} \bar{\sigma}_s^2$, respectively. These crossover points are displayed in Table 2 for a range of α values. The crossover points occur more than halfway through the life of the horizon until αT exceeds 2.06 for the spot risk and 2.375 for the running risk. For larger values of α , σ_t^2 crosses $\bar{\sigma}_t^2$ before $T/3$; because $\bar{\sigma}_t^2$ increases in $[0, T/3)$, the two crossover points in Table 2 are the same for larger α .

For an arbitrary hedging strategy g , the spot variance is

$$\sigma_t^2(g) = \sigma^2 \int_0^t \left[g(s) - \frac{1}{\alpha} (1 - e^{-\alpha(t-s)}) \right]^2 ds \quad (20)$$

which reduces to the expression in (17) as $\alpha \downarrow 0$. For each $\tau \in [0, T]$, the partial-horizon strategy g_τ given by

$$g_\tau(s) = \begin{cases} \frac{1}{\alpha} (1 - \exp(-\alpha(\tau - s))), & 0 \leq s \leq \tau; \\ 0, & \tau < s \leq T \end{cases}$$

makes the spot variance 0 at τ . The maximum spot variance under g_τ occurs at either $\tau/3$ or T ; the spot variances at these points are

$$\frac{\sigma^2}{2\alpha^3} \left(1 - e^{-\frac{2\alpha\tau}{3}} \right)^3 \quad (21)$$

and

$$\frac{\sigma^2}{\alpha^3} \left[-\frac{1}{2} (e^{-\alpha\tau} - e^{-\alpha T})^2 + e^{-\alpha(T-\tau)} - 1 + \alpha(T - \tau) \right], \quad (22)$$

Table 3. *Optimal fixed hedging horizons (as a fraction of T) and fixed hedge ratios.*

Reversion rate αT	Optimal fixed horizon	Optimal fixed fraction
0	0.733	0.630
0.10	0.732	0.633
0.5	0.727	0.647
1	0.724	0.665
2	0.728	0.697
5	0.790	0.770
10	0.881	0.857
100	0.994	0.989

respectively. The optimal τ – the one that minimizes the maximum spot variance – makes these two expressions equal. Numerical values are summarized in Table 3. The optimal horizon is rather insensitive to α . This is due, in part, to the fact that it first decreases and then increases as α increases away from zero. This lack of monotonicity arises from the fact that, as α increases, both (21) and (22) decrease, but neither consistently faster than the other.

Using (20), we can find the optimal fixed-fraction hedge for each α . Fully hedging a fraction π throughout the life of the exposure corresponds to the strategy

$$g_{\pi}(s) = \frac{\pi}{\alpha} (1 - e^{-\alpha(T-s)}).$$

Substituting this strategy in (20) yields a tractable but cumbersome expression which we suppress. We use this expression to find the hedge ratio π that minimizes the maximum variance over the life of the exposure. The results appear in the third column of Table 3. For plausible speeds of mean reversion, the hedge ratio that minimizes the running risk is in the range of 63–75%, even though the minimum-variance hedge ratio in our model is always 1.

6 Most likely paths

In this section, we examine in more detail the scenarios that lead to cash shortfalls with and without a stacked hedge. We begin by considering the case $\alpha = 0$, in which the exposure V_s is just a Wiener process. An event $\mathcal{A}_x$ like “a shortfall of magnitude greater than x occurs in $[0, T]$ ” is a set of sample paths of the Wiener process. There is often a path in a set like $\mathcal{A}_x$ that is the most likely path in the sense that when $\mathcal{A}_x$ occurs, it occurs with the Wiener process staying close to this

path. This tendency to follow the most likely path becomes most pronounced as the event becomes rare, which corresponds to x becoming large in our setting. These statements are made precise by the theory of large deviations; see Dembo and Zeitouni (1998) and Stroock (1984) for background. This is a highly technical topic, so we will keep our discussion informal and proceed as directly as possible to the calculation of most likely paths.

We noted in Section 3 that the limit

$$\lim_{x \rightarrow \infty} \frac{1}{x^2} \log P(\mathcal{A}_x) = -\gamma$$

gives the exponential rate of decrease of $P(\mathcal{A}_x)$ in x^2 . The most likely path $\phi^* \in \mathcal{A}_x$ has the following property: if we define a strip around ϕ^* of width ϵ , then the probability that the Wiener process stays within this strip throughout $[0, T]$ decays at an exponential rate nearly equal to that of $P(\mathcal{A}_x)$, the difference vanishing as $\epsilon \downarrow 0$. Moreover, the probability that the Wiener process leaves this strip *conditional* on $\mathcal{A}_x$ occurring vanishes exponentially as x increases. Thus, given that $\mathcal{A}_x$ occurs, with high probability it occurs by the Wiener process staying close to the most likely path.

Finding the most likely path is a problem in the calculus of variations. For any absolutely continuous function ϕ on $[0, T]$, denote by $\dot{\phi}$ its derivative with respect to time. The most likely path in $\mathcal{A}_x$ solves

$$\text{minimize}_{\phi \in \mathcal{A}_x} \quad \frac{1}{2} \int_0^T [\dot{\phi}(t)]^2 dt. \quad (23)$$

This is known as Schilder's Theorem; see Dembo and Zeitouni (1998) or Stroock (1984) (especially pp. 66–7 for the mean-reverting case). Membership in $\mathcal{A}_x$ defines a constraint on ϕ . Still with $\alpha = 0$, for the unhedged exposure

$$\mathcal{A}_x = \left\{ \phi : \sigma \int_0^t \phi(s) ds > x, \text{ for some } t \in [0, T] \right\},$$

since this defines a cash shortfall in this setting. (In this and all subsequent cases, the requirement $\phi(0) = 0$ is implicit.) In the fully hedged case, a shortfall occurs when $(T - t)V_t > x$, so (recalling that $V_t = -\sigma W_t$)

$$\mathcal{A}_x = \{ \phi : \sigma \phi(t) < -x/(T - t) \text{ for some } t \in [0, T] \}.$$

The solutions to (23) in these two cases are displayed in Figure 4a, b; the derivations are given in Appendix C. In each case, if ϕ^* is the minimizing path, then

$$\gamma = \frac{1}{2} \int_0^T [\dot{\phi}^*(t)]^2 dt,$$

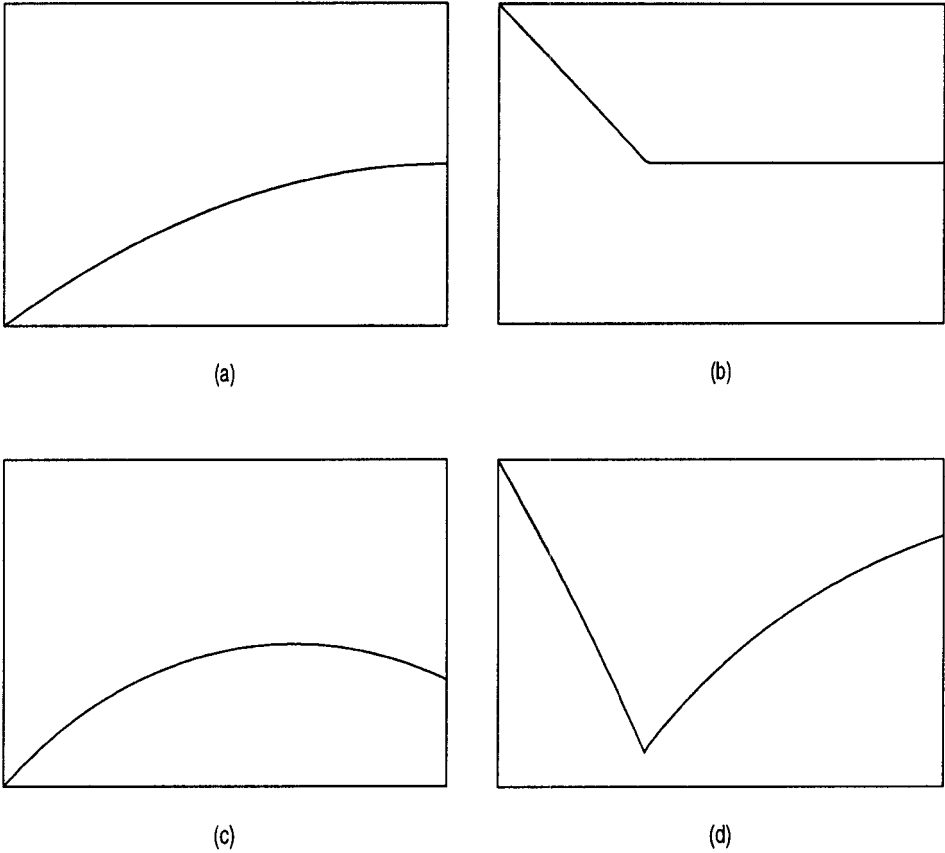


Fig. 4. Most likely paths of $S_t - E[S_t]$ to a cash shortfall. (a) and (b) are with $\alpha = 0$, (c) and (d) with $\alpha = 2$. (a) and (c) are for unhedged exposures, (b) and (d) are for fully hedged exposures.

with γ as defined in (15). In other words, the exponential rate of decrease of the shortfall probability is the also the “cost” of the minimum-cost path to a shortfall.

We now consider the case $\alpha > 0$. In light of the relation

$$V_t = -\sigma \int_0^t e^{-\alpha(t-s)} dW_s,$$

any event defined in terms of V can be expressed through conditions on W . More specifically, to each path ψ of V there corresponds a path ϕ of W via

$$\psi(t) = -\sigma \int_0^t e^{-\alpha(t-s)} \dot{\phi}(s) ds;$$

i.e.,

$$\dot{\psi}(t) = -\alpha\psi(t) - \sigma\dot{\phi}(s)$$

and therefore

$$\dot{\phi}(t) = -\frac{1}{\sigma}[\dot{\psi}(t) + \alpha\psi(t)]^2. \quad (24)$$

If we now let $\mathcal{A}_x$ be the set of ψ paths resulting in a shortfall of magnitude greater than x , then substituting (24) in (23) we arrive at the objective

$$\text{minimize}_{\psi \in \mathcal{A}_x} \quad \frac{1}{2\sigma^2} \int_0^T [\dot{\psi}(t) + \alpha\psi(t)]^2 dt \quad (25)$$

to determine the most likely path. In the unhedged case the constraint is

$$\mathcal{A}_x = \left\{ \psi : \int_0^t \psi(s) ds > x, \text{ for some } t \in [0, T] \right\},$$

whereas in the hedged case it is (see (19))

$$\mathcal{A}_x = \{\psi : \sigma\psi(t) < -\alpha x / (1 - \exp[-\alpha(T-t)]) \text{ for some } t \in [0, T]\}.$$

In each of these problems, x merely serves to scale the solution: the solution for arbitrary x is just x times the solution for $x = 1$; hence, it suffices to give the solution for $x = 1$. The volatility parameter σ is also a scale parameter and may therefore be set to 1 as well. With these simplifications, we present the solutions to the problems above:

- $\alpha = 0$, unhedged:

$$\phi(t) = \frac{3}{T^2}t - \frac{3}{2T^2}t^2;$$

- $\alpha = 0$, hedged:

$$\phi(t) = \begin{cases} -(9/2T^2)t, & 0 \leq t \leq T/3; \\ -3/2T, & T/3 < t \leq T. \end{cases}$$

- $\alpha > 0$, unhedged:

$$\psi(t) = ae^{\alpha t} + be^{-\alpha t} + c,$$

where $a = \alpha / ((3 - 2\alpha T) \exp(\alpha T) + \exp(-\alpha T) - 4)$, $b = (2 \exp(\alpha T) - 1)a$ and $c = -(a + b)$.

- $\alpha > 0$, hedged:

$$\psi(t) = \begin{cases} -2c_1 \sinh(t) & 0 \leq t \leq T/3; \\ -c_2 e^{-\alpha t} & T/3 < t \leq T, \end{cases}$$

with $c_1 = \alpha \exp(-\alpha T/3)(1 - \exp(-2\alpha T/3))^{-2}$ and $c_2 = (\exp(2\alpha T/3) - 1)c_1$.

These paths are graphed in Figure 4(a)–(d), the last two with $\alpha = 2$. The graphs are all on the same (dimensionless) scale, but with the origin in the upper-left corner of (b) and (d) and the lower-left corner of (a) and (c). In each case, the curve shows the most likely path by which the commodity price S_t deviates from the expected price $E[S_t]$ in generating a cash shortfall. Appropriately, in the unhedged cases (a) and (c) the shortfall results from an unexpected price increase and in the hedged cases (b) and (d) it results from an unexpected decrease: the rolling stack creates a large long position in the commodity early in the life of the exposure. In (a), the price increases throughout the life the exposure, leveling off at the end, where the optimal path has derivative zero. With mean reversion, (c) shows that the most likely scenario has the price deviation reaching a maximum before T ; the curvature of the path increases with α . The graphs in (b) and (d) show the rather different risks to which the firm is most exposed under a full hedge. In both cases, there is a sharp drop in price until $T/3$ where the shortfall occurs. In (b), the price then stays flat, whereas in (d) it reverts towards its mean. Indeed, after $T/3$, the paths in (b) and (d) are unconstrained by the corresponding event $\mathcal{A}_x$, so the paths follow their mean behavior; the most likely paths are interesting only up to $T/3$ in these cases. Figure (d) is reminiscent of the sharp drop followed by a gradual recovery in the price of oil around the time of MGRM's crisis.

7 Assessing the approximations

The analysis in Sections 3–6 relied on two approximations to the model initially developed in Section 2: we replaced the discrete-time model with a continuous-time one, and we replaced the exact (unknown) risk of a cash shortfall with the running maximum variance, which is valid when the magnitude x of the shortfall is large. In this section, we examine the validity of these approximations.

We begin with a closer look at approximations based on (15) and the surrounding discussion, still in continuous time. It follows from Theorem D.3 of Piterbarg (1996) that for the unhedged exposure

$$C_t - E[C_t] = \int_0^t V_s ds,$$

the shortfall probability satisfies

$$\lim_{x \rightarrow \infty} \frac{P(\min_{0 \leq s \leq t} \{C_s - E[C_s]\} < -x)}{\Phi(-x/v_t)} = 1, \quad (26)$$

for each $t \in (0, T]$, indicating that the running maximum standard deviation v_t is an even better measure of the running risk than suggested by (15) and (16), in the

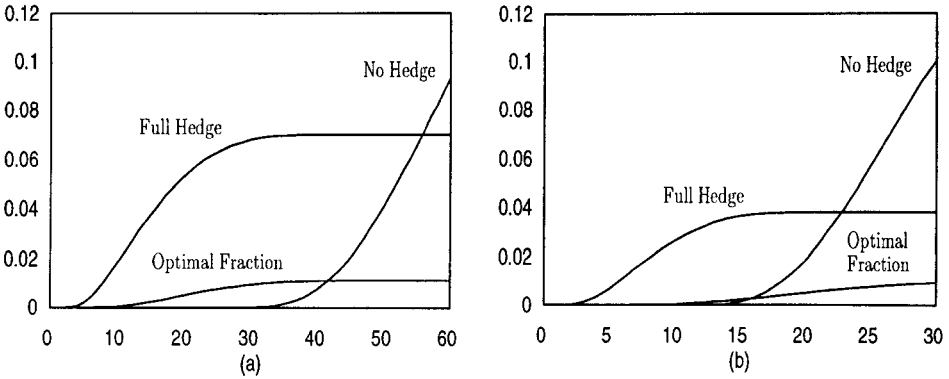


Fig. 5. Cumulative probability over time of a cash shortfall, estimated by simulation. In (a), $\alpha = 0$ and the horizon is 60 periods; in (b), $\alpha T = 2$ and the horizon is 30 periods.

unhedged case.¹⁰ In the hedged case, with an exposure of

$$\bar{C}_t - E[\bar{C}_t] = -\frac{1}{\alpha} V_t (1 - e^{-\alpha(T-t)}),$$

Theorem D.4 of Piterbarg (1996) gives

$$\Phi(-x/v_t) \leq P(\min_{0 \leq s \leq t} \{\bar{C}_s - E[\bar{C}_s]\} < -x) \leq \text{constant} \cdot x \Phi(-x/v_t), \quad (27)$$

but not the analog of (26). This suggests that the running maximum variance may underestimate the risk of the hedge, relative to no hedge, when x is not too large.

To assess the reliability of risk comparisons based on the running maximum variance, we conducted simulation experiments to estimate shortfall probabilities directly for the discrete-time model. The graphs in Figure 5 are indicative of a large number of experiments with different parameter values. The curves in the graphs show estimated cumulative probabilities of a shortfall over time with no hedge, a full hedge, and the optimal hedge ratio from Sections 4 and 5. The graphs in (a) are based on 60 periods (intended to suggest a five-year exposure hedged with one-month contracts) and $\alpha = 0$, those in (b) use 30 periods and $\alpha T = 2$. The magnitude of the shortfall was chosen to get a cumulative probability of roughly 10%. The overall appearance of the graphs is strikingly similar to the comparison of the running maximum variances in Figure 1. Indeed, the simulation results suggest that Figure 1 even understates the risk of a full hedge, consistent with the comments following (27). The general pattern we have observed based on these and other simulation results is that the riskiness of the full hedge (relative to no

¹⁰ Piterbarg formulates his result in the case that the point of maximal variance is in the interior of the time interval over which the maximum is computed, but then notes that the result extends to the case in which the maximum is attained at the boundary, as in our setting.

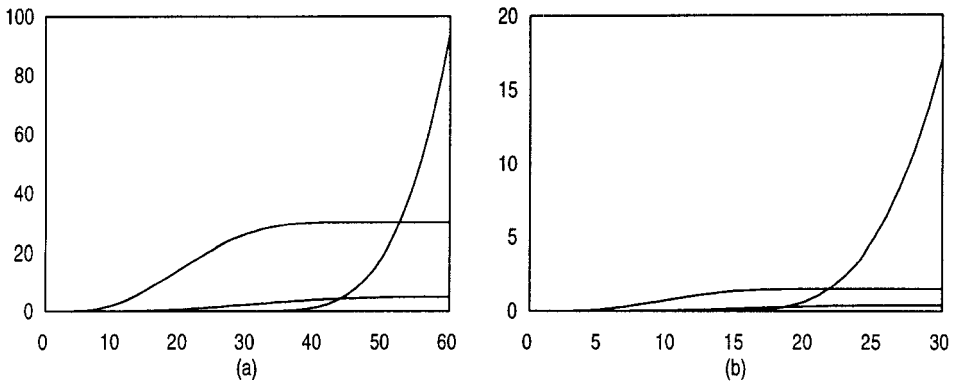


Fig. 6. Cumulative expected cash shortfall with no hedge, a full hedge, and the optimal-fraction hedge. (a) and (b) are based on the same parameters as in Figure 5. As before, the curves are ordered with the optimal-fraction hedge having smallest cumulative risk, the full hedge in the middle, and no hedge having the largest cumulative risk.

hedge) decreases with the magnitude of the shortfall and with the speed of mean reversion.

Figure 5 also indicates that substantial risk reduction can be achieved by using the optimal fixed-fraction hedge rather than a hedge ratio of 1. It should be possible to get further risk reduction for any number of periods N by solving numerically for the strategy $(g_1, \dots, g_N)$ that minimizes the maximum variance over the hedging horizon. This is an easily solved optimization problem; we have found that the resulting strategy is surprisingly erratic and does not appear to lend itself to simple specification. Of course, even this strategy is at best the optimal deterministic strategy; in practice, a firm is likely to adjust its hedge in light of new price information.

The shortfall probability is open to criticism as a measure of risk because it treats all shortfalls of magnitude greater than x equally. A simple alternative weights shortfalls in proportion to the amount by which their magnitudes exceed x . Let ϵ_n denote the exposure at the end of period n , hedged or not. By the expected cumulative shortfall to time k we mean

$$\sum_{n=1}^k E[\max(0, -x - \epsilon_n)].$$

Artzner *et al.* (1996) have developed an axiomatic approach to risk measures in which the only “coherent” measures of risk are generalizations of this expression with $x = 0$.

Figure 6 shows cumulative expected shortfalls estimated through simulation with a full hedge, no hedge, and the optimal fixed-fraction hedge. The parameters

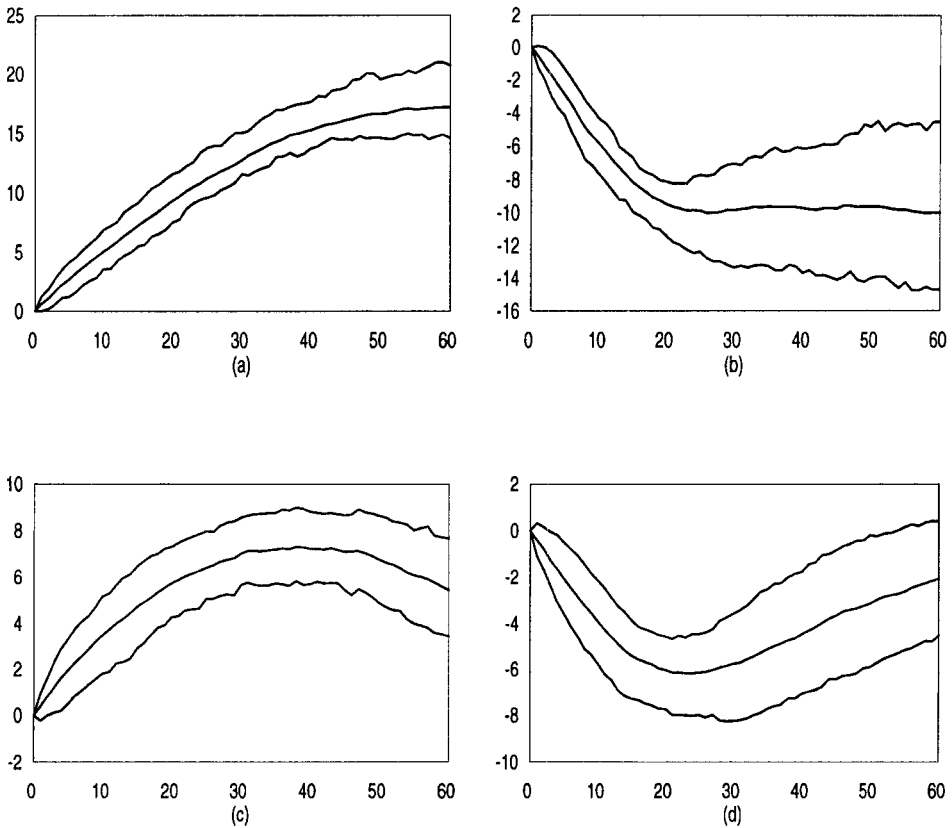


Fig. 7. Simulated paths on which a shortfall occurs. In each case, the center path is the average over all simulated paths on which a shortfall occurs, and the band around the center path shows the interquartile range. (a) and (b) are for $\alpha = 0$, (c) and (d) for $\alpha = 2$.

are exactly as in Figure 5. Again, the overall behavior of the risks is strikingly similar to that in Figure 1. The similarity is even more notable given that the motivation in Section 3 focused exclusively on the shortfall probability. These results suggest that the running maximum variance is a reasonably robust measure of risk.

We next turn to the most likely paths found in Section 6. That analysis was also based on continuous time and large x . To determine whether the paths found there are relevant to the original setting, we again simulated the original discrete-time model, with and without mean reversion, with and without hedging. For each case, we simulated roughly 20 000 paths, and saved those on which a shortfall occurred. The magnitudes of the required shortfalls were varied for different cases to keep the probability of a shortfall in the range of 2–5%. The saved paths approximate the conditional law of the exposure process given a shortfall. In Figure 7 we

have graphed the mean and the 25th and 75th percentiles (computed separately for each time period) of the paths. These show good qualitative agreement with the theoretical paths in Figure 4. As explained in Section 6, the paths in (b) and (d) are constrained only up until a shortfall occurs (near one-third of the horizon), so only this portion of the path is interesting. After the first third of the horizon, the spread in (b) reflects the ordinary $\sqrt{n}$ diffusion associated with a random walk. Indeed, the contrast in (b) before and after the first third shows the extent to which the occurrence of a shortfall alters the usual evolution of the path.

8 Concluding remarks

We have proposed a measure of liquidity risk that approximates the probability of a cash shortfall any time in the life of an exposure, and used it to compare the risks in various strategies for a firm hedging long-term commodity contracts with short-dated futures. The implications of our analysis include an assessment of the cashflow risks produced by a seemingly perfect terminal hedge of the type used by Metallgesellschaft. We have also identified the particular price patterns to which a hedged or unhedged firm is most exposed, and examined the impact of mean reversion in the spot price.

Although we focused on a rather specific context, our analysis is relevant to other settings in which the variance of a position may fail to be monotone over time. Swaps, for example, typically have this property, and, like the fully hedged position in our context, have zero terminal variance. Indeed, our basic setup applies to the cumulative payments on a floating-for-fixed interest rate swap with the floating rate described by the Vasicek (1977) model. Hedging strategies based on discrete rebalancing can also be expected to have nonmonotone variance. The current and growing emphasis – in the finance industry, among regulators, and even in corporate finance – on measuring value-at-risk over multiple horizons suggests broader potential application for the perspective developed here.

Acknowledgements.

I thank Frank Edwards for discussions that motivated this work and Suresh Sundaresan for helpful discussions and detailed comments. For additional comments and helpful discussions I thank Sid Browne, John Parsons, Larry Shepp, and Tim Zajic.

Appendix A: Futures and forwards

This section gives a brief summary of some concepts and terminology pertinent to futures and forward contracts. More thorough treatments of these topics are given

in, for example, Duffie (1989), Edwards and Ma (1992), and Stoll and Whaley (1993).

A forward contract is an agreement between two parties to make a transaction at a fixed price and date in the future. The long party commits to buying a specified quantity of, e.g., a commodity or financial asset from the short party at a specified delivery price. The forward price is the delivery price that makes the value of the contract zero. If a forward contract specifies the current forward price at the time of the agreement as the delivery price (the typical case), then the parties enter the agreement with no exchange of payments. At later dates, the forward price may change whereas the contractual delivery price will not. If the forward price rises, the forward contract – worth zero at inception – will take on positive value for the long party and negative value for the short party. Conversely, if the forward price drops, the value of the forward contract becomes positive for the short party and negative for the long party.

A futures contract is similarly a commitment to execute a sale at a specified price and date in the future; the futures price is the delivery price that makes entry into a futures contract costless. Whereas forward contracts are arranged directly between the parties involved, futures contracts are traded through exchanges. This distinction has many implications for the design of the contracts and hence for hedging strategies that use them. Forward contracts can be highly customized, specifying the precise quantity, grade, delivery date and delivery location that suits the parties involved. In contrast, futures contracts must be standardized for exchange trading and yet meet the needs of many market participants; they thus admit a relatively small number of maturities, fixed quantities, flexibility in the timing of delivery and the precise underlying grade or asset to be delivered.

The most important distinction for the purposes of this article is that futures contracts are marked-to-market and forward contracts are not. With a forward contract, no payments are made at the inception of a contract and no payments are made subsequently until the contract matures, at which time the two parties execute the agreed-upon transaction. A party entering into a futures contract neither makes nor receives a payment upon entry, but on each subsequent day the exchange will credit the party for any profits and charge the party for any losses on its position. These transactions are made through a margin account, the precise mechanics of which can be somewhat involved. A simple example should nevertheless serve to illustrate the key point.

Consider a futures or forward contract maturing in three days and suppose the current futures or forward price is 100. Suppose that over the next three days the futures or forward price fluctuates to 98, 101, and then 103. At the end of the third day, the contract matures and thus reduces to a commitment to buy immediately rather than at some point in the future. Accordingly, 103 must be the spot price

(the price for immediate purchase) at the end of the third day. Consider the case of a forward contract: the contract specifies a delivery price of 100 though the spot price is 103, so the long party can buy at 100 and then sell at 103 for a profit of 3 at the end of the third day. In the case of a futures contract, at the end of the first day the exchange would require a payment of 2 from the long party, reflecting the drop in the futures price to 98. At the end of the next day, the exchange would credit the long party 3, reflecting the increase to 101, and on the next day the exchange would make a further payment of 2. The long party could close its position without taking physical delivery of the underlying, earning a profit of $-2+3+2 = 3$. Thus, in this example, the final profit resulting from the two contracts is the same, but the futures contract entails intermediate cashflows whereas the forward contract does not. It is precisely this distinction that gives rise to the possibility of a cash shortfall in offsetting a short forward position with a long futures position. It should be noted that this distinction in the timing of cashflows also leads to the conclusion that futures prices and forward prices will not generally be equal (as they are in the example) if interest rates are correlated with the underlying asset, though we will not address that issue here.

We briefly consider the relation between futures prices and the price of the underlying asset or commodity. Fix a date T and let F_t denote the time- t futures price for a contract maturing at T . Let S_t denote the price of the underlying at time t . Under simplifying assumptions (including costless transactions and unlimited short-selling) the futures and spot price are related via $F_t = S_t e^{c(T-t)}$, where c is the *cost of carry*. The cost of carry could be positive or negative and reflects both costs and benefits associated with holding the underlying, such as financing and storage costs and any dividends paid by the underlying. In a world with a deterministic cost of carry, changes in the futures price are perfectly correlated with changes in the spot price, so the risk in one can be eliminated through trading in the other.

The term *basis* refers broadly to differences between futures and spot prices. The relevant spot price may not be precisely the one underlying the futures contracts. For example, hedging an exposure to the price of jet fuel with futures contracts on heating oil is said to entail *basis risk* due to imperfect correlation between the futures price of heating oil and the spot price of jet fuel. The simplest definitions of basis take it to be $S_t - F_t$ or $F_t - S_t$ (consistent with $b_{n,n+1}$ in Section 2), but other definitions are used as well. Duffie (1989), for example, defines the basis to be $F_T - S_T$ even at time $t < T$. This difference would generally be nonzero (but unknown) if, e.g., S_t is the price of jet fuel and F_t is the futures price for heating oil.

A related ambiguity concerns the terms *backwardation* and *contango*. Broadly speaking, these describe conditions in which futures prices are, respectively, lower

than or higher than spot prices. According to the interesting discussion in Section 4.3 of Duffie (1989), modern usage associates these terms with the conditions $E_t[S_T] > F_t$ and $E_t[S_T] < F_t$ respectively. An advantage of defining these terms through the older conditions $S_t > F_t$ and $S_t < F_t$ is that it becomes possible to observe whether in fact a futures market is in backwardation or contango. With this definition, the oil market and many other commodity markets are more often in backwardation than contango.

Appendix B: The rolling stack and conditional expectations

In this appendix, we argue that (6) and (7) are the key properties underlying the perfect terminal hedging property of the rolling stack.

Consider, again, the setting leading to (3) and (4). Suppose the X_n have mean zero and the $b_{n-1,n}$ are all zero, as in the Mello–Parsons setting, and compute the conditional expectation of the terminal value of the unhedged position, given the price history to time k :

$$\begin{aligned}
 E_k[C_N] &= E_k \left[\sum_{n=1}^N (a - S_n) \right] \\
 &= \sum_{n=1}^k (a - S_n) + (a - S_k)(N - k) \\
 &= N(c - a) + \sum_{i=1}^k (k - i + 1)X_i + (N - k) \sum_{i=1}^k X_i \\
 &= N(c - a) + \sum_{i=1}^k (N - i + 1)X_i.
 \end{aligned}$$

Comparing the last two terms with (4) and (3) (at $k = N$) we conclude that under the rolling stack hedge

$$E_k[C_N] = E[C_k] + H_k. \quad (28)$$

More generally (i.e., dropping the assumption that $E[X_n] = 0$ and $b_{n,n+1} = 0$), whenever we can find a hedging strategy with cumulative cashflows H_k satisfying

$$H_k - E[H_k] = E[C_k] - E_k[C_N], \quad (29)$$

we get (using (29) with $k = N$ for the third equality)

$$\begin{aligned}
 \bar{C}_N = C_N + H_N &= E_N[C_N] + H_N \\
 &= E[C_N] + E[H_N] = E[\bar{C}_N],
 \end{aligned}$$

showing that the hedged cash balance $\bar{C}_N$ is riskless at the terminal date N . Equation (28) is a special case of (29) with $E[H_k] = 0$ because we took all $b_{n,n+1}$ to be zero. At intermediate dates, the exposure (actual cash balance minus expected) resulting from a hedge satisfying (29) is

$$\begin{aligned}\bar{C}_k - E[\bar{C}_k] &= C_k + H_k - E[C_k] - E[H_k] \\ &= C_k - E[C_k] + E[C_k] - E_k[C_N] \\ &= C_k - E_k[C_N],\end{aligned}$$

as claimed in (7). Thus, under any hedging strategy satisfying (29), the resulting exposure at intermediate times is given directly by (7). The same argument applies if the discrete time index is replaced with a continuous one. We used this shortcut in (10), (11) and (19).

Appendix C: Derivation of optimal paths

The derivations of the optimal paths use standard techniques from the calculus of variations, especially Sections 2.12 and 3.14 from Gelfand and Fomin (1963) for the unhedged and hedged cases, respectively. We detail the cases with $\alpha > 0$; the calculations for $\alpha = 0$ are similar but slightly simpler.

When there is no hedge, it is easy to see that we can replace the inequality constraint defining $\mathcal{A}_x$ with an equality, since the integral of the optimal path will not be any larger than required by the constraint. We thus need to find an extremal for

$$\frac{1}{2} \int_0^T [\dot{\psi}(t) + \alpha \psi(t)]^2 + \lambda [\psi(t) - x] dt,$$

with λ a Lagrange multiplier. As already noted, we may take $x = 1$ since x merely scales the path. The Euler equations give

$$\alpha^2 \psi - \ddot{\psi} = \text{constant}, \quad \psi(0) = 0 \quad (30)$$

$$\dot{\psi}(T) + \alpha \psi(T) = 0 \quad (31)$$

$$\int_0^T \psi = 1. \quad (32)$$

From (30) we obtain the general solution

$$\psi(t) = ae^{\alpha t} + be^{-\alpha t} - (a + b).$$

From (31) we get $b = (2 \exp(\alpha T) - 1)a$, and by eliminating b we can solve for a using (32).

Finding the optimal path in the hedged case is a free-endpoint problem because we do not know in advance the time τ at which

$$\psi(\tau) = h(\tau) \equiv -\frac{\alpha}{1 - \exp(T - \tau)}; \quad (33)$$

i.e., the time at which the shortfall occurs. The Euler equations give

$$\alpha^2 \psi - \ddot{\psi} = 0, \quad \psi(0) = 0$$

with the general solution $\psi(t) = 2c_1 \sinh(t)$. To find c_1 and τ we use (33) and the transversality condition

$$\frac{\alpha}{2} \psi(\tau) + \dot{h}(\tau) - \frac{1}{2} \dot{\psi}(\tau) = 0.$$

Some algebra shows that c_1 is as given in Section 6 and $\tau = T/3$. On $(\tau, T]$, the minimum-cost path should contribute no cost at all since the constraint for $\mathcal{A}_x$ has already been met. A zero cost path must have $\dot{\psi} + \alpha\psi = 0$; i.e., $\psi(t) = \psi(\tau) \exp(-\alpha(t - \tau))$, so that $c_2 = \psi(\tau) \exp(\alpha\tau)$.

References

- Adler, R.J., 1990, *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, Institute of Mathematical Statistics, Hayward, California.
- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D., 1996, A characterization of measures of risk, Working Paper, Universite Louis Pasteur, Strasbourg, France.
- Benson, A.W., 1994, MG Refining and Marketing Inc: hedging strategies revisited, *Plaintiff's reply to defendants MG Corp. and MGR&M, Civil Action No. JFM-94-484*, U.S. District Court of Maryland.
- Bessembinder, H., Coughenour, J.F., Seguin, P.J., and Smoller, M.M., 1995, Mean reversion in equilibrium asset prices: evidence from the futures term structure, *Journal of Finance*, **50**, 361–75.
- Brennan, M.J., 1991, The price of convenience and the valuation of commodity contingent claims, in (s.), *Stochastic Models and Option Values* ed. D. Lund and B. Øskendal, North-Holland, New York.
- Brennan, M.J., and Crew, N., 1997, Hedging long maturity commodity commitments with short-dated futures contracts, in *Mathematics of Derivative Securities*, M.A.H. Dempster and S.R. Pliska, eds., Cambridge University Press.
- Carverhill, A., 1998, Commodity futures and forwards: the HJM approach, Working Paper, Department of Finance, University of Science of Technology, Hong Kong.
- Culp, C.L., and Miller, M.H., 1995, Metallgesellschaft and the economics of synthetic storage, *J. Applied Corporate Finance*, **7**, 62–76.
- Dembo, A., and Zeitouni, O., 1998, *Large Deviations Techniques and Applications*, Second Edition, Springer-Verlag, New York.
- Duffie, D., 1989, *Futures Markets*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Edwards, F.A., and Canter, M.S., 1995, The collapse of Metallgesellschaft: unhedgeable risks, poor hedging strategy, or just bad luck?, *Journal of Futures Markets*, **15**, 211–64.
- Edwards, F.A., and C.W. Ma, 1992, *Futures and Options*, McGraw-Hill, New York.

- Frye, J., 1997 Principals of risk: finding VAR through factor-based interest rate scenarios, in *VAR: Understanding and Applying Value-at-Risk*, Risk Publications, London.
- Garbade, K.D., 1993, A two-factor, arbitrage-free model of fluctuations in crude oil futures prices, *Journal of Derivatives*, 1, 86–97.
- Gelfand, I.M., and Fomin, S.V., 1963, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Gibson, R., and Schwartz, E.S., 1990, Stochastic convenience yield and the pricing of oil contingent claims, *Journal of Finance*, 45, 959–76.
- Hilliard, J.E., 1996, Analytics underlying the Metallgesellschaft hedge: short term futures in a multi-period environment, Working paper, University of Georgia, Athens, Georgia.
- Jorion, P. 1997. *Value at Risk: The New Benchmark for Controlling Derivatives Risk*. McGraw-Hill, New York.
- Karatzas, I., and Shreve, S., 1991, *Brownian Motion and Stochastic Calculus*, 2nd Edition, Springer-Verlag, New York.
- Larcher, G. and Leobacher, G., 2000, An optimal strategy for hedging with short-term futures contracts, Working Paper, University of Salzburg, Austria.
- Marcus, M.B., and Shepp, L.A., 1971, Sample behavior of Gaussian processes, *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability*, 2, 423–42.
- Mello, A.S., and Parsons, J.E., 1995a, Maturity structure of a hedge matters: lessons from the Metallgesellschaft debacle, *Journal of Applied Corporate Finance*, 8, 106–20.
- Mello, A.S., and Parsons, J.E., 1995b, Funding risk and hedge valuation, Working Paper, University of Wisconsin.
- Mello, A.S., and Parsons, J.E., 1996, When hedging is risky: an example, Working Paper, University of Wisconsin.
- Neuberger, A., 1999, Hedging long term exposures with multiple short term futures contracts, *Review of Financial Studies*, 12, 429–60.
- Picault, E., 1998, Calculating value-at-risk with Monte Carlo simulation, in *Monte Carlo: Methodologies and Applications for Pricing and Risk Management*, ed. B. Dupire, Risk Publications, London.
- Piterbarg, V.I., 1996, *Asymptotic Methods in the Theory of Gaussian Processes and Fields*, American Mathematical Society, Providence, Rhode Island.
- Ross, S.A., 1995, Hedging long run commitments: exercises in incomplete market pricing, Working paper, Yale University.
- Stoll, H.R., and Whaley, R.E., 1993, *Futures and Options: Theory and Applications*, South-Western Publishing, Cincinnati, Ohio.
- Stroock, D.W., 1984, *An Introduction to the Theory of Large Deviations*, Springer-Verlag, Berlin.
- Talagrand, M., 1988, Small tails for the supremum of a Gaussian process, *Annales Institute Henri Poincaré*, 24, 307–15.
- Vasicek, O.A., 1977, An equilibrium characterization of the term structure, *Journal of Financial Economics*, 5, 177–88.
- Wakeman, L. 1999. Credit enhancement, In *Risk Management and Analysis, Vol 1*, ed. C. Alexander, 255–76. Wiley, Chichester, England.
- Wilson, T. 1999. Value at risk, In *Risk Management and Analysis, Vol 1*, ed. C. Alexander, 61–124. Wiley, Chichester, England.

Numerical Comparison of Local Risk-Minimisation and Mean-Variance Hedging

David Heath, Eckhard Platen and Martin Schweizer

1 Introduction

At present there is much uncertainty in the choice of the pricing measure for the hedging of derivatives in incomplete markets. Incompleteness can arise for instance in the presence of stochastic volatility, as will be studied in the following. This chapter provides comparative numerical results for two important hedging methodologies, namely local risk-minimisation and global mean-variance hedging.

We first describe the theoretical framework that underpins these two approaches. Some comparative studies are then presented on expected squared total costs and the asymptotics of these costs, differences in prices and optimal hedge ratios. In addition, the density functions for squared total costs and proportional transaction costs are estimated as well as mean transaction costs as a function of hedging frequency. Numerical results are obtained for variations of the Heston and the Stein–Stein stochastic volatility models.

To produce accurate and reliable estimates, combinations of partial differential equation and simulation techniques have been developed that are of independent interest. Some explicit solutions for certain key quantities required for mean-variance hedging are also described. It turns out that mean-variance hedging is far more difficult to implement than what has been attempted so far for most stochastic volatility models. In particular the mean-variance pricing measure is in many cases difficult to identify and to characterise. Furthermore, the corresponding optimal hedge, due to its global optimality properties, no longer appears as a simple combination of partial derivatives with respect to state variables. It has more the character of an optimal control strategy.

The importance of this chapter is that it documents for some typical stochastic volatility models some of the quantitative differences that arise for two major hedging approaches. We conclude by drawing attention to certain observations that have implications for the practical implementation of stochastic volatility models.

2 A Markovian stochastic volatility framework

We consider a frictionless market in continuous time with a single primary asset available for trade. We denote by $S = \{S_t, 0 \leq t \leq T\}$ the price process for this asset defined on the filtered probability space $(\Omega, \mathcal{F}, P)$ with filtration $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$ satisfying the usual conditions for some fixed but arbitrary time horizon $T \in (0, \infty)$.

We introduce the discounted price process $X = \{X_t = S_t/B_t, 0 \leq t \leq T\}$, where $B = \{B_t, 0 \leq t \leq T\}$ represents the savings account that accumulates interest at the continuously compounding interest rate.

We consider a general two-factor stochastic volatility model defined by stochastic differential equations (SDEs) of the form

$$\begin{aligned} dX_t &= X_t (\mu(t, Y_t) dt + Y_t dW_t^1) \\ dY_t &= a(t, Y_t) dt + b(t, Y_t)(\varrho dW_t^1 + \sqrt{1 - \varrho^2} dW_t^2) \end{aligned} \quad (2.1)$$

for $0 \leq t \leq T$ with given deterministic initial values $X_0 \in (0, \infty)$ and $Y_0 \in (0, \infty)$. Here the function μ is a given appreciation rate. The volatility component Y evolves according to a separate SDE with drift function a , diffusion function b and constant correlation $\varrho \in [-1, 1]$. W^1 and W^2 denote independent standard Wiener processes under P . The component Y allows for an additional source of randomness but is not available as a traded asset.

To ensure that this Markovian framework provides a viable asset price model we assume that appropriate conditions hold for the functions μ, a, b so that the system of SDEs (2.1) admits a unique strong continuous solution for the vector process (X, Y) with a strictly positive discounted price process X and a volatility process Y . We take the filtration $\mathbb{F}$ to be the P -augmentation of the natural filtration generated by W^1 and W^2 .

In order to price and hedge derivatives in an arbitrage free manner we assume that there exists an equivalent local martingale measure (ELMM) Q . This is a probability measure Q with the same null sets as P and such that X is a local martingale under Q .

We denote by $\mathbb{P}$ the set of all ELMMs Q . Our financial market is characterised by the system (2.1) together with the filtration $\mathbb{F}$ and is called incomplete if $\mathbb{P}$ contains more than one element.

In this chapter we are in principle interested in the hedging of European style contingent claims with an $\mathcal{F}_T$ -measurable square integrable random payoff H based on the dynamics given by (2.1). A specific choice for H which we will use later on for our numerical examples is the European put option with payoff

given by

$$H = h(X_T) = (K - X_T)^+. \quad (2.2)$$

The requirement of $\mathcal{F}_T$ -measurability and square integrability for the payoff H allows for many types of path dependent contingent claims and possibly even dependence on the evolution of the volatility process Y .

Subject to certain restrictions on the functions μ , a , b and parameter ϱ we can ensure, via an application of the Girsanov transformation, that there is an ELMM Q .

The condition that X should be a local Q -martingale fixes the effect of the Girsanov transformation on W^1 but allows for different transformations on the independent W^2 . Consequently if $|\varrho| < 1$ the set $\mathbb{P}$ contains more than one element and our financial market is therefore incomplete.

In order to price and hedge derivatives in this incomplete market setting we need to somehow fix the ELMM Q . Currently there is no general agreement on how to choose a specific ELMM Q and a number of alternatives are being considered in the literature.

In this chapter we will consider two quadratic approaches to hedging in incomplete markets; these are local risk-minimisation and mean-variance hedging. For either of these two approaches we require hedging strategies of the form $\varphi = (\vartheta, \eta)$, where ϑ is a predictable X -integrable process and η is an adapted process such that the value process $V(\varphi) = \{V_t(\varphi), 0 \leq t \leq T\}$ with

$$V_t(\varphi) = \vartheta_t X_t + \eta_t \quad (2.3)$$

is right-continuous for $0 \leq t \leq T$. Using the hedging strategy $\varphi = (\vartheta, \eta)$ means that we form at time t a portfolio with ϑ_t units of the traded risky asset X_t and η_t units of the savings account.

The cost process $C(\varphi) = \{C_t(\varphi), 0 \leq t \leq T\}$ is then given by

$$C_t(\varphi) = V_t(\varphi) - \int_0^t \vartheta_s dX_s \quad (2.4)$$

for $0 \leq t \leq T$ and $\varphi = (\vartheta, \eta)$. A hedging strategy φ is self-financing if $C(\varphi)$ is P -a.s. constant over the time interval $[0, T]$ and φ is called mean self-financing if $C(\varphi)$ is a P -martingale.

3 Local risk-minimisation

Intuitively the goal of local risk-minimisation is to minimise the local risk defined as the conditional second moment of cost increments under the measure P at each time instant.

With local risk-minimisation we only consider hedging strategies which replicate the contingent claim H at time T ; that is we only allow hedging strategies φ such that

$$V_T(\varphi) = H \quad P\text{-a.s.} \quad (3.1)$$

Subject to certain technical conditions it can be shown that finding a locally risk-minimising strategy is equivalent to finding a decomposition of H in the form

$$H = H_0^{\text{lr}} + \int_0^T \xi_s^{\text{lr}} dX_s + L_T^{\text{lr}}, \quad (3.2)$$

where H_0^{lr} is constant, ξ^{lr} is a predictable process satisfying suitable integrability properties and $L^{\text{lr}} = \{L_t^{\text{lr}}, 0 \leq t \leq T\}$ is a square integrable P -martingale with $L_0^{\text{lr}} = 0$ and such that the product process $L^{\text{lr}}M$ is in addition a P -martingale, where M is the martingale part of X . The representation (3.2) is usually referred to as the Föllmer–Schweizer decomposition of H , see Föllmer & Schweizer (1991).

The locally risk-minimising hedging strategy is then given by

$$\vartheta_t^{\text{lr}} = \xi_t^{\text{lr}} \quad (3.3)$$

and

$$\eta_t^{\text{lr}} = V_t(\varphi^{\text{lr}}) - \vartheta_t^{\text{lr}} X_t, \quad (3.4)$$

where

$$V_t(\varphi^{\text{lr}}) = C_t(\varphi^{\text{lr}}) + \int_0^t \vartheta_s^{\text{lr}} dX_s \quad (3.5)$$

with

$$C_t(\varphi^{\text{lr}}) = H_0^{\text{lr}} + L_t^{\text{lr}} \quad (3.6)$$

for $0 \leq t \leq T$.

As is shown in Föllmer & Schweizer (1991) and Schweizer (1995) there exists a measure $\hat{P}$, the so-called minimal ELMM, such that

$$V_t(\varphi^{\text{lr}}) = E_{\hat{P}}[H \mid \mathcal{F}_t] \quad (3.7)$$

for $0 \leq t \leq T$, where the conditional expectation in (3.7) is taken under $\hat{P}$. The measure $\hat{P}$ is identified, subject to certain integrability conditions, by the Radon–Nikodým derivative

$$\frac{d\hat{P}}{dP} = \hat{Z}_T, \quad (3.8)$$

where

$$\hat{Z}_t = \exp \left(-\frac{1}{2} \int_0^t \left(\frac{\mu(s, Y_s)}{Y_s} \right)^2 ds - \int_0^t \frac{\mu(s, Y_s)}{Y_s} dW_s^1 \right) \quad (3.9)$$

for $0 \leq t \leq T$.

Assuming $\hat{Z}$ is a P -martingale, the Girsanov transformation can be used to show that the processes $\hat{W}^1$ and $\hat{W}^2$ defined by

$$\hat{W}_t^1 = W_t^1 + \int_0^t \frac{\mu(s, Y_s)}{Y_s} ds \quad (3.10)$$

and

$$\hat{W}_t^2 = W_t^2 \quad (3.11)$$

for $0 \leq t \leq T$ are independent Wiener processes under $\hat{P}$. Consequently, using $\hat{W}^1$ and $\hat{W}^2$, the system of stochastic differential equations (2.1) becomes

$$\begin{aligned} dX_t &= X_t Y_t d\hat{W}_t^1 \\ dY_t &= \left(a(t, Y_t) - \frac{\varrho}{Y_t} (b \mu)(t, Y_t) \right) dt \\ &\quad + b(t, Y_t) \left(\varrho d\hat{W}_t^1 + \sqrt{1 - \varrho^2} d\hat{W}_t^2 \right) \end{aligned} \quad (3.12)$$

for $0 \leq t \leq T$.

Taking contingent claims of the form $H = h(X_T)$ for some given function $h : [0, \infty) \rightarrow \mathbb{R}$ and using the Markov property we can rewrite (3.7) in the form

$$\begin{aligned} V_t(\varphi^{\text{tr}}) &= E_{\hat{P}}[h(X_T) | \mathcal{F}_t] \\ &= v_{\hat{P}}(t, X_t, Y_t) \end{aligned} \quad (3.13)$$

for some function $v_{\hat{P}}(t, x, y)$ defined on $[0, T] \times (0, \infty) \times \mathbb{R}$. Subject to certain regularity conditions we can show that $v_{\hat{P}}$ is the solution to the partial differential equation (PDE)

$$\frac{\partial v_{\hat{P}}}{\partial t} + \left(a - \frac{\varrho b \mu}{y} \right) \frac{\partial v_{\hat{P}}}{\partial y} + \frac{1}{2} \left(x^2 y^2 \frac{\partial^2 v_{\hat{P}}}{\partial x^2} + b^2 \frac{\partial^2 v_{\hat{P}}}{\partial y^2} + 2 \varrho x y b \frac{\partial^2 v_{\hat{P}}}{\partial x \partial y} \right) = 0 \quad (3.14)$$

on $(0, T) \times (0, \infty) \times \mathbb{R}$ with boundary condition

$$v_{\hat{P}}(T, x, y) = h(x) \quad (3.15)$$

for $x \in (0, \infty)$, $y \in \mathbb{R}$. Solving this PDE yields the pricing function (3.13) for local risk-minimisation.

Now it follows by application of Itô's formula together with (3.14) that

$$V_t(\varphi^{\text{lr}}) = V_0(\varphi^{\text{lr}}) + \int_0^t \vartheta_s^{\text{lr}} dX_s + L_t^{\text{lr}}, \quad (3.16)$$

where

$$\vartheta_t^{\text{lr}} = \frac{\partial v_{\hat{P}}}{\partial x}(t, X_t, Y_t) + \frac{\varrho}{X_t Y_t} b(t, Y_t) \frac{\partial v_{\hat{P}}}{\partial y}(t, X_t, Y_t) \quad (3.17)$$

and

$$L_t^{\text{lr}} = \int_0^t \sqrt{1 - \varrho^2} b(s, Y_s) \frac{\partial v_{\hat{P}}}{\partial y}(s, X_s, Y_s) dW_s^2 \quad (3.18)$$

for $0 \leq t \leq T$.

Using (3.6) and (3.18) we see that the conditional expected squared cost on the interval $[t, T]$ for the locally risk-minimising strategy φ^{lr} , denoted by R_t^{lr} , is given by

$$\begin{aligned} R_t^{\text{lr}} &= E \left[\left(C_T(\varphi^{\text{lr}}) - C_t(\varphi^{\text{lr}}) \right)^2 \mid \mathcal{F}_t \right] \\ &= E \left[\int_t^T (1 - \varrho^2) \left(b(s, Y_s) \frac{\partial v_{\hat{P}}}{\partial y}(s, X_s, Y_s) \right)^2 ds \mid \mathcal{F}_t \right]. \end{aligned} \quad (3.19)$$

4 Mean-variance hedging

In this section we consider an alternative approach to hedging in incomplete markets based on what is called mean-variance hedging. Intuitively the goal here is to minimise the global quadratic risk over the entire time interval $[0, T]$. This contrasts with local risk-minimisation which focuses on minimisation of the second moments of infinitesimal cost increments.

With mean-variance hedging we allow strategies which do not fully replicate the contingent claim H at time T . However, we minimise

$$E \left[\left(H - V_0 - \int_0^T \vartheta_s dX_s \right)^2 \right] \quad (4.1)$$

over an appropriate choice of initial value V_0 and hedge ratio ϑ . The pair of initial value and hedge ratio process which minimises this quantity is called the mean-variance optimal strategy and is denoted by $(V_0^{\text{mvo}}, \vartheta^{\text{mvo}})$ with

$$R_0^{\text{mvo}} = E \left[\left(H - V_0^{\text{mvo}} - \int_0^T \vartheta_s^{\text{mvo}} dX_s \right)^2 \right]. \quad (4.2)$$

Given an initial value V_0 and hedge ratio ϑ we can always construct a self-financing strategy $\varphi = (\vartheta, \eta)$ by choosing

$$\eta_t = V_0 + \int_0^t \vartheta_s dX_s - \vartheta_t X_t \quad (4.3)$$

for $0 \leq t \leq T$. The quantity

$$H - V_T(\varphi) = H - V_0 - \int_0^T \vartheta_s dX_s \quad (4.4)$$

appearing in (4.1) is then the net loss or shortfall at time T using the strategy φ with payment H . For a more precise specification of mean-variance hedging see Heath, Platen & Schweizer (2000).

Using (2.4), (3.1) and the first equation in (3.19) we see that

$$\begin{aligned} R_0^{\text{lr}} &= E \left[\left(H - V_0(\varphi^{\text{lr}}) - \int_0^T \vartheta_u^{\text{lr}} dX_u \right)^2 \right] \\ &\geq E \left[\left(H - V_0^{\text{mvo}} - \int_0^T \vartheta_u^{\text{mvo}} dX_u \right)^2 \right] = R_0^{\text{mvo}}. \end{aligned}$$

Thus, mean-variance hedging by definition delivers expected squared costs which are less than or equal to those obtained for the locally risk-minimising strategy.

Under suitable conditions it can be shown that the contingent claim H admits a decomposition of the form

$$H = \tilde{H}_0 + \int_0^T \tilde{\xi}_s dX_s + \tilde{L}_T, \quad (4.5)$$

where

$$V_0^{\text{mvo}} = \tilde{H}_0 = E_{\tilde{P}}[H], \quad (4.6)$$

$\tilde{\xi}$ is a predictable process satisfying suitable integrability properties and $\tilde{L}$ is a $\tilde{P}$ -martingale with $\tilde{L}_0 = 0$. The ELMM $\tilde{P}$ in (4.6) is the so-called variance-optimal measure; it appears naturally as the solution of a problem dual to minimising (4.1).

If we choose a self-financing strategy $\varphi^{\text{mvo}} = (\vartheta^{\text{mvo}}, \eta^{\text{mvo}})$ with η^{mvo} defined as in (4.3) then using (4.5) and (4.6) the net loss at time T is given by

$$\begin{aligned} H - V_T(\varphi^{\text{mvo}}) &= H - V_0^{\text{mvo}} - \int_0^T \vartheta_s^{\text{mvo}} dX_s \\ &= \tilde{L}_T + \int_0^T (\tilde{\xi}_s - \vartheta_s^{\text{mvo}}) dX_s. \end{aligned} \quad (4.7)$$

Under suitable conditions and with $\varrho = 0$ it can be shown that $\tilde{P}$ can be identified from its Radon–Nikodým derivative in the form

$$\frac{d\tilde{P}}{dP} = \tilde{Z}_T, \quad (4.8)$$

where

$$\begin{aligned} \tilde{Z}_t = \exp & \left(- \int_0^t \frac{\mu(s, Y_s)}{Y_s} dW_s^1 - \int_0^t \tilde{v}_s dW_s^2 \right. \\ & \left. - \frac{1}{2} \int_0^t \left[\left(\frac{\mu(s, Y_s)}{Y_s} \right)^2 + (\tilde{v}_s)^2 \right] ds \right) \end{aligned} \quad (4.9)$$

with

$$\tilde{v}_t = b(t, Y_t) \frac{\partial J}{\partial y}(t, Y_t) \quad (4.10)$$

and

$$J(t, y) = -\log E \left[\exp \left(- \int_t^T \left(\frac{\mu(s, Y_s^{t,y})}{Y_s^{t,y}} \right)^2 ds \right) \right] \quad (4.11)$$

for $0 \leq t \leq T$. Here we denote by $Y^{t,y}$ the volatility process that starts at time t with value y and evolves according to the SDE (2.1).

Applying the Feynman–Kac formula to the function $\exp(-J)$ and using a transformation of variables back to the function J it can be shown that, under appropriate conditions for a, b and μ , J satisfies the PDE

$$\frac{\partial J}{\partial t} + a \frac{\partial J}{\partial y} + \frac{1}{2} b^2 \frac{\partial^2 J}{\partial y^2} - \frac{1}{2} b^2 \left(\frac{\partial J}{\partial y} \right)^2 + \left(\frac{\mu}{y} \right)^2 = 0 \quad (4.12)$$

on $(0, T) \times \mathbb{R}$ with boundary conditions

$$J(T, y) = 0.$$

Assuming $\tilde{Z}$ is a P -martingale, an application of the Girsanov transformation shows that the processes $\tilde{W}^1$ and $\tilde{W}^2$ defined by

$$\tilde{W}_t^1 = W_t^1 + \int_0^t \frac{\mu(s, Y_s)}{Y_s} ds \quad (4.13)$$

and

$$\tilde{W}_t^2 = W_t^2 + \int_0^t \tilde{v}_s ds \quad (4.14)$$

for $0 \leq t \leq T$ are independent Wiener processes under $\tilde{P}$. Hence with respect to $\tilde{W}^1$ and $\tilde{W}^2$ the system of stochastic differential equations (2.1) becomes

$$\begin{aligned} dX_t &= X_t Y_t d\tilde{W}_t^1 \\ dY_t &= \left[a(t, Y_t) - b^2(t, Y_t) \frac{\partial J}{\partial y}(t, Y_t) \right] dt + b(t, Y_t) d\tilde{W}_t^2 \end{aligned} \quad (4.15)$$

for $0 \leq t \leq T$. Note that we have assumed $\varrho = 0$.

As in the case for local risk-minimisation we consider European contingent claims of the form $H = h(X_T)$. For this type of payoff and again using the Markov property and prescription (4.3) we can express by (4.5) and (4.6) the initial value $V_0(\varphi^{\text{mvo}})$ in the form

$$V_0(\varphi^{\text{mvo}}) = V_0^{\text{mvo}} = E_{\tilde{P}}[H] = v_{\tilde{P}}(0, X_0, Y_0) \quad (4.16)$$

for some function $v_{\tilde{P}}(t, x, y)$ defined on $[0, T] \times (0, \infty) \times \mathbb{R}$ such that

$$v_{\tilde{P}}(t, X_t, Y_t) = E_{\tilde{P}}[H | \mathcal{F}_t]. \quad (4.17)$$

Subject to certain regularity conditions, it can be shown that $v_{\tilde{P}}$ is the solution of the PDE

$$\frac{\partial v_{\tilde{P}}}{\partial t} + \left[a - b^2 \frac{\partial J}{\partial y} \right] \frac{\partial v_{\tilde{P}}}{\partial y} + \frac{1}{2} x^2 y^2 \frac{\partial^2 v_{\tilde{P}}}{\partial x^2} + \frac{1}{2} b^2 \frac{\partial^2 v_{\tilde{P}}}{\partial y^2} = 0 \quad (4.18)$$

on $(0, T) \times (0, \infty) \times \mathbb{R}$ with boundary condition

$$v_{\tilde{P}}(T, x, y) = h(x) \quad (4.19)$$

for $x \in (0, \infty)$, $y \in \mathbb{R}$.

Similar to the case for local risk-minimisation we can apply the Itô formula combined with (4.15), (4.16) and (4.18) to obtain

$$v_{\tilde{P}}(t, X_t, Y_t) = V_0^{\text{mvo}} + \int_0^t \tilde{\xi}_s dX_s + \tilde{L}_t, \quad (4.20)$$

where

$$\tilde{\xi}_t = \frac{\partial v_{\tilde{P}}}{\partial x}(t, X_t, Y_t) \quad (4.21)$$

and

$$\tilde{L}_t = \int_0^t b(s, Y_s) \frac{\partial v_{\tilde{P}}}{\partial y}(s, X_s, Y_s) d\tilde{W}_s^2 \quad (4.22)$$

for $0 \leq t \leq T$.

Also, under suitable conditions, it can be shown that the expected squared cost over the interval $[0, T]$ is given by

$$R_0^{\text{mvo}} = E \left[\int_0^T e^{-J(s, Y_s)} b^2(s, Y_s) \left(\frac{\partial v_{\tilde{P}}}{\partial y}(s, X_s, Y_s) \right)^2 ds \right]. \quad (4.23)$$

Furthermore, the mean-variance optimal hedge ratio ϑ^{mvo} is given in feedback form by

$$\vartheta_t^{\text{mvo}} = \tilde{\xi}_t + \frac{\mu(t, Y_t)}{X_t Y_t^2} \left(v_{\tilde{P}}(t, X_t, Y_t) - \tilde{H}_0 - \int_0^t \vartheta_s^{\text{mvo}} dX_s \right). \quad (4.24)$$

Thus in the case of mean-variance hedging the optimal hedge ratio ϑ^{mvo} is in general not equal to $\tilde{\xi}$ which is the integrand appearing in the decomposition (4.5). This might not have been expected based on the results obtained for local risk-minimisation and is due to the fact that ϑ_t^{mvo} has more the character of an optimal control variable.

Finally, in the case where $\tilde{P} = \hat{P}$, so that $v_{\tilde{P}} = v_{\hat{P}}$, and, again subject to certain conditions, see Heath, Platen & Schweizer (2000), it can be shown that

$$R_0^{\text{mvo}} = E \left[\int_0^T e^{-J(s, Y_s)} (1 - \varrho^2) b^2(s, Y_s) \left(\frac{\partial v_{\hat{P}}}{\partial y}(s, X_s, Y_s) \right)^2 ds \right], \quad (4.25)$$

which is similar to (4.23) but includes the case $\varrho \neq 0$.

5 Some specific models

In this section we will consider the application of both local risk-minimisation and mean-variance hedging to four stochastic volatility models. The purpose of this study is to compare various quantities for the two hedging approaches and the given models. This will provide insight into qualitative and quantitative differences for the two quadratic hedging approaches.

The models which we examine are based on the Stein & Stein (1991) and Heston (1993) type stochastic volatility models with two different specifications for the appreciation rate function μ .

The four models with their specifications are summarised in Table 1. Here S1 and S2 are the two Stein–Stein type models and H1 and H2 are the two Heston type models. We assume that the constants $\delta, \beta, k, \kappa, \theta, \Sigma$ are non-negative, with Δ and γ real valued and $\varrho \in [-1, 1]$. Note that non-zero correlation is allowed only for the H1 model. For the H1 and H2 models an SDE for the volatility component Y can be obtained via Itô's formula as follows:

$$dY_t = \left(\frac{4\kappa(\theta - Y_t^2) - \Sigma^2}{8Y_t} \right) dt + \frac{\Sigma}{2} \left(\varrho dW_t^1 + \sqrt{1 - \varrho^2} dW_t^2 \right). \quad (5.1)$$

Table 1. *Model specifications.*

Model	Volatility dynamics Y	Appreciation Rate μ
S1	$dY_t = \delta (\beta - Y_t) dt + k dW_t^2$	$\mu(t, Y_t) = \Delta Y_t$
S2	as above	$\mu(t, Y_t) = \gamma (Y_t)^2$
H1	$d(Y_t)^2 = \kappa (\theta - (Y_t)^2) dt + \Sigma Y_t (\varrho dW_t^1 + \sqrt{1 - \varrho^2} dW_t^2)$	$\mu(t, Y_t) = \Delta Y_t$
H2	$d(Y_t)^2 = \kappa (\theta - (Y_t)^2) dt + \Sigma Y_t dW_t^2$	$\mu(t, Y_t) = \gamma (Y_t)^2$

For the S1 and H1 models it can be shown, see Heath, Platen & Schweizer (2000), that $\tilde{P} = \hat{P}$ and that

$$J(t, y) = \Delta^2(T - t) \quad (5.2)$$

for $(t, y) \in [0, T] \times \mathbb{R}$. By (3.19) and (4.25) this means that

$$\begin{aligned} R_0^{\text{mvo}} &= E \left[\int_0^T e^{-\Delta^2(T-s)} (1 - \varrho^2) b^2(s, Y_s) \left(\frac{\partial v_{\tilde{P}}}{\partial y}(s, X_s, Y_s) \right)^2 ds \right] \\ &\geq e^{-\Delta^2 T} R_0^{\text{lr}}. \end{aligned} \quad (5.3)$$

In addition it can be shown that the locally risk-minimising strategy is given by (3.17).

In the next section we compute the locally risk-minimising strategies for both the S1 and H1 models based on the formulae (3.12), (3.14), (3.17) and (3.19). We note that the derivations and technical details provided in the papers Heath, Platen & Schweizer (2000) and Schweizer (1991) do not fully cover the case of $\varrho \neq 0$ for the H1 model that have also been included for comparative purposes in our study. However, the numerical results obtained do not indicate any particular problems with this case.

For the S2 and H2 models it can be shown, see again Heath, Platen & Schweizer (2000), that both the locally risk-minimising and mean-variance optimal hedging strategies exist for the case of a European put option. Note that for mean-variance hedging existence of the optimal strategy is established only for a sufficiently small time horizon T . However, also in this case the numerical experiments have been successfully performed for long time scales without apparent difficulties, as will be seen in the next section.

For the S2 and H2 models we have from (4.11) and Table 1 the function

$$J(t, y) = -\log E \left[\exp \left(-\gamma^2 \int_t^T (Y_s^{t,y})^2 ds \right) \right]. \quad (5.4)$$

Fortunately for both models this function can be computed explicitly, see again Heath, Platen & Schweizer (2000). In the case of the S2 model the J function in (5.4) is denoted by the symbol J_{S2} and has the form

$$J_{S2}(t, y) = f_0(T - t) + f_1(T - t) \frac{y}{k} + f_2(T - t) \frac{y^2}{k^2}. \quad (5.5)$$

For the S2 model we have $a(t, y) = \delta(\beta - y)$ and $b(t, y) = k$. Using these specifications for the drift and diffusion coefficients and substituting (5.5) into (4.12) we can show that the functions f_0 , f_1 and f_2 satisfy the ordinary differential equations (ODEs)

$$\begin{aligned} \frac{d}{d\tau} f_0(\tau) + f_1(\tau) \left(\frac{1}{2} f_1(\tau) - \frac{\beta \delta}{k} \right) - f_2(\tau) &= 0, \\ \frac{d}{d\tau} f_1(\tau) + f_1(\tau) (\delta + 2 f_2(\tau)) - \frac{2\beta\delta}{k} f_2(\tau) &= 0, \\ \frac{d}{d\tau} f_2(\tau) + 2 f_2(\tau) (\delta + f_2(\tau)) - k^2 \gamma^2 &= 0, \end{aligned} \quad (5.6)$$

with boundary conditions

$$f_0(0) = f_1(0) = f_2(0) = 0. \quad (5.7)$$

These equations can be solved explicitly, yielding

$$\begin{aligned} f_2(\tau) &= \frac{\lambda \gamma_1 e^{-2\gamma_1 \tau}}{\lambda + \gamma_1 - \lambda e^{-2\gamma_1 \tau}} - \lambda, \\ f_1(\tau) &= \frac{1}{1 + 2\lambda \psi(\tau)} \left((2D - D') e^{-2\gamma_1 \tau} - 2D e^{-2\gamma_1 \tau} \right) + D', \\ f_0(\tau) &= \frac{1}{2} \log(1 + 2\lambda \psi(\tau)) - \left(\lambda + \frac{\delta^2 \beta^2}{2k^2} \left(\frac{\delta^2}{\gamma_1^2} - 1 \right) \right) \tau - \frac{2D^2 \psi(\tau)}{1 + 2\lambda \psi(\tau)} \\ &\quad + \frac{\delta^2 \beta}{k \gamma_1^2} \left(\frac{1}{1 + 2\lambda \psi(\tau)} \left(2D e^{-\gamma_1 \tau} - \left(D - \frac{1}{2} D' \right) e^{-2\gamma_1 \tau} \right) - \left(D + \frac{1}{2} D' \right) \right) \end{aligned}$$

with constants

$$\gamma_1 = \sqrt{2k^2 \gamma^2 + \delta^2}, \quad \lambda = \frac{\delta - \gamma_1}{2},$$

$$D = \frac{\delta \beta}{2k} \left(1 - \frac{\delta^2}{\gamma_1^2} \right), \quad D' = \frac{\delta \beta}{k} \left(1 - \frac{\delta}{\gamma_1} \right)$$

and function

$$\psi(\tau) = \frac{1 - e^{-2\gamma_1\tau}}{2\gamma_1}.$$

Although the calculations are somewhat lengthy it can be verified by direct substitution that these analytic expressions are indeed the solution of (5.6)–(5.7). This was also confirmed for the models considered in the next section by solving (5.6)–(5.7) numerically and comparing these results with those obtained from the analytic solution. Furthermore, the ODE formulation can be used in situations where we replace one or more of the constant coefficients δ , β or k with time-dependent deterministic functions satisfying suitable regularity conditions.

The $\tilde{P}$ dynamics for the volatility component Y for the S2 model can now be obtained from (4.15) with the formula

$$\frac{\partial J_{S2}}{\partial y}(t, y) = \frac{f_1(T-t)}{k} + \frac{2 f_2(T-t) y}{k^2}. \quad (5.8)$$

For the H2 model the J function in (5.4), denoted by J_{H2} , is given by the expression

$$J_{H2}(t, y) = g_0(T-t) + g_1(T-t) y^2. \quad (5.9)$$

Using the H2 model specifications $a(t, y) = (4\kappa(\theta - y^2) - \Sigma^2)/8y$ and $b(t, y) = \Sigma/2$ and substituting (5.9) into (4.12) we see that the functions g_0 and g_1 satisfy the ODEs

$$\begin{aligned} \frac{d}{d\tau} g_0(\tau) - \kappa \theta g_1(\tau) &= 0, \\ \frac{d}{d\tau} g_1(\tau) + g_1(\tau) \left(\kappa + \frac{1}{2} \Sigma^2 g_1(\tau) \right) - \gamma^2 &= 0 \end{aligned} \quad (5.10)$$

with boundary conditions

$$g_0(0) = g_1(0) = 0. \quad (5.11)$$

These equations can also be solved explicitly with

$$\begin{aligned} g_0(\tau) &= -\frac{2\kappa\theta}{\Sigma^2} \ln \left(\frac{2\Gamma e^{\frac{\Gamma+\kappa}{2}\tau}}{(\Gamma+\kappa)(e^{\Gamma\tau}-1) + 2\Gamma} \right), \\ g_1(\tau) &= \frac{2\gamma^2(e^{\Gamma\tau}-1)}{(\Gamma+\kappa)(e^{\Gamma\tau}-1) + 2\Gamma} \end{aligned}$$

and

$$\Gamma = \sqrt{2\gamma^2\Sigma^2 + \kappa^2}.$$

It can be shown by direct substitution that these analytic expressions are the solutions of (5.10) – (5.11). Also these ODEs can under appropriate conditions be used in versions of the H2 model with time-dependent deterministic parameters.

The $\tilde{P}$ dynamics for the volatility component Y for the H2 model can now be obtained from (4.15) with

$$\frac{\partial J_{H2}}{\partial y}(t, y) = 2 g_2(T - t) y. \quad (5.12)$$

For a justification of the approach using PDEs which is applied in the next section to all four combinations of models, see Heath and Schweizer (2000).

6 Computation of expected squared costs, prices and hedge ratios

The purpose of this section is to compare actual numerical results for both hedging approaches for the models previously introduced. Emphasis will be placed on experiments which highlight differences in key quantities such as prices, expected squared total costs and hedge ratios. For the four models and two hedging frameworks extensive experimentation has been performed with different parameter sets. Only a small subset of these results can be presented in this chapter. Nevertheless these results indicate some crucial differences between the two approaches that might be of more general interest. In total eight different hedging problems had to be solved with corresponding numerical tools developed. For all numerical experiments considered here the contingent claim was taken to be a European put, see (2.2). This ensures the payoff function h is bounded and avoids integrability problems.

To solve numerically the PDEs (3.14)–(3.15) and (4.18)–(4.19) we employed finite difference approximations based on the Crank–Nicolson scheme. Some experimentation was also performed using the fully implicit scheme. To handle the two-dimensional structures appearing in (3.14) and (4.18) we used the method of fractional steps or operator splitting. For a discussion on these and related techniques, see Fletcher (1988), Sections 8.2–8.5, and Hoffman (1993), Chapters 11 and 14.

Fractional step methods are usually easier to implement in the case where there is no correlation in the diffusion terms, that is $\varrho = 0$, and thus the term in (3.14) corresponding to the cross-term partial derivative $\frac{\partial^2 v_{\tilde{P}}}{\partial x \partial y}$ is zero. In the H1 model which allows for non-zero correlation we obtained an orthogonalised system of equations by introducing the transformation

$$Z_t = \ln(X_t) - \frac{\varrho}{\Sigma} Y_t^2 \quad (6.1)$$

for $0 \leq t \leq T$ and $\Sigma > 0$.

By Itô's formula, together with (3.12) and (5.1), the evolution of Z is governed by the SDE

$$\begin{aligned} dZ_t = & \left[\left(\frac{\varrho \kappa}{\Sigma} - \frac{1}{2} \right) Y_t^2 + \varrho^2 \Delta Y_t - \frac{\varrho \kappa \theta}{\Sigma} \right] dt \\ & + Y_t \left[(1 - \varrho^2) d\hat{W}_t^1 - \varrho \sqrt{1 - \varrho^2} d\hat{W}_t^2 \right] \end{aligned} \quad (6.2)$$

for $0 \leq t \leq T$. Using this transformation for a European put option with strike price K we obtain from the Kolmogorov backward equation a transformed function $u_{\hat{P}}$ defined on $[0, T] \times \mathbb{R} \times \mathbb{R}$ which is the solution of the PDE

$$\begin{aligned} \frac{\partial u_{\hat{P}}}{\partial t} + & \left[\left(\frac{\varrho \kappa}{\Sigma} - \frac{1}{2} \right) y^2 + \varrho^2 \Delta y - \frac{\varrho \kappa \theta}{\Sigma} \right] \frac{\partial u_{\hat{P}}}{\partial z} \\ & + \left(\frac{4\kappa\beta - \Sigma^2}{8y} - \frac{\kappa y}{2} - \frac{\varrho \Sigma \Delta}{2} \right) \frac{\partial u_{\hat{P}}}{\partial y} \\ & + \frac{1}{2} y^2 (1 - \varrho^2) \frac{\partial^2 u_{\hat{P}}}{\partial z^2} + \frac{\Sigma^2}{8} \frac{\partial^2 u_{\hat{P}}}{\partial y^2} = 0 \end{aligned} \quad (6.3)$$

on $(0, T) \times \mathbb{R} \times \mathbb{R}$ with boundary condition

$$u_{\hat{P}}(T, z, y) = \left(K - \exp \left(z + \frac{\varrho y^2}{\Sigma} \right) \right)^+ . \quad (6.4)$$

In terms of the original pricing function $v_{\hat{P}}$ we have the relation

$$v_{\hat{P}}(t, x, y) = u_{\hat{P}}(t, \ln(x) - \frac{\varrho y^2}{\Sigma}, y). \quad (6.5)$$

As noted previously, for the H1 model we have $\tilde{P} = \hat{P}$ and the corresponding locally risk-minimising and mean-variance prices are the same.

For the numerical experiments described in this paper the following default values were used: For the Heston and Stein–Stein models $\kappa = 5.0$, $\theta = 0.04$, $\Sigma = 0.6$, $\delta = 5.0$, $\beta = 0.2$ and $k = 0.3$. Models other than the H1 model have $\varrho = 0.0$ and for the appreciation rate μ from Table 1 we took $\Delta = 0.5$ and $\gamma = 2.5$. Other default parameters were $X_0 = 100.0$ and $Y_0 = 0.2$ as initial values for X and Y and strike $K = 100.0$ and time to maturity $T = 1.0$ for option parameters.

To compute the expected squared costs on the interval $[0, T]$ given by (3.19) and (4.23), respectively, we introduce the functions ζ^{lr} and ζ^{mvo} defined on $[0, T] \times (0, \infty) \times \mathbb{R}$ given by

$$\zeta^{\text{lr}}(t, x, y) = (1 - \varrho^2) b^2(t, y) \left(\frac{\partial v_{\hat{P}}}{\partial y}(t, x, y) \right)^2 \quad (6.6)$$

and

$$\zeta^{\text{mvo}}(t, x, y) = (1 - \varrho^2) e^{-J(t, y)} b^2(t, y) \left(\frac{\partial v_{\tilde{P}}}{\partial y}(t, x, y) \right)^2 \quad (6.7)$$

for $(t, x, y) \in [0, T] \times (0, \infty) \times \mathbb{R}$.

By (3.19) and (6.6) it follows that

$$R_t^{\text{lr}} = E \left[\int_t^T \zeta^{\text{lr}}(s, X_s, Y_s) ds \mid \mathcal{F}_t \right].$$

We can now apply the Kolmogorov backward equation together with (2.1) to show that there is a function r^{lr} defined on $[0, T] \times (0, \infty) \times \mathbb{R}$ such that

$$r^{\text{lr}}(t, X_t, Y_t) = R_t^{\text{lr}}$$

and r^{lr} is the solution to the PDE

$$\frac{\partial r^{\text{lr}}}{\partial t} + x \mu \frac{\partial r^{\text{lr}}}{\partial x} + a \frac{\partial r^{\text{lr}}}{\partial y} + \frac{1}{2} \left(x^2 y^2 \frac{\partial^2 r^{\text{lr}}}{\partial x^2} + b^2 \frac{\partial^2 r^{\text{lr}}}{\partial y^2} + 2 x y b \varrho \frac{\partial^2 r^{\text{lr}}}{\partial x \partial y} \right) + \zeta^{\text{lr}} = 0 \quad (6.8)$$

on $(0, T) \times (0, \infty) \times \mathbb{R}$ with boundary condition

$$r^{\text{lr}}(T, x, y) = 0 \quad (6.9)$$

for $(x, y) \in (0, \infty) \times \mathbb{R}$. If we set $R_t^{\text{mvo}} := E \left[\int_t^T \zeta^{\text{mvo}}(s, X_s, Y_s) ds \mid \mathcal{F}_t \right]$ for $0 \leq t \leq T$ a completely analogous result holds for a function r^{mvo} with ζ^{mvo} replacing ζ^{lr} in (6.8).

Here we have used the system of equations (2.1) because for both hedging approaches the expected squared costs are computed under the real-world measure P . Note that for numerical solvers applied to (6.8) together with (6.9) the solutions to the pricing functions $v_{\hat{P}}$ and $v_{\tilde{P}}$ need to be pre-computed or at least made available at the current time step. For the H1 model with $\varrho \neq 0$ the transformed variable Z_t from (6.1) can be introduced to obtain orthogonalised equations for both hedging approaches, as has been explained for the pricing function $v_{\hat{P}}$.

To illustrate the difference in expected squared costs ($R_0^{\text{lr}} - R_0^{\text{mvo}}$) over the time interval $[0, T]$ we show in Figure 1 for the H1 model these differences using different values for the correlation parameter ϱ and time to maturity T . The absolute values of expected squared costs increase as T increases. For $T = 1.0$ and $\varrho = 0.0$ the computed values for prices and expected squared costs were $V_0(\varphi^{\text{lr}}) = V_0(\varphi^{\text{mvo}}) = 7.691$, $R_0^{\text{lr}} = 4.257$ and $R_0^{\text{mvo}} = 3.685$. For $T = 1.0$ and $\varrho = -0.5$ the computed values were $V_0(\varphi^{\text{lr}}) = V_0(\varphi^{\text{mvo}}) = 10.662$, $R_0^{\text{lr}} = 4.429$ and $R_0^{\text{mvo}} = 3.836$. Both R_0^{lr} and R_0^{mvo} tend to zero as $|\varrho|$ tends to 1, as can be expected from equations (3.19) and (4.24). This is also apparent from the fact that $|\varrho| = 1$ results in a complete market.

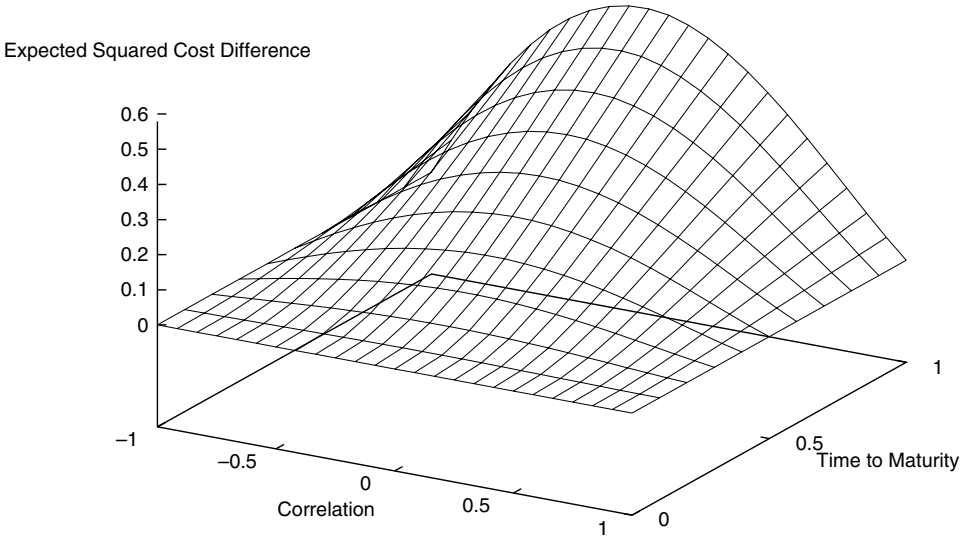


Fig. 1. Expected squared cost differences ($R_0^{\text{lr}} - R_0^{\text{mvo}}$) for the H1 model.

For increasing time to maturity T our numerical results indicate that R_0^{mvo} tends to zero. A similar remark has also been made by Hipp (1993). This observation is highlighted in Figure 2 which displays both R_0^{lr} and R_0^{mvo} over the time interval $[0, 100]$. In this sense the market can be considered as being “asymptotically complete” with respect to the mean-variance criterion. Similar results, which raise interesting questions concerning asymptotic completeness, are obtained for the other models H1, S2 and H2.

For the S2 and H2 models the drift specifications in Table 1 imply that $\hat{P} \neq \tilde{P}$ and consequently different prices are usually obtained for the two distinct measures and hedging strategies. Figure 3 illustrates these price differences for the model H2 using different values for time to maturity T and moneyness $\ln(X_0/K)$.

For at-the-money options typical price differences of the order of 2–3% were obtained. For example, with input values $T = 1.0$ and $X_0 = K = 100.0$ the computed prices were $V_0(\varphi^{\text{lr}}) = 7.6945$ and $V_0(\varphi^{\text{mvo}}) = 7.892$. However, for an out-of-the money put option with $T = 1.0$ and $\ln(X_0/K) = 0.3$ greater relative price differences were obtained with output values $V_0(\varphi^{\text{lr}}) = 0.764$ and $V_0(\varphi^{\text{mvo}}) = 0.848$. For all data points computed, local risk-minimisation prices were lower than corresponding mean-variance prices, hence the differences shown in Figure 3 are negative. This means that for the parameter set and model considered here there is no obvious best candidate when choosing between the two

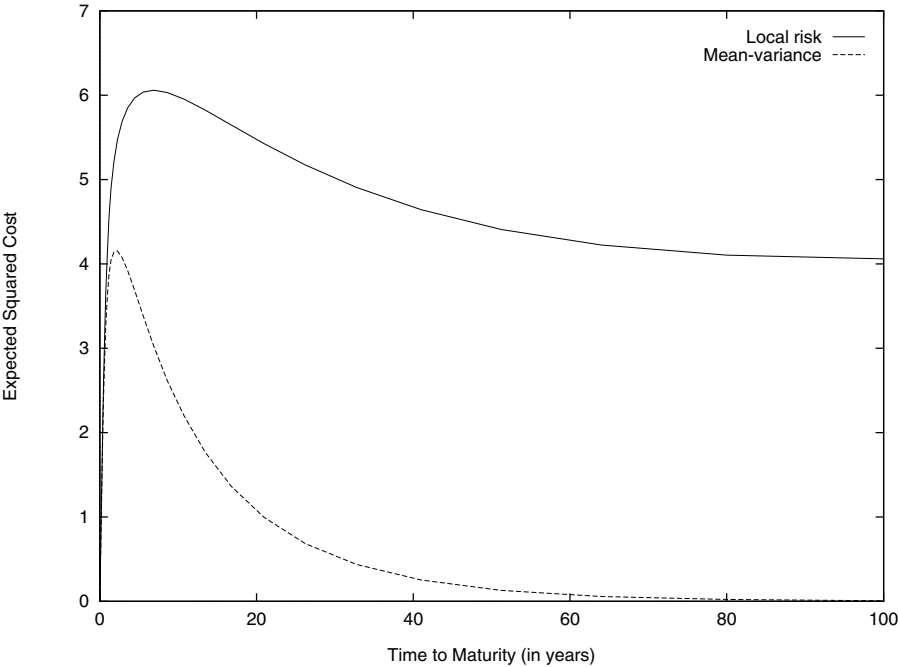


Fig. 2. Expected squared costs R_0^{lr} and R_0^{mvo} over long time periods for the S1 model.

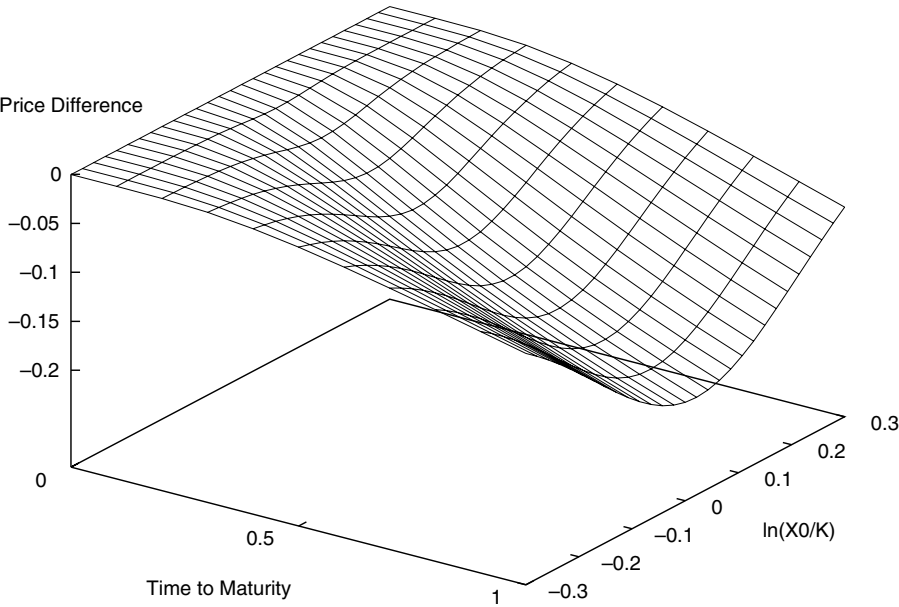


Fig. 3. Price difference $(V_0(\varphi^{\text{lr}}) - V_0(\varphi^{\text{mvo}}))$ for the H2 model.

hedging approaches. Mean-variance hedging delivers lower expected squared costs but it also results in what seem to be systematically different prices. Observe that put–call parity enforces lower prices for calls as opposed to higher prices for puts.

As is apparent from (5.3) the quantity $e^{-\Delta^2 T}$ provides a lower bound for the ratio $R_0^{\text{mvo}}/R_0^{\text{lr}}$ and the linear drift models H1 and S1. This bound is very good for small values of T ; for example, with $T = 0.01$ the computed ratio and bound for the S1 model were $R_0^{\text{mvo}}/R_0^{\text{lr}} = 0.9982$ and $e^{-\Delta^2 T} = 0.9982$. With $T = 1.0$ the corresponding values were $R_0^{\text{mvo}}/R_0^{\text{lr}} = 0.8672$ and $e^{-\Delta^2 T} = 0.7788$.

We will now consider the computation of hedge ratios ϑ^{lr} and ϑ^{mvo} for the locally risk-minimising and mean-variance optimal hedging strategies given by (3.17) and (4.24), respectively. Our aim will be to obtain approximate hedge ratios at equi-spaced discrete times $0 = t_0 < t_1 < \dots < t_N = T$ with step size $t_i - t_{i-1} = T/N$ for $i \in \{1, \dots, N\}$ using simulation techniques. Noting the form of (3.17) and (4.24) it is apparent that the price functions $v_{\hat{P}}$ and $v_{\tilde{P}}$ need to be pre-computed in order to calculate hedge ratios.

Once $v_{\hat{P}}$ and $v_{\tilde{P}}$ are determined, say on a discrete grid by a numerical solver, the partial derivatives appearing in (3.17) and (4.24) can be approximated using finite differences.

To simulate a given sample path for the vector (X, Y) under the measure P , an order 1.0 weak predictor–corrector numerical scheme, see Kloeden & Platen (1999), Section 15.5, was applied to the system of equations (2.1) to obtain a set of estimates $(\bar{X}_{t_i}, \bar{Y}_{t_i})$ for (X_{t_i}, Y_{t_i}) for $i \in \{0, \dots, N\}$ with $\bar{X}_0 = X_0$ and $\bar{Y}_0 = Y_0$. From these a set of approximate values $\bar{\vartheta}_{t_i}^{\text{lr}}$ for the hedge ratio $\vartheta_{t_i}^{\text{lr}}$ and $\bar{\xi}_{t_i}$ for the integrand $\tilde{\xi}_{t_i}$, $i \in \{0, \dots, N\}$ were obtained. One problem with this procedure is that the set of points $(t_i, \bar{X}_{t_i}, \bar{Y}_{t_i})$ for $i \in \{0, \dots, N\}$ may not lie on the grid used to compute $v_{\hat{P}}$ and $v_{\tilde{P}}$. This difficulty can be overcome by the application of multi-dimensional interpolation methods. Note that all three measures P , $\hat{P}$ and $\tilde{P}$ are used with these calculations: P is needed to simulate paths for the vector (X, Y) and $\hat{P}$ and $\tilde{P}$ are used to approximate the pricing functions $v_{\hat{P}}$ and $v_{\tilde{P}}$, respectively.

The estimates $\bar{\vartheta}_{t_i}^{\text{mvo}}$, $i \in \{0, \dots, N\}$ for the mean-variance optimal hedge ratio can now be obtained from the Euler type approximation scheme, see (4.24),

$$\bar{\vartheta}_{t_i}^{\text{mvo}} = \bar{\xi}_{t_i} + \frac{\mu(t_i, \bar{Y}_{t_i})}{\bar{X}_{t_i} \bar{Y}_{t_i}^2} \left(v_{\tilde{P}}(t_i, \bar{X}_{t_i}, \bar{Y}_{t_i}) - v_{\tilde{P}}(0, X_0, Y_0) - \sum_{j=0}^{i-1} \bar{\vartheta}_{t_j}^{\text{mvo}} (\bar{X}_{t_{j+1}} - \bar{X}_{t_j}) \right) \quad (6.10)$$

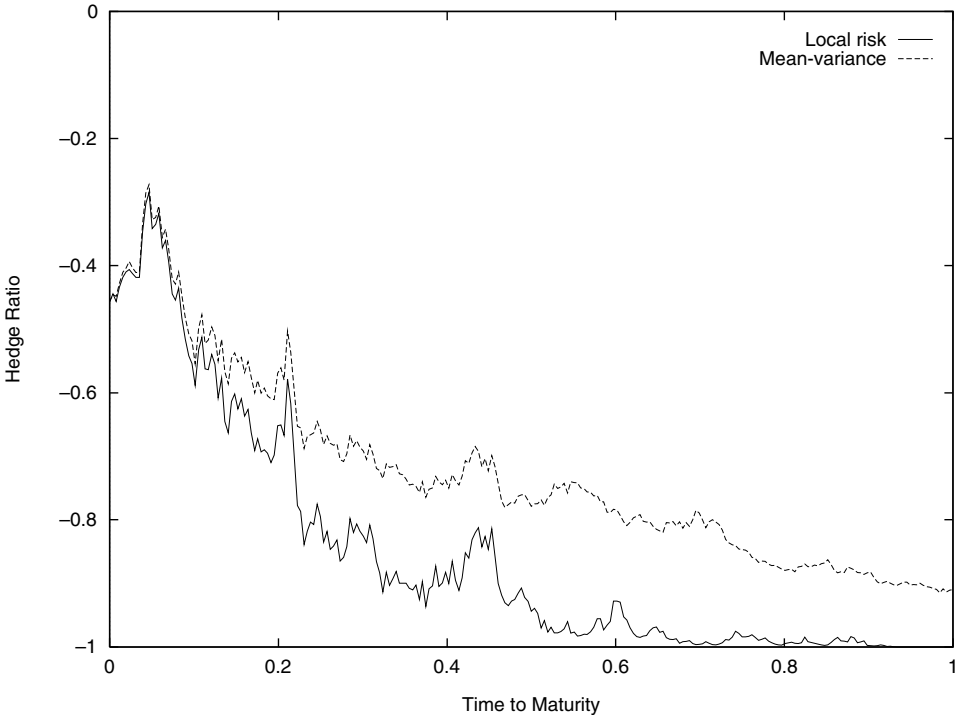


Fig. 4. Hedge ratios for the S2 model: sample path ending in the money.

for $i \in \{1, \dots, N\}$. In the case of the S2 and H2 models we have $\hat{P} \neq \tilde{P}$. In general this means that $v_{\hat{P}} \neq v_{\tilde{P}}$ and $\frac{\partial v_{\hat{P}}}{\partial x} \neq \frac{\partial v_{\tilde{P}}}{\partial x}$ and consequently it follows from (3.17), (4.21) and (4.24) with $\varrho = 0$ that for the initial hedge ratios $\bar{\vartheta}_0^{\text{lr}} \neq \bar{\vartheta}_0^{\text{mvo}}$. For models S1 and H1, since $v_{\hat{P}} = v_{\tilde{P}}$, we then get equal initial hedge ratios $\bar{\vartheta}_0^{\text{lr}} = \bar{\vartheta}_0^{\text{mvo}}$. This equality does not in general hold for $t \in (0, T)$.

Figures 4 and 5 plot the linearly interpolated hedge ratios $\bar{\vartheta}_t^{\text{lr}}$ and $\bar{\vartheta}_t^{\text{mvo}}$, $i \in \{0, \dots, N\}$, for a European put option for the S2 model. Figure 4 displays hedge ratios for a sample path ending in the money whereas Figure 5 shows hedge ratios for a different sample path ending out of the money. The trajectories for $X/100$ and Y for both sample paths are illustrated in Figure 6. Note that the mean-variance optimal hedge ratio takes values in the open interval $(0, -1)$ at maturity. This indicates that there is no full replication of the contingent claim.

In the case of the linear drift models S1 and H1 the factor $\mu(t_i, \bar{Y}_{t_i})/(\bar{X}_{t_i} \bar{Y}_{t_i}^2)$ appearing in (6.10) reduces to $\Delta/\bar{X}_{t_i} \bar{Y}_{t_i}$. This factor becomes $\gamma/\bar{X}_{t_i}$ for the quadratic drift models S2 and H2. For the given default parameter set the

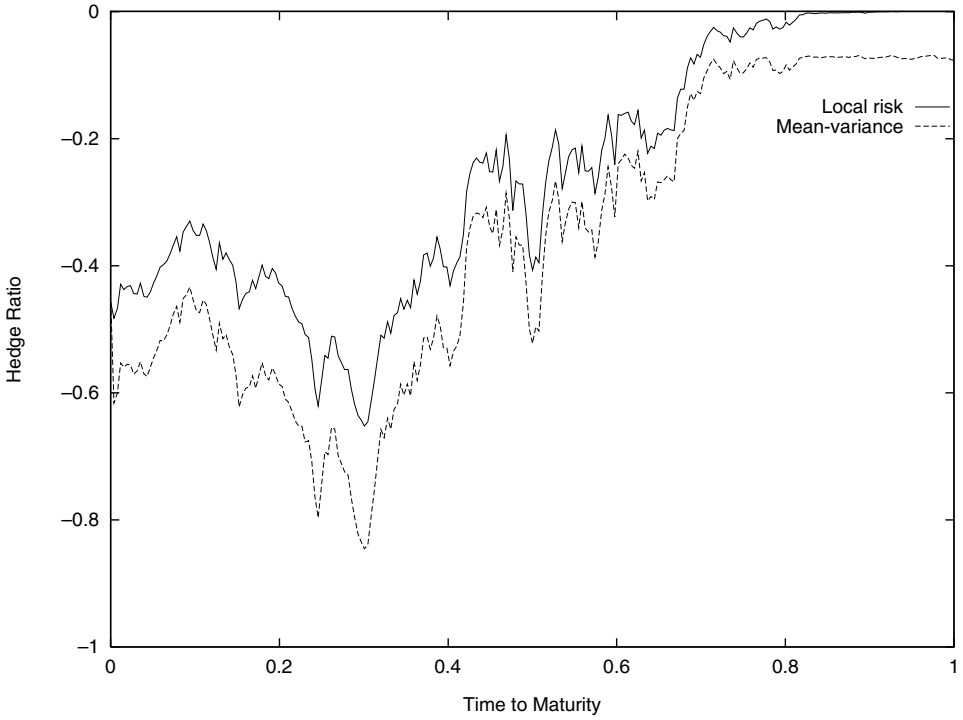


Fig. 5. Hedge ratios for the S2 model: sample path ending out of the money.

approximate volatility values $\bar{Y}_{t_i}, i \in \{0, \dots, N\}$ can be quite small. Consequently for the linear drift models large fluctuations in the mean-variance optimal hedge ratios, compared to what is obtained under the locally risk-minimising criterion, can occur. Simulation experiments have shown that these differences are not so apparent for the quadratic drift models.

7 Distributions of squared costs

So far we have examined differences in expected squared costs for the two hedging approaches. It is also interesting to consider the distributions under the real-world measure P of the quantities

$$\varepsilon_t^{\text{lr}} = \int_0^t \zeta^{\text{lr}}(s, X_s, Y_s) ds \quad (7.1)$$

and

$$\varepsilon_t^{\text{mvo}} = \int_0^t \zeta^{\text{mvo}}(s, X_s, Y_s) ds \quad (7.2)$$

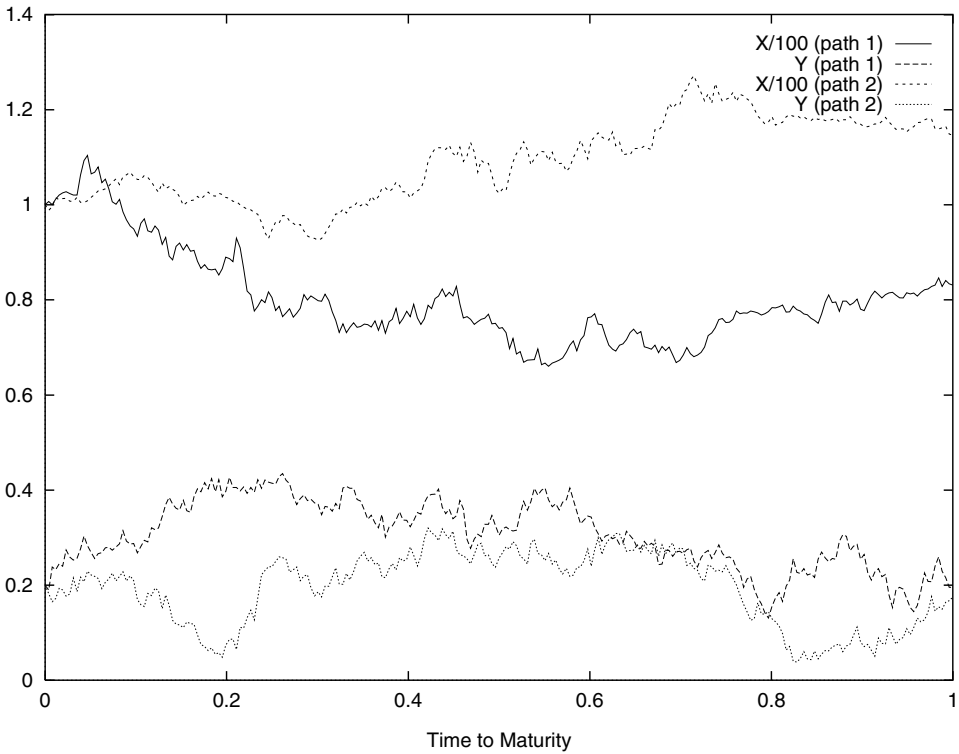


Fig. 6. Two pairs of sample paths for the S2 model.

for $0 \leq t \leq T$, where ζ^{lr} and ζ^{mvo} are given by (6.6) and (6.7), respectively. In view of (3.17) and (4.24) these terms provide a measure for the squared costs on $[0, t]$ under local risk-minimisation and mean-variance hedging, respectively. To estimate the distributions of the random variables $\varepsilon_T^{\text{lr}}$ and $\varepsilon_T^{\text{mvo}}$ we used an order 1.0 weak predictor–corrector numerical scheme, see again Kloeden & Platen (1999), Section 15.5, to obtain a set of estimates $(\bar{X}_{t_i}, \bar{Y}_{t_i})$ for (X_{t_i}, Y_{t_i}) where, as in our hedging simulation experiments, $\{t_i; i \in \{0, \dots, N\}\}$ is a set of increasing equi-spaced discrete times with $t_0 = 0$ and $t_N = T$. This enables us to compute a set of independent realisations of the random vector $(\bar{X}_{t_i}, \bar{Y}_{t_i})$ denoted by $(\bar{X}_{t_i}(\omega_j), \bar{Y}_{t_i}(\omega_j))$ for $i \in \{0, \dots, N\}$ and $j \in \{1, \dots, M\}$. From these, by applying a numerical integration routine using (7.1) and (7.2) we can generate a set of independent realisations $(\bar{\varepsilon}_T^{\text{lr}}(\omega_j), \bar{\varepsilon}_T^{\text{mvo}}(\omega_j))$ for the estimate $(\bar{\varepsilon}_T^{\text{lr}}, \bar{\varepsilon}_T^{\text{mvo}})$ of the squared costs.

We can also obtain sample path estimates of $(\varepsilon_T^{\text{lr}}, \varepsilon_T^{\text{mvo}})$ by using stochastic numerical methods applied to the full vector of components $(X, Y, \varepsilon^{\text{lr}}, \varepsilon^{\text{mvo}})$. Note

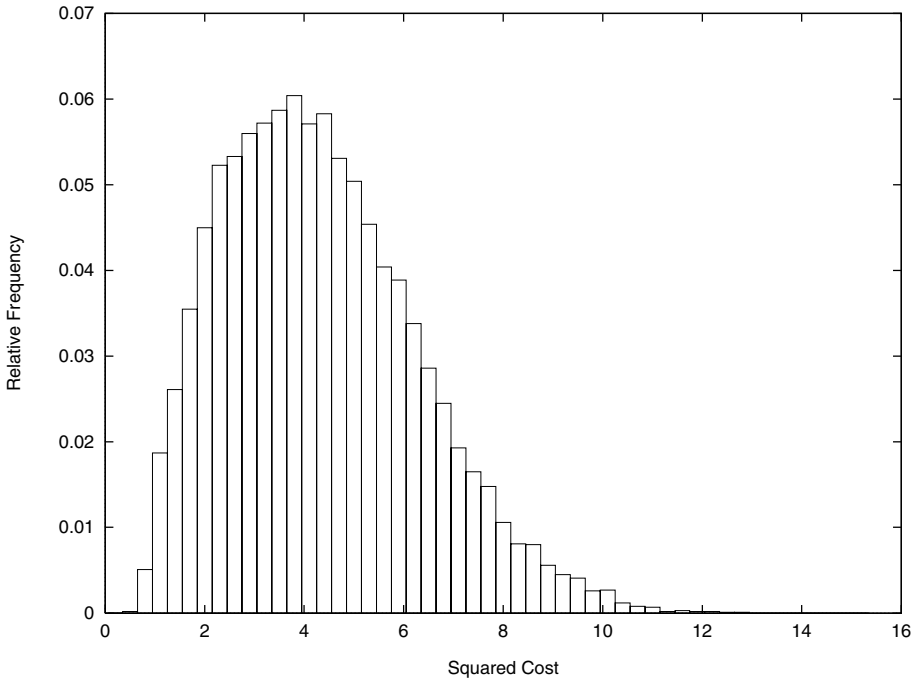


Fig. 7. Squared cost histogram of $\varepsilon_T^{\text{lr}}$ for the H1 model.

that the approximation of the integrands ζ^{lr} and ζ^{mvo} appearing in (7.1) and (7.2) requires access to the solution of the pricing functions $v_{\hat{P}}$ and $v_{\tilde{P}}$. As was the case for the computation of hedge ratios, all three measures P , $\hat{P}$ and $\tilde{P}$ are involved in these calculations and multi-dimensional interpolation is needed to obtain values for $\zeta^{\text{lr}}(t_i, \bar{X}_i, \bar{Y}_i)$ and $\zeta^{\text{mvo}}(t_i, \bar{X}_i, \bar{Y}_i)$, $i \in \{0, \dots, N\}$ along the paths of the simulated trajectories.

To obtain an estimate of the probability density function for the variates $\varepsilon_T^{\text{lr}}$ and $\varepsilon_T^{\text{mvo}}$ we use a histogram with K disjoint adjacent subintervals using the sample data $(\bar{\varepsilon}_T^{\text{lr}}(\omega_j), \bar{\varepsilon}_T^{\text{mvo}}(\omega_j))$ for $j \in \{1, \dots, M\}$. The overall procedure can be enhanced by the inclusion of anti-thetic variates for both the X and Y components of our underlying diffusion process. Figure 7 shows the histogram of relative frequencies obtained for the squared costs $\varepsilon_T^{\text{lr}}$ and the H1 model under the local risk-minimisation criterion with $N = 256$, $M = 16384$ and $K = 50$. Figure 8 shows the corresponding results for $\varepsilon_T^{\text{mvo}}$. Histograms produced for the other three model combinations S1, H2 and S2 show a slightly more symmetric form for the density function. Similar results in a jump-diffusion model have been obtained by Grünewald & Trautmann (1997).

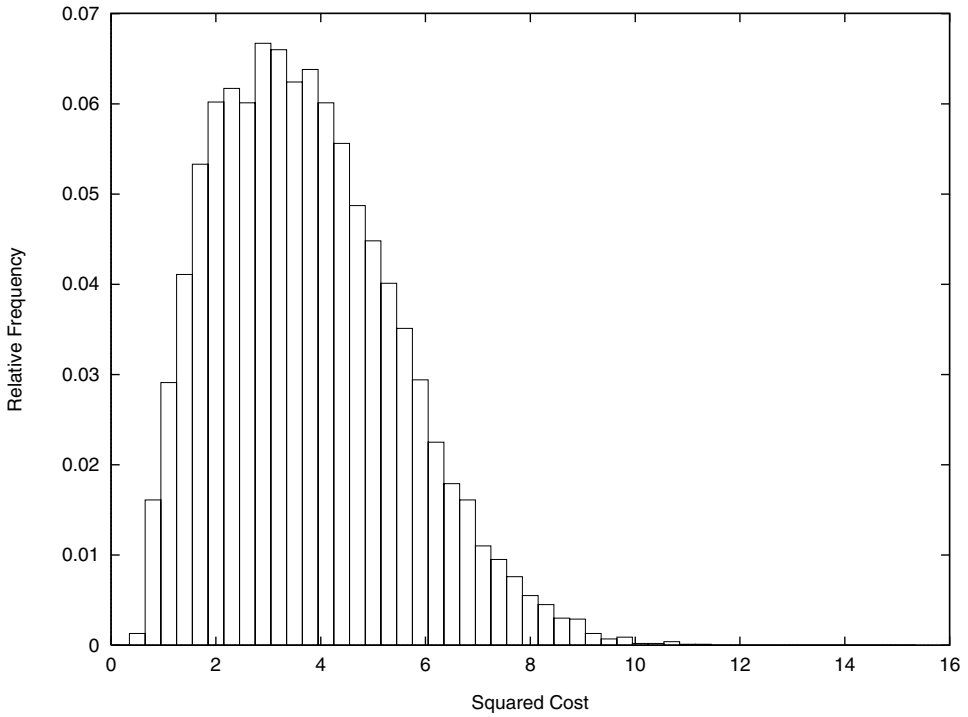


Fig. 8. Squared cost histogram of $\varepsilon_T^{\text{mvo}}$ for the H1 model.

Of course the simulated data can be also used to compute the sample means

$$\frac{1}{M} \sum_{j=1}^M \bar{\varepsilon}_T^{\text{lr}}(\omega_j) \quad \text{and} \quad \frac{1}{M} \sum_{j=1}^M \bar{\varepsilon}_T^{\text{mvo}}(\omega_j)$$

for local risk-minimisation and mean-variance hedging, respectively. These provide estimates for the expected squared costs $R_0^{\text{lr}} = E[\varepsilon_T^{\text{lr}}]$ and $R_0^{\text{mvo}} = E[\varepsilon_T^{\text{mvo}}]$ which have been previously approximated via PDE methods, see (6.8)–(6.9). Consequently our Monte Carlo simulation can also be used to check our PDE results. A summary of these results using different values for $\ln(X_0/K)$ with K fixed for the H1 model is given in Table 2.

The statistical errors reported in Table 2 were obtained at an approximate 99% confidence level. This was achieved by dividing the total number of outcomes into batches with sample means taken within each batch to form asymptotically Gaussian statistics. It is apparent from Table 2 that both methodologies produce consistent results at least within the tolerance bounds computed for the Monte Carlo estimates. As an indication of the computing power required to produce these estimates, we mention that the expected squared costs obtained from PDE

Table 2. *Expected squared cost estimates using PDEs and Monte Carlo for the H1 model.*

$\ln(X_0/K)$	PDE		Monte Carlo		Stat. error-99%	
	R_0^{lr}	R_0^{mvo}	R_0^{lr}	R_0^{mvo}	R_0^{lr}	R_0^{mvo}
0.3	0.775	0.672	0.789	0.685	0.023	0.020
0.2	1.812	1.566	1.836	1.587	0.027	0.024
0.1	3.294	2.843	3.310	2.856	0.026	0.024
0.0	4.257	3.685	4.273	3.697	0.074	0.066
-0.1	3.682	3.207	3.703	3.225	0.056	0.050
-0.2	2.278	2.003	2.293	2.016	0.025	0.022
-0.3	1.099	0.976	1.117	0.992	0.027	0.025

methods were computed in approximately 2 seconds (calculations performed on a Pentium MMX 233 MHz notebook). The Monte Carlo estimates using 16384 sample paths were computed in about 35 seconds.

8 Other numerical results

In Section 6 we considered the computation of approximate hedge ratios $\bar{\vartheta}^{\text{lr}}$ and $\bar{\vartheta}^{\text{mvo}}$ on a sample path by sample path basis. However, we would like to compare the variability of the competing hedge ratios using a more global criterion. One way of doing this is to assume proportional transaction costs.

A strategy ϑ applied at equi-spaced discrete transaction times $0 \leq t_0 < t_1 < \dots < t_N = T$ would, in addition to the pure hedging costs, incur transaction expenses

$$\lambda S_N(\vartheta)$$

for some $\lambda > 0$, where $S_N(\vartheta)$ is given by

$$S_N(\vartheta) = \sum_{i=1}^N |\vartheta_{t_i} - \vartheta_{t_{i-1}}| X_{t_i}.$$

Since ϑ will typically be of infinite variation, we expect $S_N(\vartheta)$ to diverge as $N \rightarrow +\infty$. Consequently direct comparison of $S_N(\vartheta^{\text{lr}})$ and $S_N(\vartheta^{\text{mvo}})$ is difficult as both quantities become unbounded as N becomes large. However, the transaction cost ratio

$$r_N(\vartheta^{\text{lr}}, \vartheta^{\text{mvo}}) = \frac{S_N(\vartheta^{\text{lr}})}{S_N(\vartheta^{\text{mvo}})}$$

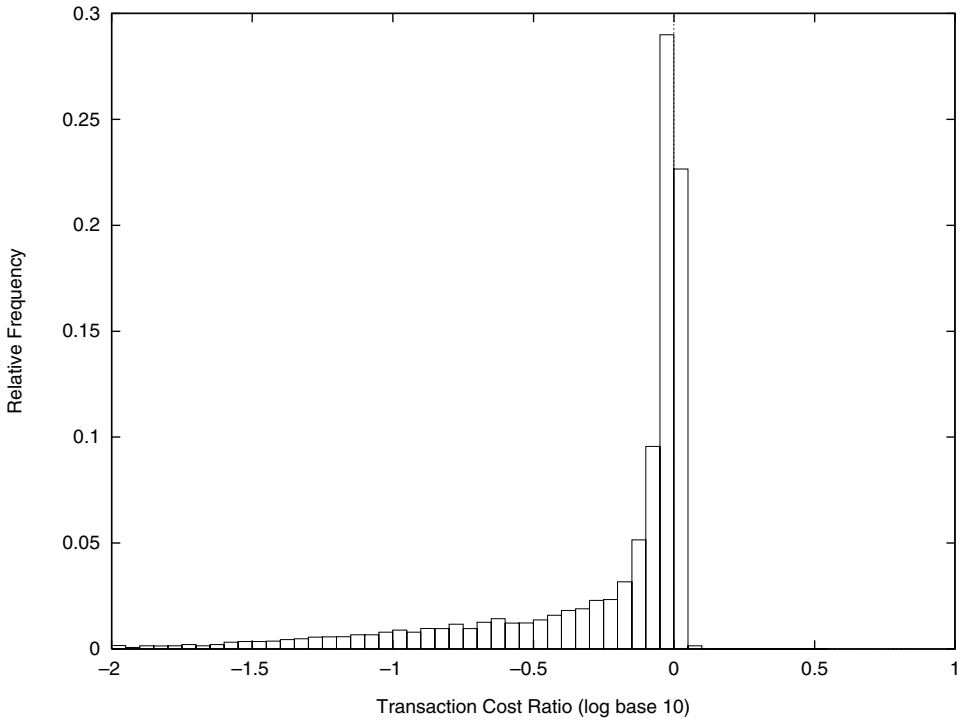


Fig. 9. Transaction cost ratio histogram of $\log_{10}(r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}}))$ for the S1 model.

can be examined and compared, at least on the basis of simulation experiments. To do this we fix N and generate approximate hedge ratios $(\bar{v}_{t_i}^{\text{lr}}, \bar{v}_{t_i}^{\text{mvo}})$, for $i \in \{0, \dots, N\}$, using the simulation methods outlined previously. These computations are performed with respect to the real-world measure P . The simulation data obtained enables us to determine $r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}})$ for a number of different sample paths and therefore to examine numerically the distributional properties of the estimate $r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}})$.

Figure 9 shows a histogram of relative frequencies for $\log_{10}(r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}}))$ and the S1 model formed with $N = 250$ transaction times and $M = 16384$ sample paths. As for our squared cost estimates, we used anti-thetic variates for each of the X and Y components in our underlying diffusion process. The value $N = 250$ corresponds approximately to daily hedging for the default time to maturity $T = 1$. Note that relative frequencies for the variable $\log_{10}(r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}}))$ rather than $r_N(\bar{v}^{\text{lr}}, \bar{v}^{\text{mvo}})$ are used. This is introduced to rescale the output so that it can be conveniently displayed in the form illustrated in Figure 9.

Figure 10 shows the corresponding histogram of relative frequencies for

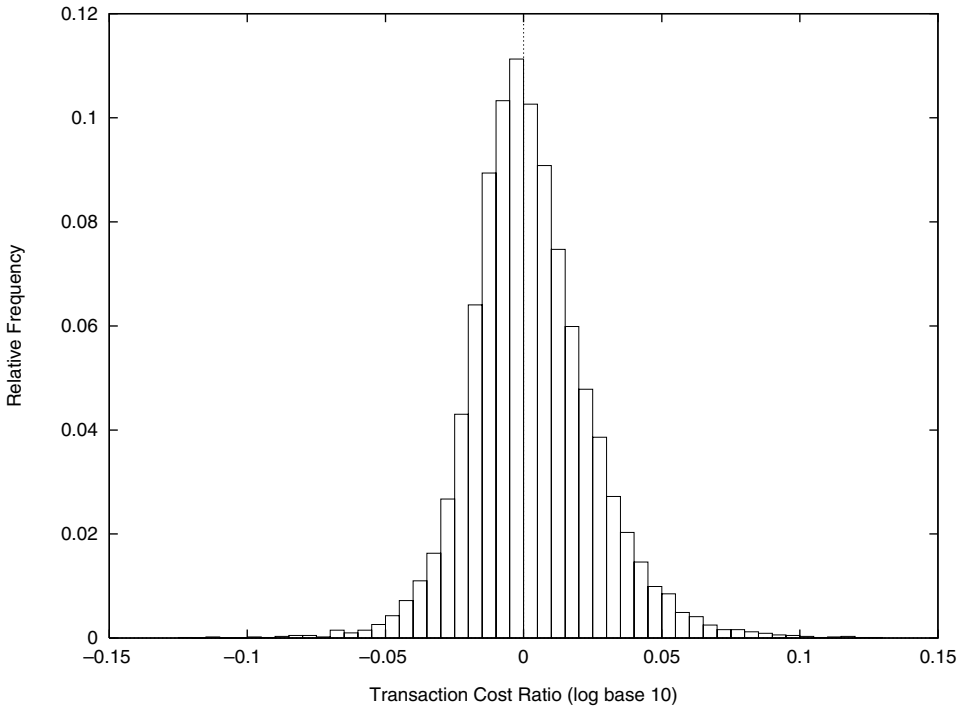


Fig. 10. Transaction cost ratio histogram of $\log_{10}(r_N(\bar{\vartheta}^{\text{lr}}, \bar{\vartheta}^{\text{mvo}}))$ for the H2 model.

$\log_{10}(r_N(\bar{\vartheta}^{\text{lr}}, \bar{\vartheta}^{\text{mvo}}))$ using the H2 model and the same transaction times and sample paths. Note that the variability of transaction cost ratios in this model is much smaller than in the first one. In Figure 9 the range of values for $\log_{10}(r_N(\bar{\vartheta}^{\text{lr}}, \bar{\vartheta}^{\text{mvo}}))$ varies from -2 to 1 whereas in Figure 10 the range is from -0.15 to 0.15 . Experimentation with the other model combinations H1 and S2 produced results which are similar to those obtained for S1 and H2 models, respectively. These results demonstrate that the distributional properties of $r_N(\bar{\vartheta}^{\text{lr}}, \bar{\vartheta}^{\text{mvo}})$ are highly dependent on our choice of the appreciation rate μ .

Experimentation with different choices of N does not seem to change these results dramatically. For example we can compute the sample mean $A(\bar{r}_N)$ of transaction cost ratios using the formula

$$A(\bar{r}_N) = \frac{1}{M} \sum_{i=1}^M \frac{S_N(\bar{\vartheta}^{\text{lr}}(\omega_j))}{S_N(\bar{\vartheta}^{\text{mvo}}(\omega_j))}.$$

Figure 11 shows the result of plotting $A(\bar{r}_N)$ for the S1, H1 and H2 models. The error-bars displayed indicate approximate confidence intervals at a 99% level. The values for the S2 model are omitted because these are very close to those for the

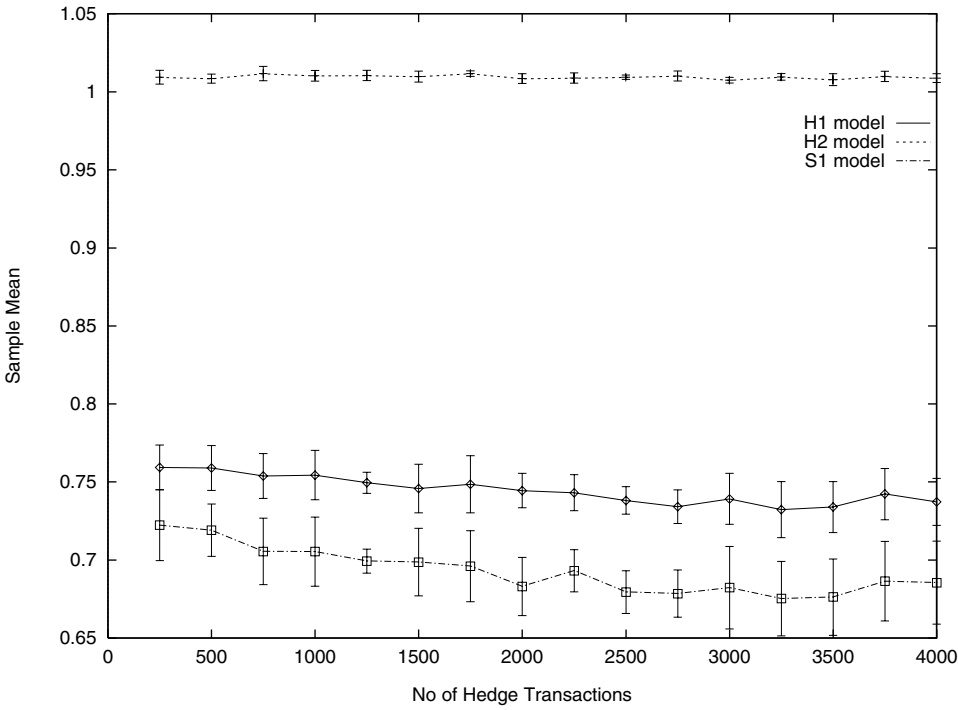


Fig. 11. Sample means and confidence intervals for $A(\bar{r}_N)$.

H2 model. The value $N = 4000$ would correspond to half-hourly hedging with an eight hour trading day and 250 trading days per year.

9 Conclusion

This chapter documents some of the differences between local risk-minimisation and mean-variance hedging for some specific stochastic volatility models. We have shown that reliable and accurate estimates for prices, hedge ratios, total expected squared costs and other quantities can be obtained for both hedging approaches. Over long time periods it seems that the mean-variance criterion leads to a form of asymptotic completeness which is not the case for local risk-minimisation. For the quadratic drift models S2 and H2 mean-variance hedging delivers lower expected squared costs and seems to change prices in a systematic way.

Relative frequency histograms of squared costs show forms which are similar for both hedging approaches, with relative frequencies for mean-variance hedging having, in general, a more compressed shape compared to those for local risk-minimisation.

However, relative frequency histograms for transaction cost ratios show highly

variable patterns which seem to depend mainly on the choice of the appreciation rate and which do not change significantly as the hedging frequency is increased.

Some of the results described in this chapter raise a number of interesting theoretical and practical issues for future research such as the assessment of long term performance and extension of the numerical methods outlined in this chapter to include more general specifications for the appreciation rate.

Acknowledgements

The authors gratefully acknowledge support by the School of Mathematical Sciences and the Faculty of Economics and Commerce of the Australian National University, the Schools of Mathematical Sciences and Finance and Economics of the University of Technology Sydney, the Fachbereich Mathematik of the Technical University of Berlin and the Deutsche Forschungsgemeinschaft.

References

- Fletcher, C.A.J. (1988), *Computational Techniques for Fluid Dynamics* (2nd ed.), Volume 1 of *Springer Ser. Comput. Phys.*, Springer.
- Föllmer, H. & Schweizer, M. (1991), Hedging of contingent claims under incomplete information. In M. Davis and R. Elliott (eds.), *Applied Stochastic Analysis*, Volume 5 of *Stochastics Monogr.*, pp. 389–414. Gordon and Breach, London/New York.
- Grünewald, B. & Trautmann, S. (1997), Varianzminimierende Hedgingstrategien für Optionen bei möglichen Kurssprüngen. *Bewertung und Einsatz von Finanzderivaten, Zeitschrift für betriebswirtschaftliche Forschung* **38**, 43–87.
- Heath, D., Platen, E. & Schweizer, M. (1998), A comparison of two quadratic approaches to hedging in incomplete markets. Preprint, Technical University of Berlin; to appear in *Mathematical Finance*.
- Heath, D. & Schweizer, M. (2000), Martingales versus PDEs in finance: An equivalence result with examples. *Journal of Applied Probability* **37**, 947–57.
- Heston, S.L. (1993), A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financial Studies* **6**(2), 327–43.
- Hipp, C. (1993), Hedging general claims. In *Proceedings of the 3rd AFIR Colloquium, Rome*, Volume 2, pp. 603–13.
- Hoffman, J.D. (1993), *Numerical Methods for Engineers and Scientists*. McGraw-Hill, Inc.
- Kloeden, P.E. & Platen, E. (1999), *Numerical Solution of Stochastic Differential Equations*, Volume 23 of *Appl. Math.*, Springer.
- Schweizer, M. (1991), Option hedging for semimartingales. *Stochastic Process. Appl.* **37**, 339–63.
- Schweizer, M. (1995), On the minimal martingale measure and the Föllmer–Schweizer decomposition. *Stochastic Anal. Appl.* **13**, 573–99.
- Stein, E.M. & Stein, J.C. (1991), Stock price distributions with stochastic volatility: An analytic approach. *Rev. Financial Studies* **4**, 727–52.

A Guided Tour through Quadratic Hedging Approaches

Martin Schweizer

0 Introduction

The goal of this chapter is to give an overview of some results and developments in the area of pricing and hedging options by means of a quadratic criterion. To put this into a broader perspective, we start in this section with some general ideas and financial motivation before turning to more precise mathematical descriptions. We remark that this borrows extensively from the financial introduction of Delbaen, Monat, Schachermayer, Schweizer and Stricker (1997).

To describe a financial market operating in continuous time, we begin with a probability space $(\Omega, \mathcal{F}, P)$, a time horizon $T \in (0, \infty)$ and a filtration $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$. Intuitively, $\mathcal{F}_t$ describes the information available at time t . We have $d + 1$ basic (primary) assets available for trade with price processes $S^i = (S_t^i)_{0 \leq t \leq T}$ for $i = 0, 1, \dots, d$. To simplify the presentation, we assume that one asset, say S^0 , has a strictly positive price. We then use S^0 as numeraire and immediately pass to quantities discounted with S^0 . This means that asset 0 has (discounted) price 1 at all times and the other assets' (discounted) prices are $X^i = S^i/S^0$ for $i = 1, \dots, d$. Without further mention, all subsequently appearing quantities will be expressed in discounted units.

One central problem of financial mathematics in such a framework is the pricing and hedging of contingent claims by means of dynamic trading strategies based on X . The best-known example of a contingent claim is a European call option on asset i with expiration date T and strike price K , say. The net payoff at T to its owner is the random amount $H(\omega) = \max(X_T^i(\omega) - K, 0) = (X_T^i(\omega) - K)^+$. More generally, a contingent claim here is simply an $\mathcal{F}_T$ -measurable random variable H describing the net payoff at T of some financial instrument. Hence our claims are of European type in the sense that the date of the payoff is fixed; but the amount to be paid may depend on the whole history of X up to time T , or even on more if $\mathbb{F}$ contains additional information. The problems of pricing and hedging

H can then be formulated as follows: what price should the seller of H charge the buyer at time 0? And having sold H , how can he insure or cover himself against the random loss at time T ?

A natural way to approach these questions is to consider dynamic portfolio strategies of the form $(\theta, \eta) = (\theta_t, \eta_t)_{0 \leq t \leq T}$, where θ is a d -dimensional predictable process and η is adapted. In such a strategy, θ_t^i describes the number of units of asset i held at time t and η_t is the amount invested in asset 0 at time t . Predictability of θ is a mathematical formulation of the informational constraint that θ is not allowed to anticipate the movement of X . At any time t , the value of the portfolio (θ_t, η_t) is given by $V_t = \theta_t^{\text{tr}} X_t + \eta_t$ and the cumulative gains from trade up to time t are $G_t(\theta) = \int_0^t \theta_s dX_s$. To have the last expression well-defined, we assume that X is a semimartingale and $G(\theta)$ is then the stochastic integral of θ with respect to X . The cumulative costs up to time t incurred by using (θ, η) are given by $C_t = V_t - \int_0^t \theta_s dX_s = V_t - G_t(\theta)$. A strategy is called self-financing if its cumulative cost process C is constant over time or equivalently if its value process V is given by

$$V_t = V_0 + \int_0^t \theta_s dX_s = V_0 + G_t(\theta), \quad (0.1)$$

where $V_0 = C_0$ is the initial outlay required to start the strategy. After time 0, such a strategy is self-supporting: any fluctuations in X can be neutralized by rebalancing θ and η in such a way that no further gains or losses result. Note that a self-financing strategy is completely described by V_0 and θ since the self-financing constraint determines V , hence also η .

Now fix a contingent claim H and suppose there exists a self-financing strategy (V_0, θ) whose terminal value V_T equals H with probability one. If our financial market model does not allow arbitrage opportunities, it is clear that the price of H must be given by V_0 and that θ furnishes a hedging strategy against H . This was the basic insight leading to the celebrated Black–Scholes formula for option pricing; see Black and Scholes (1973) and Merton (1973) who solved this problem for the case where X is a one-dimensional geometric Brownian motion and $H = (X_T - K)^+$ is a European call option. The mathematical structure of the problem and its connections to martingale theory were subsequently worked out and clarified by J. M. Harrison and D. M. Kreps; a detailed account can be found in Harrison and Pliska (1981). Following their terminology, we call a contingent claim H attainable if there exists a self-financing strategy with $V_T = H$ P -a.s. By (0.1), this means that H can be written as

$$H = H_0 + \int_0^T \theta_s^H dX_s \quad P\text{-a.s.}, \quad (0.2)$$

i.e., as the sum of a constant H_0 and a stochastic integral with respect to X . We speak of a complete market if every contingent claim is attainable. Recall that we do not give precise definitions here; for a rigorous mathematical formulation, one has to be rather careful about the integrability conditions imposed on H and θ^H .

The importance of the concept of a complete market stems from the fact that it allows the pricing and hedging of contingent claims to be done in a preference-independent fashion. However, completeness is a rather delicate property which is typically destroyed as soon as one considers even minor modifications of a basic complete model. For instance, geometric Brownian motion (the classical Black–Scholes model) becomes incomplete if the volatility is influenced by a second stochastic factor or if one adds a jump component to the model. If one insists on a preference-free approach under incompleteness, one can study the range of possible prices for H which are consistent with absence of arbitrage in a market containing X , the riskless asset 1 and H as traded instruments; this is the idea behind the concept of super-replication. An alternative is to introduce subjective criteria according to which strategies are chosen and option prices are computed. The goal of this chapter is to explain two such criteria in more detail. For a very recent similar survey, see also Pham (2000). A numerical comparison study can be found in chapter 14 of this book.

For a non-attainable contingent claim, it is by definition impossible to find a strategy with final value $V_T = H$ which is at the same time self-financing. A first possible approach is to insist on the terminal condition $V_T = H$; since η is allowed to be adapted, this can always be achieved by choice of η_T . But because such strategies cannot be self-financing in general, a “good” strategy should now have a “small” cost process C . Measuring the riskiness of a strategy by a quadratic criterion was first proposed by Föllmer and Sondermann (1986) for the case where X is a martingale and subsequently extended to the general semimartingale case in Schweizer (1988, 1991). Under some technical assumptions, such a *locally risk-minimizing* strategy can be characterized by two properties: its cost process C must be a martingale (so that the strategy is no longer self-financing, but still remains mean-self-financing) and this martingale must be orthogonal to the martingale part M of the price process X . Translating this into conditions on the contingent claim H shows that there exists a locally risk-minimizing strategy for H if and only if H admits a decomposition of the form

$$H = H_0 + \int_0^T \theta_s^H dX_s + L_T^H \quad P\text{-a.s.}, \quad (0.3)$$

where L^H is a martingale orthogonal to M . The decomposition (0.3) has been called the *Föllmer–Schweizer decomposition* of H ; it can be viewed as a general-

ization to the semimartingale case of the classical Galtchouk–Kunita–Watanabe decomposition from martingale theory. Its financial importance lies in the fact that it directly provides the locally risk-minimizing strategy for H : the stock component θ is given by the integrand θ^H and η is determined by the requirement that the cost process C should coincide with $H_0 + L^H$. Note also that the special case (0.2) of an attainable claim simply corresponds to the absence of the orthogonal term L_T^H . In particular cases, one can give more explicit constructions for the decomposition (0.3). In the case of finite discrete time, θ^H and L^H can be computed recursively backward in time. If X is continuous, the Föllmer–Schweizer decomposition under P can be obtained as a Galtchouk–Kunita–Watanabe decomposition, computed under the so-called *minimal martingale measure* $\hat{P}$.

One drawback of the preceding approach is the fact that one has to work with strategies which are not self-financing. If one prefers to avoid intermediate costs or an unplanned income, a second idea is to insist on the self-financing constraint (0.1). The possible final outcomes of such strategies are of the form $V_0 + G_T(\theta)$ for some initial capital $V_0 \in \mathbb{R}$ and some θ in the set Θ , say, of all integrands allowed in (0.1). By definition, a non-attainable claim H is not of this form and so it seems natural to look for a best approximation of H by the terminal value $V_0 + G_T(\theta)$ of some pair (V_0, θ) . The use of a quadratic criterion to measure the quality of this approximation has been proposed by Bouleau and Lamberton (1989) if X is both a martingale and a function of a Markov process, and by Duffie and Richardson (1991) and Schweizer (1994a), among others, in more general cases. To find such a *mean-variance optimal* strategy, one has to project H in $L^2(P)$ on the space $\mathbb{R} + G_T(\Theta)$ of attainable claims. In particular, this raises the questions of whether the space $G_T(\Theta)$ of stochastic integrals of X is closed in $L^2(P)$ and what the structure of the corresponding projection is. Both these problems as well as the computation of the optimal initial capital V_0 turn out to be intimately linked to the so-called *variance-optimal martingale measure* $\tilde{P}$.

The chapter is structured as follows. Section 1 introduces some general notations and recalls a few preliminaries to complement the preceding discussion. Section 2 explains the above two approaches in the case where X is a local martingale under P ; this slightly generalizes the classical results due to Föllmer and Sondermann (1986). Section 3 discusses local risk-minimization in detail and the final Section 4 is devoted to mean-variance hedging.

1 Notations and preliminaries

In this section, we briefly introduce some notation for later use. This complements the introduction by giving precise definitions. For all standard terminology from martingale theory, we refer to Dellacherie and Meyer (1982).

Mathematically, the basic asset prices are defined on a probability space $(\Omega, \mathcal{F}, P)$ and described by the constant 1 and an $\mathbb{R}^d$ -valued stochastic process $X = (X_t)_{0 \leq t \leq T}$ adapted to a filtration $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$ satisfying the usual conditions of right-continuity and completeness. Adaptedness ensures that time t prices X_t are $\mathcal{F}_t$ -measurable, i.e., observable at time t . To exclude arbitrage opportunities, we assume that X admits an *equivalent local martingale measure* (ELMM) Q , i.e., that there exists a probability measure $Q \approx P$ such that X is a local Q -martingale. With $\mathbb{P}$ denoting the convex set of all ELMMs Q for X , we thus assume that $\mathbb{P} \neq \emptyset$. *Incompleteness* of the market given by X and $\mathbb{F}$ is in our context taken to mean that $\mathbb{P}$ contains more than one element (and therefore infinitely many). Finally, a European type *contingent claim* is an $\mathcal{F}_T$ -measurable random variable H ; it describes a random payoff to be made at time T . Before we go on with the general theory, it may be useful to illustrate the preceding concepts by a simple example.

Example Consider one risky asset ($d = 1$) with price process X and stochastic volatility Y . More precisely, let X and Y satisfy the stochastic differential equations

$$\begin{aligned} \frac{dX_t}{X_t} &= \mu(t, X_t, Y_t) dt + Y_t dW_t^1, \\ dY_t &= a(t, X_t, Y_t) dt + b(t, X_t, Y_t) dW_t^2 \end{aligned}$$

with suitable coefficient functions μ, a, b and independent Brownian motions W^1, W^2 . The filtration $\mathbb{F}$ is the one generated by W^1 and W^2 , made complete and right-continuous. A simple example of a contingent claim here is a European call option on X with strike K and maturity T ; its (net) payoff at time T is $H = (X_T - K)^+$. Note, however, that our abstract framework encompasses much more general (e.g., path-dependent) payoffs and unlike the present example usually assumes no Markovian structure.

In this example, weak assumptions on μ, a, b readily guarantee the existence of an ELMM Q . In fact, it is enough to be able to remove the drift μ by a Girsanov transformation. This uniquely determines the transformation's effect on W^1 , but imposes no restrictions on the Q -drift of W^2 . Hence there is no unique ELMM and we have an incomplete market. This is also intuitively clear because there are two sources of uncertainty W^1, W^2 , but (by assumption) only one risky asset X for trade. If Y or some other suitable asset were also tradeable, the situation would be different. This ends the present discussion of the example.

Given a contingent claim H , there are at least two things a potential seller of H may want to do: *pricing* by assigning a value to H at times $t < T$ and *hedging* by covering himself against future losses arising from a sale of H . The notion of hedge-

ing brings up the idea of trading in X and we formalize this by introducing *trading strategies*. Note first that our assumption $\mathbb{P} \neq \emptyset$ implies that X is a semimartingale under P . It thus makes sense to speak of stochastic integrals with respect to X and we denote by $L(X)$ the linear space of all $\mathbb{R}^d$ -valued predictable X -integrable processes θ ; see Dellacherie and Meyer (1982) for additional information. For $\theta \in L(X)$, the stochastic integral $\int \theta dX$ is well-defined, but some elements of $L(X)$ are too general to yield economically reasonable strategies. We shall have to impose integrability assumptions later and so we use for the moment the term “pre-strategy”.

Definition A *self-financing pre-strategy* is any pair (V_0, θ) such that $\theta \in L(X)$ and V_0 is an $\mathcal{F}_0$ -measurable random variable. Intuitively, one starts out with initial capital V_0 and then holds the dynamically varying number θ_t^i of shares of asset i at time t . The self-financing condition implies that the *value process* of (V_0, θ) is given by

$$V_t(V_0, \theta) := V_0 + \int_0^t \theta_u dX_u, \quad 0 \leq t \leq T. \quad (1.1)$$

2 The martingale case

We first discuss the two basic quadratic hedging approaches in the simple special case where X is a *local P -martingale*; this means that the original measure P itself is in $\mathbb{P}$. We denote by $[X] = ([X^i, X^j])_{i,j=1,\dots,d}$ the matrix-valued optional covariance process of X and by $L^2(X)$ the space of all $\mathbb{R}^d$ -valued predictable processes θ such that

$$\|\theta\|_{L^2(X)} := \left(E \left[\int_0^T \theta_u^{\text{tr}} d[X]_u \theta_u \right] \right)^{\frac{1}{2}} < \infty.$$

Our first result shows that the stochastic integral of θ with respect to X is well-defined for $\theta \in L^2(X)$ and has nice properties even if X is not locally square-integrable. This is because the required integrability is already built into the definition of $L^2(X)$. I thank C. Stricker for providing the proof given below.

Lemma 2.1 *Suppose that X is a local P -martingale. For any $\theta \in L^2(X)$, the process $\int \theta dX$ is well-defined and in the space $\mathcal{M}_0^2(P)$ of square-integrable P -martingales null at 0. Moreover, the space $\mathcal{I}^2(X) := \{\int \theta dX \mid \theta \in L^2(X)\}$ of stochastic integrals is a stable subspace of $\mathcal{M}_0^2(P)$.*

Proof For $\theta \in L^2(X)$, the process $\int \theta^{\text{tr}} d[X] \theta$ is integrable. Hence $\int \theta dX$ is well-defined and a local P -martingale by Theorem 4.60 of Jacod (1979), and the

Burkholder–Davis–Gundy inequality implies that $\int \theta dX$ is even in $\mathcal{M}_0^2(P)$. It is clear that $\mathcal{I}^2(X)$ is a linear subspace of $\mathcal{M}_0^2(P)$ and stable under stopping. If $Y^n = \int \theta^n dX$ is a sequence in $\mathcal{I}^2(X)$ converging to some Y in $\mathcal{M}_0^2(P)$, then Y^n also converges to Y in $\mathcal{M}_0^1(P)$ and so Corollary 2.5.2 of Yor (1978) or Corollary 4.23 of Jacod (1979) (plus Remark III.2 in Stricker (1990) to account for the fact that X is multidimensional) imply that $Y = \int \psi dX$ for some $\psi \in L(X)$. Since

$$\int_0^T (\theta_u^n - \psi_u)^{\text{tr}} d[X]_u (\theta_u^n - \psi_u) = [Y^n - Y]_T$$

converges to 0 in $L^1(P)$ by the convergence of Y^n to Y in $\mathcal{M}_0^2(P)$, we obtain that ψ is in $L^2(X)$. Hence $Y \in \mathcal{I}^2(X)$, so $\mathcal{I}^2(X)$ is closed in $\mathcal{M}_0^2(P)$ and this completes the proof. $\square$

Definition An *RM-strategy* is any pair $\phi = (\theta, \eta)$ where $\theta \in L^2(X)$ and $\eta = (\eta_t)_{0 \leq t \leq T}$ is a real-valued adapted process such that the *value process* $V(\phi) := \theta^{\text{tr}} X + \eta$ is right-continuous and square-integrable (i.e., $V_t(\phi) \in L^2(P)$ for each $t \in [0, T]$).

Intuitively, θ_t^i and η_t denote as before the respective numbers of shares of assets i and 0 held at time t . (The notation RM anticipates that we shall want to focus on risk-minimization.) But in contrast to Section 1, we now also admit strategies that are not self-financing and thus may generate profits or losses over time.

Definition For any RM-strategy ϕ , the (cumulative) *cost process* $C(\phi)$ is defined by

$$C_t(\phi) := V_t(\phi) - \int_0^t \theta_u dX_u, \quad 0 \leq t \leq T.$$

$C_t(\phi)$ describes the total costs incurred by ϕ over the interval $[0, t]$; note that these arise from trading because of the fluctuations of the price process X and are not due to transaction costs. The *risk process* of ϕ is defined by

$$R_t(\phi) := E \left[(C_T(\phi) - C_t(\phi))^2 \mid \mathcal{F}_t \right], \quad 0 \leq t \leq T.$$

Since a contingent claim H is $\mathcal{F}_T$ -measurable and η is allowed to be adapted, we can always find RM-strategies with $V_T(\phi) = H$ provided that $H \in L^2(P)$. The simplest is “wait, then pay” where $\theta \equiv 0$ and $\eta_t = H I_{\{t=T\}}$. But in general, these strategies will not be self-financing; in fact, (1.1) tells us that there is a self-financing RM-strategy ϕ with $V_T(\phi) = H$ if and only if H admits a representation as the sum of an $\mathcal{F}_0$ -measurable random variable and a stochastic integral with respect to X . In that case, the cost process $C(\phi)$ is constant and the risk process $R(\phi)$ is identically 0. For claims where this is not possible, the idea of Föllmer

and Sondermann (1986) in defining *risk-minimization* is to look among all RM-strategies with $V_T(\phi) = H$ for one which minimizes the risk process in a suitable sense.

Definition An RM-strategy ϕ is called *risk-minimizing* if for any RM-strategy $\tilde{\phi}$ such that $V_T(\tilde{\phi}) = V_T(\phi)$ P -a.s., we have

$$R_t(\phi) \leq R_t(\tilde{\phi}) \quad P\text{-a.s. for every } t \in [0, T].$$

This is not the original definition, but it amounts to the same thing:

Lemma 2.2 *An RM-strategy ϕ is risk-minimizing if and only if*

$$R_t(\phi) \leq R_t(\tilde{\phi}) \quad P\text{-a.s.}$$

for every $t \in [0, T]$ and for every RM-strategy $\tilde{\phi}$ which is an admissible continuation of ϕ from t on in the sense that $V_T(\tilde{\phi}) = V_T(\phi)$ P -a.s., $\tilde{\theta}_s = \theta_s$ for $s \leq t$ and $\tilde{\eta}_s = \eta_s$ for $s < t$.

Proof See Lemma 2.1 of Schweizer (1994b); this does not use that X is a local P -martingale. $\square$

Remark The definition in Föllmer and Sondermann (1986) of an admissible continuation of ϕ from t on is more symmetric because they stipulate that $\tilde{\theta}_s = \theta_s$ and $\tilde{\eta}_s = \eta_s$ both hold for $s < t$. In the martingale case and for continuous time, this difference does not matter, but a discrete-time setting or the subsequent generalization to local risk-minimization do need the asymmetric formulation in Lemma 2.2. This also reflects the asymmetry between the requirements on θ and η since θ must be predictable while η is allowed to be adapted.

Although RM-strategies with $V_T(\phi) = H$ will in general not be self-financing, it turns out that good RM-strategies are still “self-financing on average” in the following sense.

Definition An RM-strategy ϕ is called *mean-self-financing* if its cost process $C(\phi)$ is a P -martingale.

Lemma 2.3 *Any risk-minimizing RM-strategy ϕ is also mean-self-financing.*

Proof This proof does not use that X is a local P -martingale. Fix $t_0 \in [0, T]$ and define $\tilde{\phi}$ by setting $\tilde{\theta} := \theta$ and

$$\tilde{\theta}_t^{\text{tr}} X_t + \tilde{\eta}_t = V_t(\tilde{\phi}) := V_t(\phi) I_{[0, t_0)}(t) + E \left[V_T(\phi) - \int_t^T \theta_u dX_u \middle| \mathcal{F}_t \right] I_{[t_0, T]}(t),$$

choosing an RCLL version. Then $\tilde{\phi}$ is an RM-strategy with $V_T(\tilde{\phi}) = V_T(\phi)$ and because $C_T(\phi) = C_T(\tilde{\phi})$ and $C_{t_0}(\tilde{\phi}) = E[C_T(\tilde{\phi})|\mathcal{F}_{t_0}]$,

$$C_T(\phi) - C_{t_0}(\phi) = C_T(\tilde{\phi}) - C_{t_0}(\tilde{\phi}) + E[C_T(\tilde{\phi})|\mathcal{F}_{t_0}] - C_{t_0}(\phi)$$

implies that

$$R_{t_0}(\phi) = R_{t_0}(\tilde{\phi}) + (C_{t_0}(\phi) - E[C_T(\phi)|\mathcal{F}_{t_0}])^2.$$

Because ϕ is risk-minimizing, we conclude that

$$C_{t_0}(\phi) = E[C_T(\phi)|\mathcal{F}_{t_0}] \quad P\text{-a.s.}$$

and since t_0 is arbitrary, the assertion follows. $\square$

The key result for finding risk-minimizing RM-strategies is the well-known *Galtchouk–Kunita–Watanabe decomposition*. Because $\mathcal{I}^2(X)$ is a stable subspace of $\mathcal{M}_0^2(P)$, any $H \in L^2(\mathcal{F}_T, P)$ can be uniquely written as

$$H = E[H|\mathcal{F}_0] + \int_0^T \theta_u^H dX_u + L_T^H \quad P\text{-a.s.} \quad (2.1)$$

for some $\theta^H \in L^2(X)$ and some $L^H \in \mathcal{M}_0^2(P)$ which is strongly orthogonal to $\mathcal{I}^2(X)$; this means that $L^H \int \theta dX$ is a P -martingale for every $\theta \in L^2(X)$. The next result was obtained by Föllmer and Sondermann (1986) for $d = 1$ under the assumption that X is in $\mathcal{M}^2(P)$. The observation and proof that it holds for a general local P -martingale X seem to be new.

Theorem 2.4 *Suppose that X is a local P -martingale. Then every contingent claim $H \in L^2(\mathcal{F}_T, P)$ admits a unique risk-minimizing RM-strategy ϕ^* with $V_T(\phi^*) = H$ P -a.s. In terms of the decomposition (2.1), ϕ^* is explicitly given by*

$$\begin{aligned} \theta^* &= \theta^H, \\ V_t(\phi^*) &= E[H|\mathcal{F}_t] =: V_t^*, \quad 0 \leq t \leq T, \\ C(\phi^*) &= E[H|\mathcal{F}_0] + L^H. \end{aligned}$$

Proof Note first that the above prescription defines an RM-strategy ϕ^* with $V_T(\phi^*) = H$. Now fix $t \in [0, T]$ and any RM-strategy $\tilde{\phi}$ with $V_T(\tilde{\phi}) = H$. The same argument as in the proof of Lemma 2.3 shows that we may assume $C_t(\tilde{\phi}) = E[C_T(\tilde{\phi})|\mathcal{F}_t]$ and so we get

$$C_T(\tilde{\phi}) - C_t(\tilde{\phi}) = H - \int_t^T \tilde{\theta}_u dX_u - E[H|\mathcal{F}_t] = L_T^H - L_t^H + \int_t^T (\theta_u^H - \tilde{\theta}_u) dX_u$$

by using (2.1) and the martingale property of $\int \tilde{\theta} dX$. Because $C(\phi^*) = C_0(\phi^*) + L^H$, the orthogonality of L^H and $\mathcal{I}^2(X)$ yields

$$R_t(\tilde{\phi}) = R_t(\phi^*) + E \left[\left(\int_t^T (\theta_u^H - \tilde{\theta}_u) dX_u \right)^2 \middle| \mathcal{F}_t \right] \geq R_t(\phi^*).$$

Hence ϕ^* is risk-minimizing. If some other $\tilde{\phi}$ is also risk-minimizing, then $C(\tilde{\phi})$ must be a martingale by Lemma 2.3 and then the same argument as before gives for $t = 0$

$$R_0(\tilde{\phi}) = R_0(\phi^*) + E \left[\int_0^T (\theta_u^H - \tilde{\theta}_u)^{\text{tr}} d[X]_u (\theta_u^H - \tilde{\theta}_u) \middle| \mathcal{F}_0 \right].$$

Because $\tilde{\phi}$ is risk-minimizing, this implies $\tilde{\theta} = \theta^H = \theta^*$ and since $C(\tilde{\phi})$ is a martingale and $V_T(\tilde{\phi}) = V_T(\phi^*)$, we also obtain $\tilde{\phi} = \phi^*$. $\square$

Remark The preceding approach relies heavily on the fact that the contingent claim H only makes one payment at the terminal date T . For applications to insurance derivatives as in Møller (1998a), this is not sufficient because such products involve possible payments at any time $t \in [0, T]$. An extension of the risk-minimization concept to the case of such payment streams has been developed in Møller (1998b).

An alternative quadratic approach in the martingale case has been studied by Bouleau and Lamberton (1989). They imposed the additional condition that X is a function of some Markov process to get more explicit results, but their basic idea can also be explained in our general framework. Suppose that instead of insisting on $V_T(\phi) = H$ P -a.s., we focus on *self-financing* RM-strategies. Such a strategy is described by a pair (V_0, θ) in $L^2(\mathcal{F}_0, P) \times L^2(X)$ and its shortfall at the terminal date T is

$$H - V_T(V_0, \theta) = H - V_0 - \int_0^T \theta_u dX_u.$$

If H is attainable by such a strategy in the sense that $H = V_T(V_0, \theta)$ for some pair (V_0, θ) , the shortfall can be reduced to 0. But in general, one has a residual risk of

$$J_0(V_0, \theta) := E \left[(H - V_T(V_0, \theta))^2 \right]$$

if one uses a quadratic loss function, and the idea of Bouleau and Lamberton (1989) is to minimize this residual risk by choice of (V_0, θ) . This clearly amounts to projecting the random variable H in $L^2(P)$ on the linear space spanned by $L^2(\mathcal{F}_0, P)$ and the stochastic integrals $\int_0^T \theta_u dX_u$ with $\theta \in L^2(X)$ and, thanks to (2.1), the

solution is given by

$$\begin{aligned}\bar{V}_0 &= [H|\mathcal{F}_0], \\ \bar{\theta} &= \theta^H\end{aligned}$$

with a minimal residual risk of

$$J_0(\bar{V}_0, \bar{\theta}) = E\left[(L_T^H)^2\right] = \text{Var}\left[L_T^H\right].$$

In the next two sections, we generalize the preceding two approaches to the case where X under P is no longer a local martingale, but only a semimartingale. Risk-minimization will be replaced by local risk-minimization and extending the above projection approach leads to mean-variance hedging. We shall also see that extensions of the Galtchouk–Kunita–Watanabe decomposition play an important role and that it is often very helpful to work with a suitably chosen ELMM.

3 Local risk-minimization

Let us now consider the general situation where the original measure P is not in $\mathbb{P}$. Hence X is no longer a local P -martingale, but only a semimartingale under P . Given a contingent claim H , we could still look for risk-minimizing strategies ϕ with $V_T(\phi) = H$. But there is bad news:

Proposition 3.1 *If X is not a local P -martingale, a contingent claim H admits in general no risk-minimizing strategy ϕ with $V_T(\phi) = H$ P -a.s.*

Proof We show this by presenting an explicit counterexample given in Schweizer (1988). For simplicity, we work in discrete time. Let $X = (X_k)_{k=0,1,\dots,T}$ (with $T \in \mathbb{N}$) be a real-valued square-integrable process adapted to a filtration $\mathbb{F} = (\mathcal{F}_k)_{k=0,1,\dots,T}$ and fix $H \in L^2(\mathcal{F}_T, P)$. The example below is on a finite probability space so that all integrability requirements are satisfied.

If ϕ^* is a risk-minimizing strategy with $V_T(\phi^*) = H$ P -a.s., Lemma 2.3 implies that $C(\phi^*)$ is a P -martingale so that we get

$$R_k(\phi^*) = \text{Var}[C_T(\phi^*)|\mathcal{F}_k] = \text{Var}\left[H - \sum_{j=k+1}^T \theta_j^* \Delta X_j \mid \mathcal{F}_k\right]$$

by using $V_T(\phi^*) = H$ and omitting $\mathcal{F}_k$ -measurable terms from the conditional variance. By $\Delta X_j := X_j - X_{j-1}$, we denote the increment of X from $j-1$ to j .

Moreover,

$$\begin{aligned}\theta_k^* X_k + \eta_k^* &= V_k(\phi^*) = C_k(\phi^*) + \sum_{j=1}^k \theta_j^* \Delta X_j \\ &= E \left[H - \sum_{j=1}^T \theta_j^* \Delta X_j \middle| \mathcal{F}_k \right] + \sum_{j=1}^k \theta_j^* \Delta X_j\end{aligned}$$

shows that ϕ^* is uniquely determined by the predictable process θ^* and *vice versa*. Because ϕ^* is risk-minimizing, any mean-self-financing strategy ϕ with $V_T(\phi) = H$ will satisfy

$$\text{Var} \left[H - \sum_{j=k+1}^T \theta_j \Delta X_j \middle| \mathcal{F}_k \right] = R_k(\phi) \geq R_k(\phi^*) = \text{Var} \left[H - \sum_{j=k+1}^T \theta_j^* \Delta X_j \middle| \mathcal{F}_k \right].$$

In particular, this implies that the mapping

$$\theta_{k+1} \mapsto \text{Var} \left[H - \theta_{k+1} \Delta X_{k+1} - \sum_{j=k+2}^T \theta_j^* \Delta X_j \middle| \mathcal{F}_k \right]$$

attains its minimum at θ_{k+1}^* and so the first order condition for this problem yields

$$\theta_{k+1}^* = \frac{\text{Cov} \left(H - \sum_{j=k+2}^T \theta_j^* \Delta X_j, \Delta X_{k+1} \middle| \mathcal{F}_k \right)}{\text{Var}[\Delta X_{k+1} | \mathcal{F}_k]}. \quad (3.1)$$

This backward recursive expression determines a unique candidate for a risk-minimizing strategy ϕ^* .

For the counterexample, we take $T = 2$ and consider a random walk X starting at 0 whose (i.i.d.) increments take the values $+1, 0, -1$ with respective probabilities $1/4, 1/4, 1/2$ under P . The filtration $\mathbb{F}$ is generated by X and the contingent claim is $H = |X_2|^2$. Any predictable process θ is determined by the value of θ_1 and the three possible values of θ_2 on the sets $\{X_1 = +1\}$, $\{X_1 = 0\}$, $\{X_1 = -1\}$ generating $\mathcal{F}_1$, and we denote the latter by $\theta_2(+1)$, $\theta_2(0)$, $\theta_2(-1)$ respectively. If there is a risk-minimizing strategy ϕ^* with $V_T(\phi^*) = H$, then θ^* must be given by (3.1) and an explicit calculation yields the values $\theta_1^* = -1/11$, $\theta_2^*(+1) = 21/11$, $\theta_2^*(0) = -1/11$, $\theta_2^*(-1) = -23/11$ which lead to an initial risk of

$$R_0(\phi^*) = \frac{24}{66}.$$

But for any mean-self-financing strategy ϕ with $V_T(\phi) = H$, the initial risk $R_0(\phi)$ can also be viewed as a function of the four variables $\theta_1, \theta_2(+1), \theta_2(0), \theta_2(-1)$. The minimum of this function is found to be attained at $\bar{\theta}_1 = -1/11$, $\bar{\theta}_2(+1) =$

$59/33, \bar{\theta}_2(0) = 5/33, \bar{\theta}_2(-1) = -71/33$ and calculated as

$$R_0(\bar{\phi}) = \frac{23}{66} < R_0(\phi^*).$$

This shows that the unique candidate ϕ^* given by (3.1) is not risk-minimizing and hence there cannot exist any risk-minimizing strategy ending at H . This completes the proof. $\square$

Remark Intuitively, the reason for the failure of the risk-minimization approach in the non-martingale case is a compatibility problem. At any time t , we minimize $R_t(\phi)$ over all admissible continuations from t on and obtain a continuation which is optimal when viewed in t only. But for $s < t$, the s -optimal continuation from s on tells us what to do on the entire interval $(s, T] \supset (t, T]$ and this may be different from what the t -optimal continuation from t on prescribes. The above counterexample shows that this indeed creates a problem in general, and the remarkable result in Theorem 2.4 is that the martingale property of X guarantees the required compatibility.

Before we turn to the somewhat technical concept of local risk-minimization in continuous time, it may be useful to explain the basic ideas and results in a *discrete-time framework*; an elementary introduction can also be found in Föllmer and Schweizer (1989). We consider for this a situation where trading is only done at dates $k = 0, 1, \dots, T \in \mathbb{N}$. At time k , we choose the numbers θ_{k+1} of shares to be held over the time period $(k, k+1]$ and the number η_k of units of asset 0 to be held over $[k, k+1)$. Note that predictability of θ forces us to determine the date $k+1$ holdings θ_{k+1} already at date k . The actual time k portfolio is $\phi_k = (\theta_k, \eta_k)$ and its value is $V_k(\phi) = \theta_k^{\text{tr}} X_k + \eta_k$. Since we want to minimize risk locally, we now consider the incremental cost incurred by adjusting the portfolio from ϕ_k to ϕ_{k+1} . Because θ_{k+1} is already chosen at time k with prices given by X_k , this cost increment is

$$\begin{aligned} C_{k+1}(\phi) - C_k(\phi) &= (\theta_{k+1} - \theta_k)^{\text{tr}} X_k + \eta_{k+1} - \eta_k \\ &= V_{k+1}(\phi) - V_k(\phi) - \theta_{k+1}^{\text{tr}} (X_{k+1} - X_k) \\ &= \Delta V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \end{aligned}$$

with the difference operator $\Delta U_{k+1} := U_{k+1} - U_k$ for any discrete-time stochastic process U .

For local risk-minimization, our goal is to minimize $E[(C_{k+1}(\phi) - C_k(\phi))^2 \mid \mathcal{F}_k]$ with respect to the time k control variables θ_{k+1} and η_k . To be accurate, this requires integrability conditions on θ and η , but we leave these aside for the moment. By using the expression for $\Delta C_{k+1}(\phi)$ and the fact that the $\mathcal{F}_k$ -measurable term

$V_k(\phi)$ does not influence the conditional variance given $\mathcal{F}_k$, we can write

$$\begin{aligned} E \left[(\Delta C_{k+1}(\phi))^2 \middle| \mathcal{F}_k \right] &= \text{Var} \left[V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \middle| \mathcal{F}_k \right] \\ &\quad + \left(E \left[V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \middle| \mathcal{F}_k \right] - V_k(\phi) \right)^2. \end{aligned}$$

Because the first term on the right-hand side does not depend on η_k , it is clearly optimal to choose η_k in such a way that

$$V_k(\phi) = E \left[V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \middle| \mathcal{F}_k \right]. \quad (3.2)$$

This is equivalent to

$$0 = E \left[\Delta V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \middle| \mathcal{F}_k \right] = E[\Delta C_{k+1}(\phi) | \mathcal{F}_k]$$

so that an optimal strategy should again be mean-self-financing. Because $V_T(\phi) = H$ is fixed, (3.2) implies by a backward induction argument that for the purposes of minimizing $E[(\Delta C_{k+1}(\phi))^2 | \mathcal{F}_k]$ at time k , the value $V_{k+1}(\phi)$ may be considered as given. Thus it only remains to minimize $\text{Var} \left[V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1} \middle| \mathcal{F}_k \right]$ with respect to the $\mathcal{F}_k$ -measurable quantity θ_{k+1} , and this will be achieved if and only if

$$\text{Cov} \left(V_{k+1}(\phi) - \theta_{k+1}^{\text{tr}} \Delta X_{k+1}, \Delta X_{k+1} \middle| \mathcal{F}_k \right) = 0. \quad (3.3)$$

To simplify this, we use the Doob decomposition of X into a martingale $\bar{M}$ and a predictable process $\bar{A}$ given by $\bar{M}_0 := 0 =: \bar{A}_0$, $\Delta \bar{A}_{k+1} := E[\Delta X_{k+1} | \mathcal{F}_k]$ and $\Delta \bar{M}_{k+1} := \Delta X_{k+1} - \Delta \bar{A}_{k+1}$. Then (3.3) can be rewritten as

$$0 = \text{Cov} \left(\Delta C_{k+1}(\phi), \Delta \bar{M}_{k+1} \middle| \mathcal{F}_k \right) = E \left[\Delta C_{k+1}(\phi) \Delta \bar{M}_{k+1} \middle| \mathcal{F}_k \right],$$

which says that the product of the two martingales $C(\phi)$ and $\bar{M}$ must be a martingale or (equivalently) that $C(\phi)$ and $\bar{M}$ must be strongly orthogonal under P . Thus in discrete time

$$\begin{aligned} &\text{a suitably integrable strategy } \phi \text{ is locally risk-minimizing if and only} \\ &\text{if its cost process } C(\phi) \text{ is a martingale and strongly orthogonal to the} \\ &\text{martingale part (here } \bar{M} \text{) of } X. \end{aligned} \quad (3.4)$$

Before passing to the continuous-time case, let us point out another useful property which will have an analogue later on. Suppose for simplicity that $d = 1$. Because θ_{k+1} is $\mathcal{F}_k$ -measurable, we can solve (3.3) for θ_{k+1} to obtain

$$\theta_{k+1} = \frac{\text{Cov}(V_{k+1}(\phi), \Delta X_{k+1} | \mathcal{F}_k)}{\text{Var}[\Delta X_{k+1} | \mathcal{F}_k]} = \frac{E \left[V_{k+1}(\phi) \Delta \bar{M}_{k+1} \middle| \mathcal{F}_k \right]}{E \left[(\Delta \bar{M}_{k+1})^2 \middle| \mathcal{F}_k \right]}.$$

Using $E[\theta_{k+1} \Delta X_{k+1} | \mathcal{F}_k] = \theta_{k+1} \Delta \bar{A}_{k+1}$ and plugging into (3.2) yields

$$\begin{aligned} V_k(\phi) &= E \left[V_{k+1}(\phi) - \theta_{k+1} \Delta \bar{A}_{k+1} \middle| \mathcal{F}_k \right] \\ &= E \left[V_{k+1}(\phi) \left(1 - \frac{\Delta \bar{A}_{k+1}}{E[(\Delta \bar{M}_{k+1})^2 | \mathcal{F}_k]} \Delta \bar{M}_{k+1} \right) \middle| \mathcal{F}_k \right] \\ &= E \left[V_{k+1}(\phi) \frac{\bar{Z}_{k+1}}{\bar{Z}_k} \middle| \mathcal{F}_k \right] \end{aligned}$$

so that

$$\text{for a locally risk-minimizing strategy } \phi, \text{ the product } \bar{Z}V(\phi) \text{ is a } P\text{-martingale} \quad (3.5)$$

if the process $\bar{Z}$ is defined by the difference equation

$$\bar{Z}_{k+1} - \bar{Z}_k = \bar{Z}_k \left(\frac{\bar{Z}_{k+1}}{\bar{Z}_k} - 1 \right) = -\bar{Z}_k \bar{\lambda}_{k+1} \Delta \bar{M}_{k+1}, \quad \bar{Z}_0 = 1 \quad (3.6)$$

with the predictable process

$$\bar{\lambda}_{k+1} := \frac{\Delta \bar{A}_{k+1}}{E[(\Delta \bar{M}_{k+1})^2 | \mathcal{F}_k]} = \frac{E[\Delta X_{k+1} | \mathcal{F}_k]}{\text{Var}[\Delta X_{k+1} | \mathcal{F}_k]}, \quad k = 0, 1, \dots, T-1.$$

This property will come up again later in a continuous-time version.

Remark The above definition of local risk-minimization in discrete time is different from the original one. The idea there is to consider at time k instead of $E[(C_{k+1}(\phi) - C_k(\phi))^2 | \mathcal{F}_k]$ the risk $R_k(\phi) = E[(C_T(\phi) - C_k(\phi))^2 | \mathcal{F}_k]$. But just as before and in contrast to risk-minimization, this is viewed as a function of the time k control variables η_k and θ_{k+1} only and minimized only locally, i.e., with respect to these local variables. A more formal definition can be found in Schweizer (1988) or Lamberton, Pham and Schweizer (1998) who also prove the equivalence between the two definitions; see the remark on p. 25 of Schweizer (1988) or Proposition 2 of Lamberton, Pham and Schweizer (1998). The reason for using $R_k(\phi)$ is that this formulation can be generalized to continuous time.

Let us now turn to the case of continuous time. Because we want to work again with local variances, we require more specific assumptions on the price process X and we start by making these precise. Since $\mathbb{P} \neq \emptyset$, we know already that X is a semimartingale under P . We now assume that X is in $\mathcal{S}_{\text{loc}}^2(P)$ so that it can be decomposed as $X = X_0 + M + A$ where $M \in \mathcal{M}_{0,\text{loc}}^2(P)$ is an $\mathbb{R}^d$ -valued locally square-integrable local P -martingale null at 0 and A is an $\mathbb{R}^d$ -valued predictable process of finite variation also null at 0. We denote by $\langle M \rangle = (\langle M \rangle^{ij})_{i,j=1,\dots,d} =$

$(\langle M^i, M^j \rangle)_{i,j=1,\dots,d}$ the matrix-valued predictable covariance process of M and we suppose that A is absolutely continuous with respect to $\langle M \rangle$ in the sense that

$$A_t^i = \left(\int_0^t d\langle M \rangle_s \hat{\lambda}_s \right)^i := \sum_{j=1}^d \int_0^t \hat{\lambda}_s^j d\langle M^i, M^j \rangle_s, \quad 0 \leq t \leq T, i = 1, \dots, d$$

for some $\mathbb{R}^d$ -valued predictable process $\hat{\lambda}$ such that the *mean-variance tradeoff process*

$$\hat{K}_t := \int_0^t \hat{\lambda}_s^{\text{tr}} d\langle M \rangle_s \hat{\lambda}_s = \sum_{i,j=1}^d \int_0^t \hat{\lambda}_s^i \hat{\lambda}_s^j d\langle M^i, M^j \rangle_s$$

is finite P -a.s. for each $t \in [0, T]$. This complex of conditions on X is sometimes called the *structure condition (SC)*. Since $\mathbb{P} \neq \emptyset$, it is for instance automatically satisfied if X is continuous; see Theorem 1 of Schweizer (1995a) for this and Choulli and Stricker (1996) for more general results in this direction. Additional results on the relation between (SC) and properties of absence of arbitrage for the process X can be found in Delbaen and Schachermayer (1995). Note that the stochastic integral $\int \hat{\lambda} dM$ is well-defined under (SC) and that its variance process is $\langle \int \hat{\lambda} dM \rangle = \hat{K}$; this will be used later on.

Definition Θ_S denotes the space of all processes $\theta \in L(X)$ for which the stochastic integral $\int \theta dX$ is in the space $\mathcal{S}^2(P)$ of semimartingales. Equivalently, θ must be predictable with

$$E \left[\int_0^T \theta_s^{\text{tr}} d[M]_s \theta_s + \left(\int_0^T |\theta_s^{\text{tr}} dA_s| \right)^2 \right] < \infty.$$

(This equivalence does not use (SC); it only requires X to be a special semimartingale.)

Definition An L^2 -strategy is a pair $\phi = (\theta, \eta)$ where $\theta \in \Theta_S$ and $\eta = (\eta_t)_{0 \leq t \leq T}$ is a real-valued adapted process such that the *value process* $V(\phi) := \theta^{\text{tr}} X + \eta$ is right-continuous and square-integrable (i.e., $V_t(\phi) \in L^2(P)$ for each $t \in [0, T]$). The cost process $C(\phi)$, the risk process $R(\phi)$ and the concept of mean-self-financing are defined as in section 2. Note that in the martingale case $A \equiv 0$, we have $\Theta_S = L^2(X)$, so that the notions of RM-strategy and L^2 -strategy then coincide.

For a formal description of local risk-minimization in continuous time, we now restrict our attention to the case $d = 1$. One can proceed in a similar way and obtain analogous results for $d > 1$; the details for this have been worked out and will be presented elsewhere. The only reason for choosing $d = 1$ here is that this permits references to already published work. Let us first fix some terminology. A *partition*

of $[0, T]$ is a finite set $\tau = \{t_0, t_1, \dots, t_k\}$ of times with $0 = t_0 < t_1 < \dots < t_k = T$ and the mesh size of τ is $|\tau| := \max_{t_i, t_{i+1} \in \tau} (t_{i+1} - t_i)$. The number k of times is not fixed, but can depend on τ . A sequence $(\tau_n)_{n \in \mathbb{N}}$ of partitions is called *increasing* if $\tau_n \subseteq \tau_{n+1}$ for all n ; it *tends to the identity* if $\lim_{n \rightarrow \infty} |\tau_n| = 0$.

The next definition translates the idea that changing an optimal strategy over a small time interval should lead to an increase of risk, at least asymptotically. The form of the denominator indicates that the appropriate time scale for these asymptotics is determined by the fluctuations of X as measured by its predictable quadratic variation.

Definition A *small perturbation* is an L^2 -strategy $\Delta = (\delta, \epsilon)$ such that δ is bounded, the variation of $\int \delta dA$ is bounded (uniformly in t and ω) and $\delta_T = \epsilon_T = 0$. For any subinterval $(s, t]$ of $[0, T]$, we then define the small perturbation

$$\Delta|_{(s,t]} := (\delta I_{(s,t]}, \epsilon I_{[s,t)}).$$

The asymmetry between δ and ϵ reflects the fact that δ is predictable and ϵ merely adapted.

Definition For an L^2 -strategy ϕ , a small perturbation Δ and a partition τ of $[0, T]$, we set

$$r^\tau(\phi, \Delta) := \sum_{t_i, t_{i+1} \in \tau} \frac{R_{t_i}(\phi + \Delta|_{(t_i, t_{i+1}]}) - R_{t_i}(\phi)}{E[\langle M \rangle_{t_{i+1}} - \langle M \rangle_{t_i} | \mathcal{F}_{t_i}]} I_{(t_i, t_{i+1}]}$$

ϕ is called *locally risk-minimizing* if

$$\liminf_{n \rightarrow \infty} r^{\tau_n}(\phi, \Delta) \geq 0 \quad (P \otimes \langle M \rangle)\text{-a.e. on } \Omega \times [0, T]$$

for every small perturbation Δ and every increasing sequence $(\tau_n)_{n \in \mathbb{N}}$ of partitions tending to the identity.

Lemma 3.2 *Let $d = 1$ and suppose that $\langle M \rangle$ is P -a.s. strictly increasing. If an L^2 -strategy is locally risk-minimizing, it is also mean-self-financing.*

Proof This is Lemma 2.1 of Schweizer (1991); note that its assumption (X1) of square-integrability for M is not required in the proof. $\square$

Thanks to Lemma 3.2, we can in searching for locally risk-minimizing strategies restrict ourselves to the class of mean-self-financing strategies. Together with the terminal condition $V_T(\phi) = H$, this class can be parametrized by processes $\theta \in \Theta_S$ so that we effectively have to deal with one dimension fewer than before. To proceed, we then split $r^\tau(\phi, \Delta)$ into a term depending only on θ and δ and a

second term involving η and ϵ as well. The subsequent assumptions ensure that the second term vanishes asymptotically, and the first one is dealt with by means of differentiation results for semimartingales presented in Schweizer (1990). In the end, we then obtain the following result; note that it exactly parallels (3.4).

Theorem 3.3 *Suppose that X satisfies (SC), $d = 1$, M is in $\mathcal{M}_0^2(P)$, $\langle M \rangle$ is P -a.s. strictly increasing, A is P -a.s. continuous and $E[\widehat{K}_T] < \infty$. Let $H \in L^2(\mathcal{F}_T, P)$ be a contingent claim and ϕ an L^2 -strategy with $V_T(\phi) = H$ P -a.s. Then ϕ is locally risk-minimizing if and only if ϕ is mean-self-financing and the martingale $C(\phi)$ is strongly orthogonal to M .*

Proof This follows immediately from Proposition 2.3 of Schweizer (1991) once we note that

$$E[\widehat{K}_T] = E\left[\int_0^T |\widehat{\lambda}_u|^2 d\langle M \rangle_u\right] < \infty$$

implies that $\widehat{\lambda} \in L^2(P \otimes \langle M \rangle)$ so that $|\widehat{\lambda}| \log^+ |\widehat{\lambda}|$ is $(P \otimes \langle M \rangle)$ -integrable. Assumption (X5) of Schweizer (1991) (X continuous at T P -a.s.) is not used in the proof. $\square$

Now we return to the general case $d \geq 1$. The preceding result motivates the following:

Definition Let $H \in L^2(\mathcal{F}_T, P)$ be a contingent claim. An L^2 -strategy ϕ with $V_T(\phi) = H$ P -a.s. is called *pseudo-locally risk-minimizing* or *pseudo-optimal* for H if ϕ is mean-self-financing and the martingale $C(\phi)$ is strongly orthogonal to M .

For $d = 1$ and X sufficiently well-behaved, we have just seen that pseudo-optimal and locally risk-minimizing strategies are the same. But, in general, pseudo-optimal strategies are both easier to find and to characterize. This is shown in the next result which is due to Föllmer and Schweizer (1991).

Proposition 3.4 *A contingent claim $H \in L^2(\mathcal{F}_T, P)$ admits a pseudo-optimal L^2 -strategy ϕ with $V_T(\phi) = H$ P -a.s. if and only if H can be written as*

$$H = H_0 + \int_0^T \xi_u^H dX_u + L_T^H \quad P\text{-a.s.} \quad (3.7)$$

with $H_0 \in L^2(\mathcal{F}_0, P)$, $\xi^H \in \Theta_S$ and $L^H \in \mathcal{M}_0^2(P)$ strongly P -orthogonal to M . The strategy ϕ is then given by

$$\theta_t = \xi_t^H, \quad 0 \leq t \leq T$$

and

$$C_t(\phi) = H_0 + L_t^H, \quad 0 \leq t \leq T;$$

its value process is

$$V_t(\phi) = C_t(\phi) + \int_0^t \theta_u dX_u = H_0 + \int_0^t \xi_u^H dX_u + L_t^H, \quad 0 \leq t \leq T \quad (3.8)$$

so that η is also determined by the above description.

Proof This is Proposition (2.24) of Föllmer and Schweizer (1991), but for completeness we repeat here the simple proof. Write

$$H = V_T(\phi) = C_T(\phi) + \int_0^T \theta_u dX_u = C_0(\phi) + \int_0^T \theta_u dX_u + C_T(\phi) - C_0(\phi)$$

and use the definition of pseudo-optimality. $\square$

Quite apart from the connection to local risk-minimization, the decomposition (3.7) is in itself interesting. In the martingale case where $A \equiv 0$ and $M = X - X_0$, it is the well-known Galtchouk–Kunita–Watanabe decomposition (2.1). In the general case, it has been called in the literature the *Föllmer–Schweizer decomposition* of H and has been studied by several authors. Sufficient conditions for its existence have, for instance, been given by Buckdahn (1993), Schweizer (1994a), Monat and Stricker (1995), Schweizer (1995a), Delbaen, Monat, Schachermayer, Schweizer and Stricker (1997) or Pham, Rheinländer and Schweizer (1998). The simplest sufficient condition is that the mean-variance tradeoff process $\widehat{K}$ should be bounded uniformly in t and ω ; see Theorem 3.4 of Monat and Stricker (1995). A survey of some results on the Föllmer–Schweizer decomposition has been given by Stricker (1996).

In view of Theorem 3.3 and Proposition 3.4, finding the Föllmer–Schweizer decomposition of a given contingent claim H is important because it allows one to obtain a locally risk-minimizing strategy under some additional assumptions. In Buckdahn (1993) and Schweizer (1994a), the existence of this decomposition is proved by means of backward stochastic differential equations, whereas Monat and Stricker (1995) and Pham, Rheinländer and Schweizer (1998) use a fixed point argument. But all these results do not provide a constructive way of finding ξ^H and L^H more explicitly. Following Föllmer and Schweizer (1991) and Schweizer (1995a), we therefore explain how one can often obtain (3.7) by switching to a suitably chosen martingale measure for X ; this notably works in the case where X is continuous and has a bounded mean-variance tradeoff. Moreover, this approach is in perfect analogy to the situation in discrete time.

Inspired by the difference equation (3.6), we consider the stochastic differential equation

$$d\widehat{Z}_t = -\widehat{Z}_t \widehat{\lambda}_t dM_t, \quad \widehat{Z}_0 = 1.$$

Its unique strong solution is the stochastic exponential $\widehat{Z} = \mathcal{E}(-\int \widehat{\lambda} dM)$; if X (hence also M) is continuous, this is explicitly given by

$$\begin{aligned} \widehat{Z}_t &= \exp\left(-\int_0^t \widehat{\lambda}_u dM_u - \frac{1}{2}\left\langle \int \widehat{\lambda} dM \right\rangle_t\right) = \exp\left(-\int_0^t \widehat{\lambda}_u dM_u - \frac{1}{2}\widehat{K}_t\right), \\ 0 &\leq t \leq T. \end{aligned}$$

It is well known and easily checked that $\widehat{Z}$ is in general a locally square-integrable local P -martingale such that

$$\begin{aligned} \widehat{Z}X &\text{ is a local } P\text{-martingale,} \\ \widehat{Z} \int \theta dX &\text{ is a local } P\text{-martingale for every } \theta \in \Theta_S \end{aligned} \quad (3.9)$$

and

$$\widehat{Z}L \text{ is a local } P\text{-martingale for every } L \in \mathcal{M}_{0,\text{loc}}^2(P) \text{ strongly } P\text{-orthogonal to } M; \quad (3.10)$$

see for instance Theorem (3.5) of Föllmer and Schweizer (1991) or Schweizer (1995a). By (3.8), this implies the analogue of (3.5), that

$$\text{for a pseudo-optimal } L^2\text{-strategy } \phi \text{ for } H, \text{ the product } \widehat{Z}V(\phi) \text{ is a local } P\text{-martingale.} \quad (3.11)$$

In the situation of (3.11), $C(\phi)$ is a martingale and $\sup_{0 \leq t \leq T} |V_t(\phi)| \in L^2(P)$; hence $\widehat{Z}V(\phi)$ is then a true martingale if $\widehat{Z}$ itself is a square-integrable martingale.

So suppose now that $\widehat{Z} \in \mathcal{M}^2(P)$. A restrictive sufficient condition for this is by Theorem II.2 of Lepingle and Mémmin (1978), uniform boundedness of $\widehat{K}$ in t and ω . In concrete applications, one can also try to check square-integrability directly. If $\widehat{Z}$ is also strictly positive on $[0, T]$ (which will certainly hold if M , hence $\widehat{Z}$, is continuous), then

$$\frac{d\widehat{P}}{dP} := \widehat{Z}_T = \mathcal{E}\left(-\int \widehat{\lambda} dM\right)_T \in L^2(P) \quad (3.12)$$

defines a probability measure $\widehat{P} \approx P$ which is in $\mathbb{P}$ according to (3.9). For reasons explained below, this measure $\widehat{P}$ is called the *minimal equivalent local martingale measure* for X . Since the martingale form of (3.11) says that $V(\phi)$ is a $\widehat{P}$ -martingale for a pseudo-optimal L^2 -strategy ϕ for H , we get

$$V_t(\phi) = \widehat{E}[H|\mathcal{F}_t] =: V_t^{H,\widehat{P}}, \quad 0 \leq t \leq T \quad (3.13)$$

for such a strategy. Hence we are led to study the $\widehat{P}$ -martingale $V^{H,\widehat{P}}$ and its relation to the local $\widehat{P}$ -martingale X . Note that $H \in L^1(\widehat{P})$ because H and $\widehat{Z}_T$ are both in $L^2(P)$; hence $V^{H,\widehat{P}}$ is indeed well-defined.

In addition to the previous assumptions, suppose now also that X is *continuous*. By (3.9), X is a local $\widehat{P}$ -martingale and so $V^{H,\widehat{P}}$ admits a Galtchouk–Kunita–Watanabe decomposition under $\widehat{P}$ with respect to X as

$$V_t^{H,\widehat{P}} = V_0^{H,\widehat{P}} + \int_0^t \xi_u^{H,\widehat{P}} dX_u + L_t^{H,\widehat{P}}, \quad 0 \leq t \leq T \quad (3.14)$$

where $\xi^{H,\widehat{P}} \in L(X)$ and $L^{H,\widehat{P}}$ is a local $\widehat{P}$ -martingale null at 0 and strongly $\widehat{P}$ -orthogonal to X ; see Ansel and Stricker (1993). For $t = T$, this gives in particular a decomposition of the random variable H . Thanks to the continuity of X , $L^{H,\widehat{P}}$ is also a local P -martingale strongly P -orthogonal to X ; see Ansel and Stricker (1992) or Schweizer (1995a). In many cases, this decomposition gives us what we need; this was already observed in Theorem (3.14) of Föllmer and Schweizer (1991).

Theorem 3.5 *Suppose that X is continuous and hence satisfies (SC) (because $\mathbb{P} \neq \emptyset$). Define the strictly positive local P -martingale $\widehat{Z} := \mathcal{E}(-\int \widehat{\lambda} dM)$ and suppose that*

$$\widehat{Z} \in \mathcal{M}^2(P). \quad (3.15)$$

Define $\widehat{P}$ and $V^{H,\widehat{P}}$ as above by (3.12) and (3.13), respectively. If either

$$H \text{ admits a Föllmer–Schweizer decomposition} \quad (3.16)$$

or

$$V_0^{H,\widehat{P}} \in L^2(P), \xi^{H,\widehat{P}} \in \Theta_S \text{ and } L^{H,\widehat{P}} \in \mathcal{M}^2(P), \quad (3.17)$$

then (3.14) for $t = T$ gives the Föllmer–Schweizer decomposition of H and $\xi^{H,\widehat{P}}$ determines a pseudo-optimal L^2 -strategy for H . A sufficient condition for (3.15), (3.16) and (3.17) is that $\widehat{K}$ is uniformly bounded.

Proof This is almost a summary of the preceding arguments. If we have (3.16), then (3.10) implies that L^H is a local $\widehat{P}$ -martingale and strongly $\widehat{P}$ -orthogonal to X , since $\langle L^H, X \rangle = \langle L^H, M \rangle = 0$ by the continuity of X . By the uniqueness of the Galtchouk–Kunita–Watanabe decomposition, (3.7) and (3.14) for $t = T$ must therefore coincide. If we have (3.17), the argument just before Theorem 3.5 shows that (3.14) for $t = T$ gives a Föllmer–Schweizer decomposition for H which by uniqueness must again coincide with (3.7). The assertion about $\xi^{H,\widehat{P}}$ is then immediate from Proposition 3.4, and that boundedness of $\widehat{K}$ is sufficient

follows from Theorem II.2 of Lepingle and Mémin (1978), Theorem 3.4 of Monat and Stricker (1995) and Lemma 6 of Pham, Rheinländer and Schweizer (1998) respectively. $\square$

The basic message of Theorem 3.5 is that for X continuous, finding a locally risk-minimizing strategy essentially boils down to finding the Galtchouk–Kunita–Watanabe decomposition of H under the minimal ELMM $\hat{P}$. This is very useful because the density process $\hat{Z}$ of $\hat{P}$ with respect to P can immediately be written down explicitly and we can directly see the dynamics of X under $\hat{P}$. In particular, finding (3.14) can often be reduced to solving a partial differential equation if H can be written as a function of the final value of some (possibly multidimensional) process which has a Markovian structure under $\hat{P}$. This is explained in Pham, Rheinländer and Schweizer (1998) and for the case of a stochastic volatility model in more detail also in Heath, Platen and Schweizer (2000).

Remark We emphasize that by its very nature, local risk-minimization is a *hedging approach* designed to control the riskiness of a strategy as measured by its local cost fluctuations. If there is an optimal strategy ϕ , we can use $V_t(\phi)$ as a *value* or price of H at time t , but two things about this should be kept in mind: such a valuation is a by-product of the method, not its primary objective, and it is only a valuation with respect to the (subjective) criterion of local risk-minimization.

If we can obtain the Föllmer–Schweizer decomposition of H via the Galtchouk–Kunita–Watanabe decomposition of H under $\hat{P}$, we know from (3.13) that the value process of the corresponding pseudo-optimal strategy ϕ is given by the conditional expectations of H under $\hat{P}$. Together with the preceding remark, this shows that $V^{H, \hat{P}}$ can be interpreted as an *intrinsic valuation process* for H and identifies $\hat{P}$ as the valuation operator naturally associated with the criterion of local risk-minimization. It seems therefore appropriate to comment briefly on the origins and properties of $\hat{P}$ and in particular on the terminology “minimal ELMM”.

The first formal definition of a minimal martingale measure appears in Föllmer and Schweizer (1991). They consider a continuous square-integrable real-valued process X and focus on equivalent martingale measures Q for X that satisfy $\frac{dQ}{dP} \in L^2(P)$. A martingale measure Q from this class is called *minimal* if $Q = P$ on $\mathcal{F}_0$ and if any $L \in \mathcal{M}_0^2(P)$ strongly P -orthogonal to M is still a martingale under Q . Theorem (3.5) of Föllmer and Schweizer (1991) then proves that such a measure is unique and must coincide with $\hat{P}$ defined above; existence is therefore equivalent to $\hat{Z}$ being in $\mathcal{M}^2(P)$. These results have precursors in Schweizer (1988, 1991) for the special case where $\mathcal{M}^2(P)$ is generated by M and a second orthogonal P -martingale N . In that context, the “minimal” martingale measure is introduced as an equivalent probability that turns X into a martingale and preserves

the martingale property of N . The terminology “minimal” is there motivated by the fact that apart from turning X into a martingale, this measure disturbs the overall martingale and orthogonality structures as little as possible.

The original motivation in Schweizer (1988) for introducing a minimal martingale measure $\hat{P}$ was its use in finding locally risk-minimizing strategies via a variant of Theorem 3.5. It has subsequently turned out that $\hat{P}$ appears quite naturally in a number of other situations as well. Apart from local risk-minimization as discussed above, one can mention here logarithmic utility maximization problems (see Cvitanić and Karatzas (1992), Karatzas (1997), Amendinger, Imkeller and Schweizer (1998)), pricing under local utility indifference (see Davis (1994, 1997), Karatzas and Kou (1996)), equilibrium prices for assets (see Pham and Touzi (1996) or Jouini and Napp (1998)) and value preservation (see Korn (1997, 1998)). In view of this apparent ubiquity of $\hat{P}$, it is natural to ask for a more concise and transparent description of $\hat{P}$, preferably as the solution of a suitable optimization problem. This would give a more precise meaning to the sense in which $\hat{P}$ is optimal.

Proposition 3.6 *Let X be a continuous adapted process admitting at least one equivalent local martingale measure Q . If $\hat{P}$ defined by (3.12) is a probability measure equivalent to P , then $\hat{P}$ minimizes the reverse relative entropy $H(P|Q)$ over all ELMMs Q for X .*

Proof See Theorem 1 of Schweizer (1999a). □

At present, this seems to be the most general known characterization of $\hat{P}$. For the case of a multidimensional diffusion model for X , this can also be found in Section 5.6 of Karatzas (1997), and Schweizer (1999a) contains a discussion of other less general results. A counterexample in Schweizer (1999a) shows that Proposition 3.6 does not carry over to the case where X is discontinuous. Finding an analogous description of $\hat{P}$ in general seems to be an open problem.

4 Mean-variance hedging

Let us now return to the general situation where X is a semimartingale under P and H is a given contingent claim. The key difference between (local) risk-minimization and mean-variance hedging is that we no longer impose on our trading strategies the replication requirement $V_T = H$ P -a.s., but insist instead on the self-financing constraint (1.1). For a self-financing pre-strategy (V_0, θ) , the

shortfall or loss from hedging H by (V_0, θ) is then

$$H - V_T(V_0, \theta) = H - V_0 - \int_0^T \theta_u dX_u,$$

and we want to minimize the $L^2(P)$ -norm of this quantity by choosing (V_0, θ) . Note that a symmetric criterion is quite natural in the present context of hedging and pricing options because one does not know at the start whether one is dealing with a buyer or a seller; see Bertsimas, Kogan and Lo (1999) for an amplification of this point. Choosing the L^2 -norm is mainly for convenience because it allows fairly explicit results while at the same time leading to interesting mathematical questions. For brevity, we write L^2 for $L^2(P)$ if there is no risk of confusion.

We first have to be more specific about our strategies. We do not assume that $\mathcal{F}_0$ is trivial but we insist on a non-random initial capital V_0 .

Definition We denote by Θ_2 the set of all $\theta \in L(X)$ such that the stochastic integral process $G(\theta) := \int \theta dX$ satisfies $G_T(\theta) \in L^2(P)$. For a fixed linear subspace Θ of Θ_2 , a Θ -strategy is a pair $(V_0, \theta) \in \mathbb{R} \times \Theta$ and its value process is $V_0 + G(\theta)$. A Θ -strategy $(\tilde{V}_0, \tilde{\theta})$ is called Θ -mean-variance optimal for a given contingent claim $H \in L^2$ if it minimizes $\|H - V_0 - G_T(\theta)\|_{L^2}$ over all Θ -strategies (V_0, θ) , and $\tilde{V}_0$ is then called the Θ -approximation price for H .

The preceding definition depends on the choice of the space Θ of strategies allowed for trading and we shall be more specific about this later on. For the moment, however, we go in the other direction and consider an even more general framework. Suppose we have chosen a linear subspace Θ of Θ_2 . Then the linear subspace

$$\mathcal{G} := G_T(\Theta) = \left\{ \int_0^T \theta_u dX_u \mid \theta \in \Theta \right\}$$

of L^2 describes all outcomes of self-financing Θ -strategies with initial wealth $V_0 = 0$ and

$$\mathcal{A} := \mathbb{R} + \mathcal{G} = \left\{ V_0 + \int_0^T \theta_u dX_u \mid (V_0, \theta) \in \mathbb{R} \times \Theta \right\}$$

is the space of contingent claims replicable by self-financing Θ -strategies. Our goal in mean-variance hedging is to find the projection in L^2 of H on $\mathcal{A}$ and this can be studied for a general linear subspace $\mathcal{G}$ of L^2 . In analogy to the above definition, we introduce a $\mathcal{G}$ -mean-variance optimal pair $(\tilde{V}_0, \tilde{g}) \in \mathbb{R} \times \mathcal{G}$ for $H \in L^2$ and call $\tilde{V}_0$ the $\mathcal{G}$ -approximation price for H . In particular, we need no explicit model for X or Θ at this stage and either a discrete-time or a continuous-time choice for X fit equally well into this setting. This was first pointed out in Schweizer (2000) and exploited in Schweizer (1999b). Our presentation here follows the latter.

Definition We say that $\mathcal{G}$ admits no approximate profits in L^2 if $\bar{\mathcal{G}}$ does not contain the constant 1; the bar $\bar{\cdot}$ denotes the closure in L^2 .

With our preceding interpretations, this notion is very intuitive. It says that one cannot approximate (in the L^2 -sense) the riskless payoff 1 by a self-financing strategy with initial wealth 0. This is a no-arbitrage condition on the financial market underlying $\mathcal{G}$; see also Stricker (1990).

Definition A signed $\mathcal{G}$ -martingale measure is a signed measure Q on $(\Omega, \mathcal{F})$ with $Q[\Omega] = 1$, $Q \ll P$ with $\frac{dQ}{dP} \in L^2$ and

$$E_Q[g] = E \left[\frac{dQ}{dP} g \right] = 0 \quad \text{for all } g \in \mathcal{G}.$$

$\mathbb{P}_s^2(\mathcal{G})$ denotes the convex set of all signed $\mathcal{G}$ -martingale measures and an element $\tilde{P}_{\mathcal{G}}$ of $\mathbb{P}_s^2(\mathcal{G})$ is called *variance-optimal* if it minimizes $\left\| \frac{dQ}{dP} \right\|_{L^2} = \sqrt{1 + \text{Var} \left[\frac{dQ}{dP} \right]}$ over all $Q \in \mathbb{P}_s^2(\mathcal{G})$.

Lemma 4.1 Let $\mathcal{G}$ be a linear subspace of L^2 . Then:

- (a) $\mathcal{G}$ admits no approximate profits in L^2 if and only if $\mathbb{P}_s^2(\mathcal{G}) \neq \emptyset$.
- (b) If $\mathcal{G}$ admits no approximate profits in L^2 , then $\bar{\mathcal{A}} = \mathbb{R} + \bar{\mathcal{G}}$.
- (c) If $\mathcal{G}$ admits no approximate profits in L^2 , then the variance-optimal signed $\mathcal{G}$ -martingale measure $\tilde{P}_{\mathcal{G}}$ exists, is unique and satisfies

$$\frac{d\tilde{P}_{\mathcal{G}}}{dP} \in \bar{\mathcal{A}}. \quad (4.1)$$

Proof This very simple result goes back to Delbaen and Schachermayer (1996a) and Schweizer (2000); for completeness, we reproduce here the detailed proof of Schweizer (1999b). We use $(\cdot, \cdot)$ for the scalar product in L^2 .

- (1) An element Q of $\mathbb{P}_s^2(\mathcal{G})$ can be identified with a continuous linear functional Ψ on L^2 satisfying $\Psi = 0$ on $\mathcal{G}$ and $\Psi(1) = 1$ by setting $\Psi(U) = E \left[\frac{dQ}{dP} U \right] = \left(\frac{dQ}{dP}, U \right)$. Hence (a) is clear from the Hahn–Banach theorem.
- (2) Any $g \in \bar{\mathcal{G}}$ is the limit in L^2 of a sequence (g_n) in $\mathcal{G}$; hence $c + g_n = a_n$ is a Cauchy sequence in $\mathcal{A}$ and thus converges in L^2 to a limit $a \in \bar{\mathcal{A}}$ so that $c + g = a \in \bar{\mathcal{A}}$. This gives the inclusion “ $\supseteq$ ” in general. For the converse, we use the assumption that $\mathcal{G}$ admits no approximate profits in L^2 to obtain from part (a) a signed $\mathcal{G}$ -martingale measure Q . The random variable $Z := \frac{dQ}{dP}$ is then in $\mathcal{G}^\perp$ and satisfies $(Z, 1) = Q[\Omega] = 1$. For any $a \in \bar{\mathcal{A}}$, there is a sequence $a_n = c_n + g_n$ in $\mathcal{A}$ converging to a in L^2 . Since $c_n + g_n \in \mathbb{R} + \mathcal{G}$ for all n , we conclude that $c_n = (c_n + g_n, Z) = (a_n, Z)$ converges in $\mathbb{R}$ to

$(a, Z) =: c$. Therefore $g_n = a_n - c_n$ converges in L^2 to $g := a - c$ and since this limit is in $\bar{\mathcal{G}}$, we have $a = c + g \in \mathbb{R} + \bar{\mathcal{G}}$ which proves the inclusion “ $\subseteq$ ”.

- (3) Existence and uniqueness of $\tilde{P}_{\mathcal{G}}$ are clear once we observe that we have to minimize $\|Z\|$ over the closed convex set $\mathcal{Z} := \left\{ Z = \frac{dQ}{dP} \mid Q \in \mathbb{P}_s^2(\mathcal{G}) \right\}$ which is non-empty thanks to (a). For any fixed $Z_0 \in \mathcal{Z}$, the projection $\tilde{Z}$ of Z_0 in L^2 on $\bar{\mathcal{A}}$ is again in $\mathcal{Z}$; in fact, one easily verifies that $\tilde{\Psi}(U) := (\tilde{Z}, U)$ is 0 on $\mathcal{G}$ and has $\tilde{\Psi}(1) = 1$. Since part (b) tells us that $\tilde{Z} = \tilde{c} + \tilde{g}$ with $\tilde{g} \in \bar{\mathcal{G}}$, we obtain $(Z, \tilde{Z}) = \tilde{c} = (\tilde{Z}, \tilde{Z})$ for all $Z \in \mathcal{Z}$ and therefore

$$\|Z\|^2 = \|\tilde{Z}\|^2 + \|Z - \tilde{Z}\|^2 \geq \|\tilde{Z}\|^2 \quad \text{for all } Z \in \mathcal{Z}.$$

Hence we conclude that $\frac{d\tilde{P}_{\mathcal{G}}}{dP} = \tilde{Z}$ is in $\bar{\mathcal{A}}$. □

For any $g \in \mathcal{G}$ and any $Q \in \mathbb{P}_s^2(\mathcal{G})$, we have

$$1 = E_Q[1 - g] = E \left[\frac{dQ}{dP}(1 - g) \right] \leq \left\| \frac{dQ}{dP} \right\|_{L^2} \|1 - g\|_{L^2}$$

by the Cauchy–Schwarz inequality and therefore

$$\frac{1}{\inf_{Q \in \mathbb{P}_s^2(\mathcal{G})} \left\| \frac{dQ}{dP} \right\|_{L^2}} = \sup_{Q \in \mathbb{P}_s^2(\mathcal{G})} \frac{1}{\left\| \frac{dQ}{dP} \right\|_{L^2}} \leq \inf_{g \in \mathcal{G}} \|1 - g\|_{L^2}.$$

This indicates that finding the variance-optimal signed $\mathcal{G}$ -martingale measure is the dual problem to approximating in L^2 the constant 1 by elements of $\mathcal{G}$. This duality is reflected in the next result which gives the $\mathcal{G}$ -approximation price as an expectation under $\tilde{P}_{\mathcal{G}}$.

Proposition 4.2 *Suppose that $\mathcal{G}$ is a linear subspace of L^2 which admits no approximate profits in L^2 . If a contingent claim $H \in L^2$ admits a $\mathcal{G}$ -mean-variance optimal pair $(\tilde{V}_0, \tilde{g})$, the $\mathcal{G}$ -approximation price of H is given by*

$$\tilde{V}_0 = \tilde{E}_{\mathcal{G}}[H],$$

where $\tilde{E}_{\mathcal{G}}$ denotes expectation under the variance-optimal signed $\mathcal{G}$ -martingale measure $\tilde{P}_{\mathcal{G}}$.

Proof If H admits a $\mathcal{G}$ -mean-variance optimal pair $(\tilde{V}_0, \tilde{g})$, then $\tilde{V}_0 + \tilde{g}$ is the projection in L^2 of H on $\bar{\mathcal{A}} = \mathbb{R} + \bar{\mathcal{G}}$ by Lemma 4.1. Since $H - \tilde{V}_0 - \tilde{g}$ is then in the orthogonal complement of $\bar{\mathcal{A}}$, (4.1) implies that $E \left[(H - \tilde{V}_0 - \tilde{g}) \frac{d\tilde{P}_{\mathcal{G}}}{dP} \right] = 0$ and so we obtain

$$\tilde{V}_0 = E \left[(H - \tilde{g}) \frac{d\tilde{P}_{\mathcal{G}}}{dP} \right] = \tilde{E}_{\mathcal{G}}[H]$$

because $\tilde{P}_{\mathcal{G}}$ is in $\mathbb{P}_s^2(\mathcal{G})$. □

The assumption in Proposition 4.2 that H admits a $\mathcal{G}$ -mean-variance optimal pair is obviously unpleasant. We can avoid it by either working *a priori* with elements from the *closed* linear subspace $\tilde{\mathcal{A}} = \mathbb{R} + \tilde{\mathcal{G}}$ or by ensuring in some way that $\mathcal{G}$ (hence also $\mathcal{A}$) is already closed in L^2 . The simpler first solution is preferable if we are not directly interested in the structure of the optimal element $\tilde{V}_0 + \tilde{g}$. This is the case in most situations where we only want to *value* contingent claims by using some quadratic criterion; see for instance Mercurio (1996), Aurell and Simdyankin (1998), Schweizer (1999b) or Schweizer (2000). But for *hedging* purposes, we also want to understand $\tilde{g}$ itself and therefore we follow here the second idea and return to the framework with a semimartingale X and a space $\Theta \subseteq \Theta_2$ of integrands to study the closedness of $G_T(\Theta)$ in L^2 .

So let $X = (X_t)_{0 \leq t \leq T}$ be an $\mathbb{R}^d$ -valued semimartingale which is locally in $L^2(P)$ in the sense that the maximal process $X_t^* := \sup_{0 \leq s \leq t} |X_s|$, $0 \leq t \leq T$, is locally P -square-integrable. Let $(\rho_n)_{n \in \mathbb{N}}$ be a corresponding localizing sequence of stopping times. A process of the form $\theta = \xi I_{\llbracket \sigma, \tau \rrbracket}$ with $\sigma \leq \tau$ stopping times with $\tau \leq \rho_n$ for some n and with a bounded $\mathbb{R}^d$ -valued $\mathcal{F}_\sigma$ -measurable random variable ξ is called a *simple integrand*, and we denote by Θ_{simple} the linear space spanned by all simple integrands. It is evident that $\Theta_{\text{simple}} \subseteq \Theta_2$ and easy to verify that Q is an ELMM for X with $\frac{dQ}{dP} \in L^2(P)$ if and only if Q is in $\mathbb{P}_s^2(\Theta_{\text{simple}})$ and $Q \approx P$. We denote the set of all these probability measures Q by $\mathbb{P}_e^2(X)$.

Definition The *variance-optimal signed martingale measure* $\tilde{P}$ for X is defined as the variance-optimal signed $G_T(\Theta_{\text{simple}})$ -martingale measure.

In general, $\tilde{P}$ is unfortunately a signed measure. But for a continuous process X , the situation is better.

Theorem 4.3 If X is a continuous $\mathbb{R}^d$ -valued semimartingale and $\mathbb{P}_e^2(X) \neq \emptyset$, then $\tilde{P}$ is in $\mathbb{P}_e^2(X)$. In other words, the variance-optimal signed martingale measure for X is then automatically equivalent to P and in particular a probability measure.

Proof See Theorem 1.3 of Delbaen and Schachermayer (1996a). □

In order to study the closedness in L^2 of $\mathcal{G} := G_T(\Theta)$ and also to relate $\tilde{P}$ to $\tilde{P}_{\mathcal{G}}$, we now consider two specific choices of Θ .

Definition Θ_{GLP} consists of all $\theta \in L(X)$ such that $G_T(\theta)$ is in $L^2(P)$ and the process $G(\theta) = \int \theta dX$ is a uniformly Q -integrable Q -martingale for every $Q \in \mathbb{P}_e^2(X)$. Θ_S consists (as in Section 3) of all $\theta \in L(X)$ such that $G(\theta)$ is in the space $\mathcal{S}^2(P)$ of semimartingales.

The space Θ_S was introduced by Schweizer (1994a). At first sight, it appears simpler and more natural because it can be defined directly in terms of the original probability measure P . Moreover, it obviously generalizes the space $L^2(X)$ used in Section 2 for the martingale case to the semimartingale framework. The space Θ_{GLP} was first used by Delbaen and Schachermayer (1996b) and introduced to hedging by Gouriéroux, Laurent and Pham (1998). Its main advantage (as illustrated by the next two results) is that it is better adapted to duality formulations and easier to handle for certain theoretical aspects. On the other hand, proving for an explicitly given strategy θ that it is in Θ is usually much simpler for $\Theta = \Theta_S$ than for $\Theta = \Theta_{GLP}$. For additional results on the relation between Θ_S and Θ_{GLP} , see also Rheinländer (1999).

Theorem 4.4 *Let X be an $\mathbb{R}^d$ -valued semimartingale which is locally in $L^2(P)$ and assume that $\mathbb{P}_e^2(X) \neq \emptyset$. Then $G_T(\Theta_{GLP})$ is closed in $L^2(P)$. If X is continuous, we have in addition that $G_T(\Theta_{GLP}) = \overline{G_T(\Theta_{\text{simple}})}$ where the bar $\overline{}$ denotes the closure in $L^2(P)$; this implies in particular that $\tilde{P} = \tilde{P}_{G_T(\Theta_{GLP})}$.*

Proof This is due to Delbaen and Schachermayer (1996b). The first assertion follows from the equivalence of (i) and (ii) in their Theorem 1.2 (note that their D_2 is always closed in L^2) and the second uses in addition their Theorem 2.2. $\square$

For Θ_S instead of Θ_{GLP} , analyzing the closedness question is more delicate.

Definition Let $Z = (Z_t)_{0 \leq t \leq T}$ be a strictly positive P -martingale with $E[Z_0] = 1$. We say that Z satisfies the *reverse Hölder inequality* $R_2(P)$ if there is some constant C such that

$$E[Z_T^2 | \mathcal{F}_t] \leq C Z_t^2 \quad P\text{-a.s.}$$

for each $t \in [0, T]$. A probability measure $Q \approx P$ is said to satisfy $R_2(P)$ if its density process $Z_t^Q := E\left[\frac{dQ}{dP} \middle| \mathcal{F}_t\right]$, $0 \leq t \leq T$, satisfies $R_2(P)$.

Theorem 4.5 *Let X be a continuous $\mathbb{R}^d$ -valued semimartingale. Then the following statements are equivalent:*

- (a) $\mathbb{P}_e^2(X) \neq \emptyset$ and $G_T(\Theta_S)$ is closed in $L^2(P)$.
- (b) There exists some $Q \in \mathbb{P}_e^2(X)$ satisfying $R_2(P)$.
- (c) The variance-optimal martingale measure $\tilde{P}$ is in $\mathbb{P}_e^2(X)$ and satisfies $R_2(P)$.

Proof This is a partial statement of Theorem 4.1 of Delbaen, Monat, Schachermayer, Schweizer and Stricker (1997). $\square$

Once we know that $G_T(\Theta)$ is closed and does not contain 1, we can obtain Θ -mean-variance optimal Θ -strategies $(\tilde{V}_0, \tilde{\theta})$ by projecting the given contingent claim $H \in L^2$ on the space $\mathcal{A}$ of replicable claims and it becomes interesting to study the structure of the optimal integrand $\tilde{\theta}$ in more detail. Before we do this, let us briefly mention some more recent extensions of the preceding results. It is natural to replace the exponent 2 by $p \in (1, \infty)$ in the definition of Θ_S and to ask if $G_T(\Theta_S)$ is then closed in $L^p(P)$. For the case where X is continuous, this has been treated in Grandits and Krawczyk (1998) who generalized Theorem 4.5 to an arbitrary $p \in (1, \infty)$. The next step is then to eliminate the assumption that X is continuous. This has been done in Choulli, Krawczyk and Stricker (1998, 1999) who first extended the Doob, Burkholder–Davis–Gundy and Fefferman inequalities from (local) martingales to a class of semimartingales (called $\mathcal{E}$ -martingales) with a particular structure inspired by the financial background of the problem. They then used this to provide sufficient conditions for the closedness of $G_T(\Theta_S)$ in $L^p(P)$ when X is an $\mathcal{E}$ -martingale. Moreover, they also generalized earlier results by Delbaen, Monat, Schachermayer, Schweizer and Stricker (1997) on the existence and continuity of the Föllmer–Schweizer decomposition. The problem of finding necessary and sufficient conditions for $G_T(\Theta_S)$ to be closed in this general setting seems at present still open.

Let us now turn to the problem of finding the integrand $\tilde{\theta}$ in the projection of a given $H \in L^2$ on the space $\mathcal{A} = \mathbb{R} + G_T(\Theta)$. For the case where $X = (X_k)_{k=0,1,\dots,T}$ is a real-valued square-integrable process in discrete time with a bounded mean-variance tradeoff, explicit recursive formulae for $\tilde{\theta}$ have been given in Schweizer (1995b). These results are for the one-dimensional case $d = 1$; the extension to $d > 1$ has been worked out and will be presented elsewhere. See also Bertsimas, Kogan and Lo (1999) and Černý (1999) for recent results obtained via dynamic programming arguments. If $X = (X_t)_{0 \leq t \leq T}$ is an $\mathbb{R}^d$ -valued semimartingale, the above recursive expressions take under some additional assumptions the form of a backward stochastic differential equation; see Schweizer (1994a, 1996) for more details. Both types of results simplify considerably if $\log X$ is a Lévy process in either discrete or continuous time and H has a particular structure; this has been worked out by Hubalek and Krawczyk (1998). Theoretical and numerical results for mean-variance optimal strategies can be found in Biagini, Guasoni and Pratelli (2000), Guasoni and Biagini (1999) and Heath, Platen and Schweizer (2000) for the case of a stochastic volatility model, and more numerically oriented studies in diffusion or jump-diffusion models have been done by Bertsimas, Kogan and Lo (1999), Grünewald and Trautmann (1997) and Hipp (1996, 1998). Additional references can also be found after the next theorem.

The most general results on $\tilde{\theta}$ have been obtained for the case where X is continuous and $\mathbb{P}_e^2(X) \neq \emptyset$. By Theorem 4.3, the variance-optimal martingale

measure $\tilde{P}$ for X then exists and is equivalent to P . Moreover, the arguments in Delbaen and Schachermayer (1996a) also show that the process

$$\tilde{Z}_t := \tilde{E} \left[\frac{d\tilde{P}}{dP} \middle| \mathcal{F}_t \right], \quad 0 \leq t \leq T$$

can be written as

$$\tilde{Z}_t = \tilde{Z}_0 + \int_0^t \tilde{\zeta}_u dX_u, \quad 0 \leq t \leq T$$

for some $\tilde{\zeta} \in \Theta_{\text{GLP}}$. In particular, $\tilde{Z}$ is continuous. Note also that (4.1) implies that $\tilde{Z}_0$ is a non-random constant. As the next result shows, $\tilde{P}$, $\tilde{Z}$ and $\tilde{\zeta}$ all turn up in the solution of the mean-variance hedging problem.

Theorem 4.6 *Suppose that X is a continuous process such that $\mathbb{P}_e^2(X) \neq \emptyset$. Let $H \in L^2(P)$ be a contingent claim and write the Galtchouk–Kunita–Watanabe decomposition of H under $\tilde{P}$ with respect to X as*

$$H = \tilde{E}[H|\mathcal{F}_0] + \int_0^T \xi_u^{H,\tilde{P}} dX_u + L_T^{H,\tilde{P}} = V_T^{H,\tilde{P}} \quad (4.2)$$

with

$$V_t^{H,\tilde{P}} := \tilde{E}[H|\mathcal{F}_t] = \tilde{E}[H|\mathcal{F}_0] + \int_0^t \xi_u^{H,\tilde{P}} dX_u + L_t^{H,\tilde{P}}, \quad 0 \leq t \leq T.$$

Then the mean-variance optimal Θ_{GLP} -strategy for H is given by

$$\tilde{V}_0 = \tilde{E}[H] \quad (4.3)$$

and

$$\begin{aligned} \tilde{\theta}_t &= \xi_t^{H,\tilde{P}} - \frac{\tilde{\zeta}_t}{\tilde{Z}_t} \left(V_{t-}^{H,\tilde{P}} - \tilde{E}[H] - \int_0^t \tilde{\theta}_u dX_u \right) \\ &= \xi_t^{H,\tilde{P}} - \tilde{\zeta}_t \left(\frac{V_0^{H,\tilde{P}} - \tilde{E}[H]}{\tilde{Z}_0} + \int_0^{t-} \frac{1}{\tilde{Z}_u} dL_u^{H,\tilde{P}} \right), \quad 0 \leq t \leq T. \end{aligned} \quad (4.4)$$

Proof Thanks to Theorem 4.4, (4.3) follows immediately from Proposition 4.2. According to Corollary 16 of Schweizer (1996), $\tilde{\theta}$ is obtained by projecting the random variable $H - \tilde{E}[H]$ on $G_T(\Theta)$ and this is in principle dealt with in Rheinländer and Schweizer (1997). The representation (4.4) is very similar to their Theorem 6, but we cannot directly use their results since they work with Θ_S instead of Θ_{GLP} . Thus we appeal to some results from Gouriéroux, Laurent and Pham (1998) and this involves a second change of measure. Because $\tilde{Z}$ is a strictly

positive P -martingale and $\tilde{Z}_0$ is deterministic, we can define a new probability measure $\tilde{R} \approx \tilde{P} \approx P$ by setting

$$\frac{d\tilde{R}}{d\tilde{P}} := \frac{\tilde{Z}_T}{\tilde{Z}_0}.$$

Clearly, the $\mathbb{R}^{d+1}$ -valued process $Y = \begin{pmatrix} 1/\tilde{Z} \\ X/\tilde{Z} \end{pmatrix}$ is then a continuous local $\tilde{R}$ -martingale since $\tilde{P} \in \mathbb{P}_e^2(X)$. The density of $\tilde{R}$ with respect to P is $\tilde{Z}_T^2/\tilde{Z}_0$ and because $\tilde{Z}_0$ is deterministic, H is in $L^2(P)$ if and only if $H/\tilde{Z}_T$ is in $L^2(\tilde{R})$. The basic idea of Gouriéroux, Laurent and Pham (1998) is now to use $\tilde{Z}/\tilde{Z}_0$ as a new numeraire, rewrite the original problem in terms of the corresponding new quantities and apply the Galtchouk–Kunita–Watanabe decomposition theorem to $H/\tilde{Z}_T$ under $\tilde{R}$ with respect to Y . This yields

$$\frac{H}{\tilde{Z}_T} = E_{\tilde{R}} \left[\frac{H}{\tilde{Z}_T} \middle| \mathcal{F}_0 \right] + \int_0^T \psi_u dY_u + L_T \quad (4.5)$$

for some $\mathbb{R}^{d+1}$ -valued $\psi \in L(Y)$ such that $\int \psi dY \in \mathcal{M}_0^2(\tilde{R})$ and some $L \in \mathcal{M}_0^2(\tilde{R})$ strongly $\tilde{R}$ -orthogonal to Y . According to Theorem 5.1 and the subsequent remark in Gouriéroux, Laurent and Pham (1998), $\tilde{\theta}$ is then given by

$$\tilde{\theta}_t^i = \psi_t^i + \tilde{\zeta}_t^i \left(\frac{\tilde{E}[H]}{\tilde{Z}_0} + \int_0^t \psi_u dY_u - \psi_t^u Y_t \right), \quad 0 \leq t \leq T, i = 1, \dots, d \quad (4.6)$$

if we note that the relation between their terminology and ours is given by $V(\tilde{a}) = \tilde{Z}/\tilde{Z}_0$, $X^i(\tilde{a}) = \tilde{Z}_0 Y^i$ and $\tilde{a} = -\tilde{\zeta}/\tilde{Z}$. By using Proposition 8 of Rheinländer and Schweizer (1997), (4.6) can be rewritten as

$$\tilde{\theta} = \frac{\tilde{E}[H]}{\tilde{Z}_0} \tilde{\zeta} + \theta \quad (4.7)$$

with θ corresponding to ψ from (4.5) via Equation (4.6) in Rheinländer and Schweizer (1997). Hence it only remains to obtain θ or ψ in terms of the decomposition (4.5) and this is basically already contained in Rheinländer and Schweizer (1997) if one looks carefully enough. More precisely, we start from (4.5) and argue as in Proposition 10 of Rheinländer and Schweizer (1997) to express the quantities in the decomposition (4.2) in terms of ψ and L . Note that as long as we make no integrability assertions, that argument only uses Proposition 8 of Rheinländer and Schweizer (1997) which holds as soon as $\mathbb{P}_e^2(X) \neq \emptyset$; see Remark (2) following that Proposition 8. The uniqueness of the Galtchouk–Kunita–Watanabe decompo-

sition then implies that

$$L_t^{H, \tilde{P}} = \int_0^t \tilde{Z}_u dL_u, \quad 0 \leq t \leq T$$

and

$$\xi_t^{H, \tilde{P}} = \frac{V_0^{H, \tilde{P}}}{\tilde{Z}_0} \tilde{\zeta}_t + \theta_t + L_t - \tilde{\zeta}_t, \quad 0 \leq t \leq T;$$

note that we have to replace $\tilde{E}[H]$ in Equation (4.14) of Rheinländer and Schweizer (1997) by $V_0^{H, \tilde{P}}$ since $\mathcal{F}_0$ need not be trivial. Solving this for θ and plugging the result into (4.7) yields the second expression in (4.4). The first then follows similarly as in the proof of Theorem 6 of Rheinländer and Schweizer (1997); we again have to replace there $\tilde{E}[H]$ by $V_0^{H, \tilde{P}}$. $\square$

While Theorem 4.6 does give a reasonably constructive description of the strategy $\tilde{\theta}$, it is still not completely satisfactory. For continuous-time processes with discontinuous trajectories, hardly anything is known about $\tilde{\theta}$ except under quite restrictive additional assumptions on X . Fairly explicit expressions have been found by Hubalek and Krawczyk (1998) if X is an exponential Lévy process. This relies on earlier results in Schweizer (1994a) who obtained an analogue to (4.4) for the case where X has a deterministic mean-variance tradeoff; see also Grünwald (1998) who used this in a jump-diffusion setting. Somewhat more generally, Hipp (1993, 1996), Wiese (1998) and Pham, Rheinländer and Schweizer (1998) studied the special case where the minimal martingale measure $\hat{P}$ and the variance-optimal martingale measure $\tilde{P}$ coincide. But at present, finding $\tilde{\theta}$ in general is an open problem.

At least for continuous processes X , Theorem 4.6 makes it clear that a key role in determining $\tilde{\theta}$ is played by the variance-optimal martingale measure $\tilde{P}$. For one thing, we need the Galtchouk–Kunita–Watanabe decomposition of H under $\tilde{P}$ just as we needed the Galtchouk–Kunita–Watanabe decomposition of H under $\hat{P}$ in section 3 to find locally risk-minimizing strategies. (This partly explains why the case $\tilde{P} = \hat{P}$ is still solvable.) Thus we have to understand the behaviour of X under $\tilde{P}$ and therefore also the structure of $\tilde{P}$ itself in more detail. In addition, the latter is also required for finding $\tilde{\zeta}$ and $\tilde{Z}$ that appear in (4.4). We first recall a rather special case treated by Pham, Rheinländer and Schweizer (1998).

Lemma 4.7 *Suppose that X is a continuous process such that $\mathbb{P}_e^2(X) \neq \emptyset$. For $Q \in \{\hat{P}, \tilde{P}\}$, we denote by $Z_t^Q := E\left[\frac{dQ}{dP} \mid \mathcal{F}_t\right]$, $0 \leq t \leq T$, the density process of Q with respect to P . If the final value $\hat{K}_T$ of the mean-variance tradeoff is*

deterministic, then $\tilde{P} = \hat{P}$,

$$Z_t^{\tilde{P}} = Z_t^{\hat{P}} = \hat{Z}_t = \mathcal{E} \left(- \int \hat{\lambda} dM \right)_t, \quad 0 \leq t \leq T,$$

$$\tilde{Z}_t = \hat{E} \left[\frac{d\hat{P}}{dP} \middle| \mathcal{F}_t \right] = e^{\hat{K}_T} \mathcal{E} \left(- \int \hat{\lambda} dX \right)_t, \quad 0 \leq t \leq T,$$

$$\tilde{\zeta}_t = -e^{\hat{K}_T} \mathcal{E} \left(- \int \hat{\lambda} dX \right)_t \hat{\lambda}_t = -\tilde{Z}_t \hat{\lambda}_t, \quad 0 \leq t \leq T$$

and

$$\frac{Z_t^{\tilde{P}}}{\tilde{Z}_t} = e^{-(\hat{K}_T - \hat{K}_t)}, \quad 0 \leq t \leq T.$$

Proof Because X satisfies (SC), the three middle results are simply reformulations of Subsection 4.2 of Pham, Rheinländer and Schweizer (1998). The equality of $\tilde{P}$ and $\hat{P}$ is a consequence of the last remark in Section 3 of Pham, Rheinländer and Schweizer (1998) and the last result follows because

$$\tilde{Z}_t = e^{\hat{K}_T} \mathcal{E} \left(- \int \hat{\lambda} dM - \hat{K} \right)_t = e^{\hat{K}_T} Z_t^{\tilde{P}} e^{-\hat{K}_t}. \quad \square$$

Although Lemma 4.7 is a pleasingly simple result, its assumption is usually too restrictive for practical applications. More general results have been obtained by Laurent and Pham (1999) in a multidimensional diffusion model by dynamic programming arguments. They show how one can represent the ratio process $\tilde{Z}/Z^{\tilde{P}}$ as the solution of a dynamic optimization problem and how its canonical decomposition determines the ratio $\tilde{\zeta}/\tilde{Z}$. Current work in progress is aimed at extending these results to general continuous semimartingales, but there still remains a lot to be done because no really explicit results have been found so far. If we consider for instance a *stochastic volatility* model for X , the currently available techniques only work in the case where X and its volatility are uncorrelated. This unfortunately excludes most models of interest for practical applications and illustrates the need for more research in this area. For additional details and more recent work, we refer to Biagini, Guasoni and Pratelli (2000), Guasoni and Biagini (1999), Heath, Platen and Schweizer (2000) and Laurent and Pham (1999).

Acknowledgements

Instead of putting up a very long list of people who would all deserve thanks, I apologize to all those whose work I have forgotten or misrepresented in any way. Thomas Møller pointed out the need to have $\mathcal{F}_0$ non-trivial in Section 4

and Christophe Stricker was as usual extremely helpful with comments and hints on technical issues.

References

- Amendinger, J., Imkeller, P. and Schweizer, M. (1998), Additional logarithmic utility of an insider, *Stochastic Processes and their Applications* **75**, 263–86.
- Ansel, J.P. and Stricker, C. (1992), Lois de martingale, densités et décomposition de Föllmer–Schweizer, *Annales de l'Institut Henri Poincaré* **28**, 375–92.
- Ansel, J.P. and Stricker, C. (1993), Décomposition de Kunita–Watanabe, *Séminaire de Probabilités XXVII, Lecture Notes in Mathematics 1557*, Springer-Verlag, Berlin, 30–32.
- Aurell, E. and Simdyankin, S.I. (1998), Pricing risky options simply, *International Journal of Theoretical and Applied Finance* **1**, 1–23.
- Bertsimas, D., Kogan, L. and Lo, A. (1999), Hedging derivative securities and incomplete markets: an ϵ -arbitrage approach, LFE working paper No. 1027-99R, Sloan School of Management, MIT, Cambridge MA; to appear in *Operations Research*.
- Biagini, F., Guasoni, P. and Pratelli, M. (2000), Mean-variance hedging for stochastic volatility models, *Mathematical Finance* **10**, 109–23.
- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–54.
- Bouleau, N. and Lamberton, D. (1989), Residual risks and hedging strategies in Markovian markets, *Stochastic Processes and their Applications* **33**, 131–50.
- Buckdahn, R. (1993), Backward stochastic differential equations driven by a martingale, preprint, Humboldt University, Berlin (unpublished).
- Černý, A. (1999), Mean-variance hedging in discrete time, preprint, Imperial College Management School, London.
- Choulli, T., Krawczyk, L. and Stricker, C. (1998), $\mathcal{E}$ -martingales and their applications in mathematical finance, *Annals of Probability* **26**, 853–76.
- Choulli, T., Krawczyk, L. and Stricker, C. (1999), On Fefferman and Burkholder–Davis–Gundy inequalities for $\mathcal{E}$ -martingales, *Probability Theory and Related Fields* **113**, 571–97.
- Choulli, T. and Stricker, C. (1996), Deux applications de la décomposition de Galtchouk–Kunita–Watanabe, *Séminaire de Probabilités XXX, Lecture Notes in Mathematics 1626*, Springer-Verlag, Berlin, 12–23.
- Cvitanic, J. and Karatzas, I. (1992), Convex duality in constrained portfolio optimization, *Annals of Applied Probability* **2**, 767–818.
- Davis, M.H.A. (1994), A general option pricing formula, preprint, Imperial College, London.
- Davis, M.H.A. (1997), Option pricing in incomplete markets, in M.A.H. Dempster and S.R. Pliska (eds.), *Mathematics of Derivative Securities*, Cambridge University Press, Cambridge, 216–26.
- Delbaen, F., Monat, P., Schachermayer, W., Schweizer, M. and Stricker, C. (1997), Weighted norm inequalities and hedging in incomplete markets, *Finance and Stochastics* **1**, 181–227.
- Delbaen, F. and Schachermayer, W. (1995), The existence of absolutely continuous local martingale measures, *Annals of Applied Probability* **5**, 926–45.
- Delbaen, F. and Schachermayer, W. (1996a), The variance-optimal martingale measure

- for continuous processes, *BERNOULLI* **2**, 81–105; amendments and corrections (1996), *BERNOULLI* **2**, 379–80.
- Delbaen, F. and Schachermayer, W. (1996b), Attainable claims with p'th moments, *Annales de l'Institut Henri Poincaré* **32**, 743–63.
- Dellacherie, C. and Meyer, P.A. (1982), *Probabilities and Potential B*, North-Holland, Amsterdam.
- Duffie, D. and Richardson, H.R. (1991), Mean-variance hedging in continuous time, *Annals of Applied Probability* **1**, 1–15.
- Föllmer, H. and Schweizer, M. (1989), Hedging by sequential regression: an introduction to the mathematics of option trading, *ASTIN Bulletin* **18**, 147–60.
- Föllmer, H. and Schweizer, M. (1991), Hedging of contingent claims under incomplete information, in M.H.A. Davis and R.J. Elliott (eds.), *Applied Stochastic Analysis, Stochastics Monographs*, Vol. 5, Gordon and Breach, New York, 389–414.
- Föllmer, H. and Sondermann, D. (1986), Hedging of non-redundant contingent claims, in W. Hildenbrand and A. Mas-Colell (eds.), *Contributions to Mathematical Economics*, North-Holland, Amsterdam, 205–23.
- Gouriéroux, C., Laurent, J.P. and Pham, H. (1998), Mean-variance hedging and numéraire, *Mathematical Finance* **8**, 179–200.
- Grandits, P. and Krawczyk, L. (1998), Closedness of some spaces of stochastic integrals, *Séminaire de Probabilités XXXII, Lecture Notes in Mathematics 1686*, Springer-Verlag, Berlin, 73–85.
- Grünewald, B. (1998), *Absicherungsstrategien für Optionen bei Kurssprüngen*, Deutscher Universitäts Verlag, Wiesbaden.
- Grünewald, B. and Trautmann, S. (1997), Varianzminimierende Hedgingstrategien für Optionen bei möglichen Kurssprüngen, in G. Franke (ed.), *Bewertung und Einsatz von Finanzderivaten, Zeitschrift für betriebswirtschaftliche Forschung, Sonderheft* **38**, 43–87.
- Guasoni, P. and Biagini, F. (1999), Mean-variance hedging with random volatility jumps, preprint, University of Pisa; to appear in *Stochastic Analysis and Applications*.
- Harrison, J.M. and Pliska, S.R. (1981), Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–60.
- Heath, D., Platen, E. and Schweizer, M. (2000), A comparison of two quadratic approaches to hedging in incomplete markets, preprint, Technical University of Berlin; to appear in *Mathematical Finance*.
- Hipp, C. (1993), Hedging general claims, *Proceedings of the 3rd AFIR Colloquium, Rome* Vol. 2, 603–13.
- Hipp, C. (1996), Hedging and Insurance Risk, preprint 1/96, University of Karlsruhe.
- Hipp, C. (1998), Hedging general claims in diffusion models, preprint 1/98, University of Karlsruhe.
- Hubalek, F. and Krawczyk, L. (1998), Simple explicit formulae for variance-optimal hedging for processes with stationary independent increments, preprint, University of Vienna.
- Jacod, J. (1979), Calcul stochastique et problèmes de martingales, *Lecture Notes in Mathematics 714*, Springer-Verlag, Berlin.
- Jouini, E. and Napp, C. (1998), Continuous time equilibrium pricing of nonredundant assets, CREST preprint No. 9830, Paris.
- Karatzas, I. (1997), Lectures on the mathematics of finance, *CRM Monograph Series, Vol. 8*, American Mathematical Society, Providence, RI.
- Karatzas, I. and Kou, S.-G. (1996), On the pricing of contingent claims under constraints, *Annals of Applied Probability* **6**, 321–69.

- Korn, R. (1997), Value preserving portfolio strategies in continuous-time models, *Mathematical Methods of Operations Research* **45**, 1–43.
- Korn, R. (1998), Value preserving portfolio strategies and the minimal martingale measure, *Mathematical Methods of Operations Research* **47**, 169–79.
- Lamberton, D., Pham, H. and Schweizer, M. (1998), Local risk-minimization under transaction costs, *Mathematics of Operations Research* **23**, 585–612.
- Laurent, J.P. and Pham, H. (1999), Dynamic programming and mean-variance hedging, *Finance and Stochastics* **3**, 83–110.
- Lepingle, D. and Mémin, J. (1978), Sur l'intégrabilité uniforme des martingales exponentielles, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **42**, 175–203.
- Mercurio, F. (1996), Mean-variance pricing and risk preferences, Tinbergen Institute discussion paper TI 96-44/2, Erasmus University Rotterdam.
- Merton, R.C. (1973), Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–83.
- Møller, T. (1998a), Risk-minimizing hedging strategies for unit-linked life insurance contracts, *ASTIN Bulletin* **28**, 17–47.
- Møller, T. (1998b), Risk-minimizing hedging strategies for insurance payment processes, working paper No. 154, University of Copenhagen; to appear in *Finance and Stochastics*.
- Monat, P. and Stricker, C. (1995), Föllmer–Schweizer decomposition and mean-variance hedging of general claims, *Annals of Probability* **23**, 605–28.
- Pham, H. (2000), On quadratic hedging in continuous time, *Mathematical Methods of Operations Research* **51**, 315–39.
- Pham, H., Rheinländer, T. and Schweizer, M. (1998), Mean-variance hedging for continuous processes: new results and examples, *Finance and Stochastics* **2**, 173–98.
- Pham, H. and Touzi, N. (1996), Equilibrium state prices in a stochastic volatility model, *Mathematical Finance* **6**, 215–36.
- Rheinländer, T. (1999), Optimal martingale measures and their applications in mathematical finance, *PhD thesis*, Technical University of Berlin.
- Rheinländer, T. and Schweizer, M. (1997), On L^2 -projections on a space of stochastic integrals, *Annals of Probability* **25**, 1810–31.
- Schweizer, M. (1988), Hedging of options in a general semimartingale model, *Diss. ETH Zürich* 8615.
- Schweizer, M. (1990), Risk-minimality and orthogonality of martingales, *Stochastics and Stochastics Reports* **30**, 123–31.
- Schweizer, M. (1991), Option hedging for semimartingales, *Stochastic Processes and their Applications* **37**, 339–63.
- Schweizer, M. (1994a), Approximating random variables by stochastic integrals, *Annals of Probability* **22**, 1536–75.
- Schweizer, M. (1994b), Risk-minimizing hedging strategies under restricted information, *Mathematical Finance* **4**, 327–42.
- Schweizer, M. (1995a), On the minimal martingale measure and the Föllmer–Schweizer decomposition, *Stochastic Analysis and Applications* **13**, 573–99.
- Schweizer, M. (1995b), Variance-optimal hedging in discrete time, *Mathematics of Operations Research* **20**, 1–32.
- Schweizer, M. (1996), Approximation pricing and the variance-optimal martingale measure, *Annals of Probability* **24**, 206–36.
- Schweizer, M. (1999a), A minimality property of the minimal martingale measure, *Statistics and Probability Letters* **42**, 27–31.

- Schweizer, M. (1999b), Risky options simplified, *International Journal of Theoretical and Applied Finance* **2**, 59–82.
- Schweizer, M. (2000), From actuarial to financial valuation principles, preprint, Technical University of Berlin; to appear in *Insurance: Mathematics and Economics*.
- Stricker, C. (1990), Arbitrage et lois de martingale, *Annales de l'Institut Henri Poincaré* **26**, 451–60.
- Stricker, C. (1996), The Föllmer–Schweizer Decomposition, in: H.-J. Engelbert, H. Föllmer and J. Zabczyk (eds.), *Stochastic Processes and Related Topics, Stochastics Monographs, Vol. 10*, Gordon and Breach, New York, 77–89.
- Wiese, A. (1998), *Hedging stochastischer Verpflichtungen in zeitstetigen Modellen*, Verlag Versicherungswissenschaft, Karlsruhe.
- Yor, M. (1978), Sous-espaces denses dans L^1 ou H^1 et représentation des martingales, *Séminaire de Probabilités XII, Lecture Notes in Mathematics 649*, Springer-Verlag, Berlin, 265–309.

Part four

Utility Maximization

Theory of Portfolio Optimization in Markets with Frictions

Jakša Cvitanić

1 Introduction

The main topic of this survey is the problem of utility maximization from terminal wealth for a single agent in various financial markets. Specifically, given the agent's utility function $U(\cdot)$ and initial capital $x > 0$, he is trying to maximize the expected utility $E[U(X^{x,\pi}(T))]$ from his “terminal wealth”, over all “admissible” portfolio strategies $\pi(\cdot)$. The same mathematical techniques that we employ here can be used to get similar results for maximizing expected utility from consumption; we refer the interested reader to the rich literature on that problem, some of which is cited below.

The seminal papers on these problems in the continuous-time complete market model are Merton (1969, 1971). Using Itô calculus and a stochastic control/partial differential equations approach, Merton finds a solution to the problem in a Markovian model driven by a Brownian motion process, for logarithmic and power utility functions. A comprehensive survey of his work is Merton (1990). For non-Markovian models one cannot deal with the problem using partial differential equations. Instead, a martingale approach using convex duality has been developed, with remarkable success in solving portfolio optimization problems in diverse frameworks. The approach is particularly well suited for incomplete markets (in which not all contingent claims can be perfectly replicated). It consists of solving an appropriate dual problem over a set of “state-price densities” corresponding to “shadow markets” associated with the incompleteness of the original market. Given the optimal solution $\hat{Z}$ to the dual problem, it is usually possible to show that the optimal terminal wealth for the primal problem is represented as the inverse of “marginal utility” (the derivative of the utility function) evaluated at $\hat{Z}$. Early work in this spirit includes Foldes (1978a,b) and Bismut (1975), based on his stochastic duality theory in Bismut (1973). The first paper using (implicitly) the technique in its modern form, in the complete market, is Pliska (1986), followed

by Karatzas, Lehoczky and Shreve (1987) and Cox and Huang (1989, 1991). The explicit use of the duality method, and in incomplete and/or constrained market models, was applied by Xu (1990), He and Pearson (1991), Xu and Shreve (1992), Karatzas, Lehoczky, Shreve and Xu (1991), Cvitanić and Karatzas (1992, 1993), El Karoui and Quenez (1995), Jouini and Kallal (1995a), Karatzas and Kou (1996), Broadie, Cvitanić and Soner (1998). An excellent exposition of these methods can be found in Karatzas and Shreve (1998), and that of discrete-time models in Pliska (1997); see also Korn (1997). A definite treatment in a very general semimartingale framework is provided in Kramkov and Schachermayer (1998).

A similar approach works in models in which the drift of the wealth process of the agent is concave in his portfolio strategy $\pi(\cdot)$. This includes models with different borrowing and lending rates as well as some “large investor” models. An analytical approach is used in Fleming and Zariphopoulou (1991), Bergman (1995), while the tools of duality are essential in El Karoui, Peng and Quenez (1997), Cvitanić (1997), Cuoco and Cvitanić (1998).

Portfolio optimization problems under transaction costs, usually on an infinite horizon $T = \infty$, have been studied mostly in Markovian models, using PDE/variational inequalities methods. The literature includes Magill and Constantinides (1976), Constantinides (1979), Taksar, Klass and Assaf (1988), Davis and Norman (1990), Zariphopoulou (1992), Shreve and Soner (1994), and Morton and Pliska (1995). We follow the martingale/duality approach of Cvitanić and Karatzas (1996) and Cvitanić and Wang (1999), on the finite horizon $T < \infty$. While this method is powerful enough to guarantee existence and a characterization of the optimal solution, algorithms for actually finding the optimal strategy are still lacking.

In order to apply the martingale approach to portfolio optimization, we first have to resolve the problem of (super)replication of contingent claims in a given market. After presenting the continuous-time complete market model and recalling the classical Black–Scholes–Merton pricing in Sections 2 and 3, we find the minimal cost of superreplicating a given claim B under convex constraints on the proportions of wealth the agent invests in stocks, in Sections 4 and 5 (for much more general results of this kind see Föllmer and Kramkov (1997)). In the complete market this cost of superreplication of B is equal to the Black–Scholes price of B , which is equal to the expected value of B (discounted), under a change of probability measure that makes the discounted prices of stocks martingales.

In the case of a constrained market, in which the agent’s hedging portfolio has to take values in a given closed convex set K , it is shown that the minimal cost of superreplication is now a supremum of Black–Scholes prices, taken over a family of auxiliary markets, parametrized by processes $\nu(\cdot)$, taking values in the domain of the support function of the set $-K$. These markets are chosen so that the wealth

process becomes a supermartingale, under the appropriate change of measure. In the constant market parameters framework, the minimal cost for superreplicating B under constraints can be calculated as the Black–Scholes (unconstrained) price of an appropriately modified contingent claim $\hat{B} \geq B$, and the hedging portfolio for $\hat{B}$ automatically satisfies the constraints.

In Section 6 we show how the same methodology can be used to get analogous results in a market in which the drift of the wealth process is a concave function of the portfolio process.

Section 7 introduces the concept of utility functions, and the existence of an optimal constrained portfolio strategy for maximizing expected utility from terminal wealth is proved in Section 8. This is done indirectly, by first solving a dual problem, which is, loosely speaking, a problem of finding an optimal change of probability measure associated with the constrained market. The optimal portfolio policy is the one that replicates the inverse of marginal utility, evaluated at the Radon–Nikodym derivative corresponding to the optimal change of measure in the dual problem. Explicit solutions are provided in Section 9, for the case of logarithmic and power utilities. Next, in Section 10 we argue that it makes sense to price contingent claims in the constrained market by calculating the Black–Scholes price in the unconstrained auxiliary market that corresponds to the optimal dual change of measure. Although in general this price depends on the utility of the agent and his initial capital, in many cases it does not. In particular, if the constraints are given by a cone, and the market parameters are constant, the optimal dual process is independent of utility and initial capital. This approach to pricing in incomplete markets was suggested in Davis (1997) and further developed in Karatzas and Kou (1996).

In Sections 11–15 we study the superreplication and utility maximization problems in the presence of proportional transaction costs. Similarly as in the case of constraints, we identify the family of (pairs of) changes of probability measure, under which the “wealth process” is a supermartingale, and the supremum over which gives the minimal superreplication cost of a claim in this market. Representations of this type were obtained in various models in Jouini and Kallal (1995b), Kusuoka (1995), and Kabanov (1999). (It is known that in standard diffusion models this cost is simply the cost of the least expensive static (buy-and-hold) strategy which superreplicates the claim. For the case of the European call it is then equal to the price of one share of the underlying, the result which was conjectured by Davis and Clarke (1994) and proved by Soner, Shreve and Cvitanić (1995). The same result was shown to hold for more general models and claims in Levental and Skorohod (1997) and Cvitanić, Pham and Touzi (1998).) Next, we consider the utility maximization problem under transaction costs, and its dual. The nature of the optimal terminal wealth in the primal problem is shown to be the same as in

the case of constraints – it is equal to the inverse of the marginal utility evaluated at the optimal dual solution. This result is used to get sufficient conditions for the optimal policy to be the one of no trade at all – this is the case if the return rate of the stock is not very different from the interest rate of the bank account and the transaction costs are large relative to the time horizon.

The important topic which is not considered here is approximate hedging and pricing under transaction costs. Articles dealing with this problem in continuous-time include Leland (1985), Avellaneda and Parás (1993), Davis, Panas and Zariphopoulou (1993), Davis and Panas (1994), Davis and Zariphopoulou (1995), Barles and Soner (1998), Constantinides and Zariphopoulou (1999). Other related works on the the subject of transaction costs of which the reader may find useful to consult are: Bensaid, Lesne, Pagès and Scheinkman (1992), Boyle and Vorst (1992), Edirisinghe, Naik and Uppal (1993), Flesaker and Hughston (1994), Gilster and Lee (1984), Grannan and Swindle (1996), Hodges and Neuberger (1989), Hoggard, Whalley and Wilmott (1994), Merton (1989), Morton and Pliska (1995).

2 The complete market model

We introduce here the standard, Itô processes model for a financial market $\mathcal{M}$. It consists of one *bank-account* and d *stocks*. Price processes $S_0(\cdot)$ and $S_1(\cdot), \dots, S_d(\cdot)$ of these instruments are modeled by the equations

$$\begin{aligned} dS_0(t) &= S_0(t)r(t)dt, \quad S_0(0) = 1, \\ dS_i(t) &= S_i(t) \left[b_i(t)dt + \sum_{j=1}^d \sigma_{ij}(t)dW^j(t) \right], \quad S_i(0) = s_i > 0, \quad (2.1) \end{aligned}$$

for $i = 1, \dots, d$, on some given time horizon $[0, T]$, $0 < T < \infty$. Here $W(\cdot) = (W^1(\cdot), \dots, W^d(\cdot))'$ is a standard d -dimensional Brownian motion on a complete probability space $(\Omega, \mathcal{F}, P)$, endowed with a filtration $\mathbf{F} = \{\mathcal{F}_t\}_{0 \leq t \leq T}$, the P -augmentation of $\mathcal{F}^W(t) := \sigma(W(s); 0 \leq s \leq t)$, $0 \leq t \leq T$, the filtration generated by the Brownian motion $W(\cdot)$. The *coefficients* $r(\cdot)$ (interest rate), $b(\cdot) = (b_1(\cdot), \dots, b_d(\cdot))'$ (vector of stock return rates) and $\sigma(\cdot) = \{\sigma_{ij}(\cdot)\}_{1 \leq i, j \leq d}$ (matrix of stock-volatilities) of the model $\mathcal{M}$, are all assumed to be progressively measurable with respect to $\mathbf{F}$. Furthermore, the matrix $\sigma(\cdot)$ is assumed to be invertible, and all processes $r(\cdot)$, $b(\cdot)$, $\sigma(\cdot)$, $\sigma^{-1}(\cdot)$ are assumed to be bounded, uniformly in $(t, \omega) \in [0, T] \times \Omega$.

The “risk premium” process

$$\theta_0(t) := \sigma^{-1}(t)[b(t) - r(t)\mathbf{1}], \quad 0 \leq t \leq T \quad (2.2)$$

where $\mathbf{1} = (1, \dots, 1)' \in \mathbb{R}^d$, is then bounded and $\mathbf{F}$ -progressively measurable. Therefore, the process

$$Z_0(t) := \exp \left[- \int_0^t \theta'_0(s) dW_0(s) - \frac{1}{2} \int_0^t \|\theta_0(s)\|^2 ds \right], \quad 0 \leq t \leq T \quad (2.3)$$

is a P -martingale, and

$$P_0(\Lambda) := E[Z_0(T)1_\Lambda], \quad \Lambda \in \mathcal{F}_T \quad (2.4)$$

is a probability measure equivalent to P on $\mathcal{F}_T$. Under this *risk-neutral equivalent martingale measure* P_0 , the discounted stock prices $S_1(\cdot)/S_0(\cdot), \dots, S_d(\cdot)/S_0(\cdot)$ become martingales, and the process

$$W_0(t) := W(t) + \int_0^t \theta_0(s) ds, \quad 0 \leq t \leq T, \quad (2.5)$$

becomes Brownian motion, by the Girsanov theorem.

We also introduce the discount process

$$\gamma_0(t) := e^{-\int_0^t r(u) du}, \quad 0 \leq t \leq T. \quad (2.6)$$

and “state price density” process

$$H_0(t) := \gamma_0(t)Z_0(t), \quad 0 \leq t \leq T. \quad (2.7)$$

Consider now a financial agent whose actions cannot affect market prices, and who can decide, at any time $t \in [0, T]$, what proportion $\pi_i(t)$ of his (nonnegative) wealth $X(t)$ to invest in the i -th stock ($1 \leq i \leq d$). Of course these decisions can only be based on the current information $\mathcal{F}_t$, without anticipation of the future. With $\pi(t) = (\pi_1(t), \dots, \pi_d(t))'$ chosen, the amount $X(t)[1 - \sum_{i=1}^d \pi_i(t)]$ is invested in the bank. Thus, in light of the dynamics (2.1), the wealth process $X(\cdot) \equiv X^{x, \pi, c}(\cdot)$ satisfies the linear stochastic differential equation

$$\begin{aligned} dX(t) &= -dc(t) + \left[X(t)(1 - \sum_{i=1}^d \pi_i(t)) \right] r(t)dt \\ &\quad + \sum_{i=1}^d \pi_i(t) X(t-) \left[b_i(t)dt + \sum_{j=1}^d \sigma_{ij}(t) dW^j(t) \right] \\ &= -dc(t) + r(t)X(t)dt + \pi'(t)\sigma(t)X(t-)dW_0(t); \quad X(0) = x, \end{aligned}$$

where the real number $x > 0$ represents initial capital and $c(\cdot) \geq 0$ denotes the agent's *cumulative consumption process*.

We formalize the above discussion as follows.

Definition 2.1

- (i) A *portfolio process* $\pi : [0, T] \times \Omega \rightarrow \mathbb{R}^d$ is $\mathbf{F}$ -progressively measurable and satisfies $\int_0^T \|X(t)\pi(t)\|^2 dt < \infty$, almost surely (here, X is the corresponding wealth process defined below). A *consumption process* $c(\cdot)$ is a nonnegative, nondecreasing, progressively measurable process with RCLL paths, with $c(0) = 0$ and $c(T) < \infty$.
- (ii) For a given portfolio and consumption processes $\pi(\cdot)$, $c(\cdot)$, the process $X(\cdot) \equiv X^{x,\pi,c}(\cdot)$ defined by (2.9) below, is called the *wealth process* corresponding to strategy (π, c) and initial capital x .
- (iii) A portfolio-consumption process pair $(\pi(\cdot), c(\cdot))$ is called *admissible* for the initial capital x , and we write $(\pi, c) \in \mathcal{A}_0(x)$, if

$$X^{x,\pi,c}(t) \geq 0, \quad 0 \leq t \leq T \quad (2.8)$$

holds almost surely.

For the discounted version of process $X(\cdot)$, we get the equation

$$d(\gamma_0(t)X(t)) = -\gamma_0(t)dc(t) + \pi'(t)\sigma(t)\gamma_0(t)X(t-)dW_0(t). \quad (2.9)$$

It follows that $\gamma_0(\cdot)X(\cdot)$ is a nonnegative local P_0 -supermartingale, hence also a P_0 -supermartingale, by Fatou's lemma. Therefore, if τ_0 is defined to be the first time it hits zero, we have $X(t) = 0$ for $t \geq \tau_0$, so that the portfolio values $\pi(t)$ are irrelevant after that happens. Accordingly, we can and do set $\pi(t) \equiv 0$ for $t \geq \tau_0$. The supermartingale property implies

$$E^0[\gamma_0(T)X^{x,\pi,c}(T)] \leq x, \quad \forall \pi \in \mathcal{A}_0(x). \quad (2.10)$$

Here, E^0 denotes the expectation operator under the measure P_0 .

We say that a strategy $(\pi(\cdot), c(\cdot))$ results in *arbitrage* if with the initial investment $x = 0$ we have $X^{0,\pi,c}(T) \geq 0$ almost surely, but $X^{0,\pi,c}(T) > 0$ with positive probability. Notice that inequality (2.10) implies that an admissible strategy $(\pi(\cdot), c(\cdot)) \in \mathcal{A}_0(0)$ cannot result in arbitrage.

3 Pricing in the complete market

Let us suppose now that the agent promises to pay a random amount $B(\omega) \geq 0$ at time $t = T$ and that he wants to invest x dollars in the market in such a way that his profit “hedges away” all the risk, specifically that $X^{x,\pi,c}(T) \geq B$, almost surely. What is the smallest value of $x > 0$ for which such “hedging” is possible? This smallest value will then be the “price” of the *contingent claim* B at time $t = 0$.

We say that B is a *contingent claim* if it is a nonnegative, $\mathcal{F}_T$ -measurable random variable such that $0 < E^0[\gamma_0(T)B] < \infty$. The *superreplication price* of this

contingent claim is defined by

$$h(0) := \inf\{x > 0; \exists(\pi, c) \in \mathcal{A}_0(x) \text{ s.t. } X^{x,\pi,c}(T) \geq B \text{ a.s.}\}. \quad (3.1)$$

The following classical result identifies $h(0)$ as the expectation, under the risk-neutral probability measure, of the claim's discounted value; see Harrison and Kreps (1979), Harrison and Pliska (1981, 1983).

Proposition 3.1 *The infimum in (3.1) is attained, and we have*

$$h(0) = E^0[\gamma_0(T)B]. \quad (3.2)$$

Furthermore, there exists a portfolio $\pi_B(\cdot)$ such that $X_B(\cdot) \equiv X^{h(0),\pi_B,0}(\cdot)$ is given by

$$X_B(t) = \frac{1}{\gamma_0(t)} E^0[\gamma_0(T)B|\mathcal{F}_t], \quad 0 \leq t \leq T. \quad (3.3)$$

Proof Suppose $X^{x,\pi,c}(T) \geq B$ holds a.s. for some $x \in (0, \infty)$ and a suitable $(\pi, c) \in \mathcal{A}_0(x)$. Then from (2.10) we have $x \geq z := E^0[\gamma_0(T)B]$ and thus $h(0) \geq z$.

On the other hand, from the martingale representation theorem, the process

$$X_B(t) := \frac{1}{\gamma_0(t)} E^0[\gamma_0(T)B|\mathcal{F}_t], \quad 0 \leq t \leq T$$

can be represented as

$$X_B(t) = \frac{1}{\gamma_0(t)} \left[z + \int_0^t \psi'(s) dW_0(s) \right]$$

for a suitable $\{\mathcal{F}_t\}$ -progressively measurable process $\psi(\cdot)$ with values in $\mathbb{R}^d$ and $\int_0^T \|\psi(t)\|^2 dt < \infty$, a.s. Then $\pi_B(t) := (\gamma_0(t)X_B(t-))^{-1}(\sigma'(t))^{-1}\psi(t)$ is a well defined portfolio process, and we have $X_B(\cdot) \equiv X^{z,\pi_B,0}(\cdot)$, by comparison with (2.9). Therefore, $z \geq h(0)$. $\square$

Notice that

$$X_B^{h(0),\pi_B,0}(T) = B,$$

almost surely. We express this by saying that contingent claim B is *attainable*, with initial capital $h(0)$ and portfolio π_B . In this complete market model, we call $h(0)$ the *Black–Scholes price* of B and $\pi_B(\cdot)$ the *Black–Scholes hedging portfolio*.

Example 3.2 *Constant* $r(\cdot) \equiv r > 0$, $\sigma(\cdot) \equiv \sigma$ *nonsingular*. In this case, the solution $S(t) = (S_1(t), \dots, S_d(t))'$ is given by $S_i(t) = f_i(t-s, S(s), \sigma(W_0(t)-$

$W_0(s)))$, $0 \leq s \leq t$, where $f : [0, \infty) \times \mathbb{R}_+^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+^d$ is the function defined by

$$f_i(t, s, y; r) := s_i \exp \left[\left(r - \frac{1}{2} a_{ii} \right) t + y_i \right], \quad i = 1, \dots, d,$$

where $a = \sigma \sigma'$.

Consider now a contingent claim of the type $B = \varphi(S(T))$, where $\varphi : \mathbb{R}_+^d \rightarrow [0, \infty)$ is a given continuous function, that satisfies polynomial growth conditions in both $\|s\|$ and $1/\|s\|$. Then the value process of this claim is given by

$$\begin{aligned} X_B(t) &= e^{-r(T-t)} E^0[\varphi(S(T)) | \mathcal{F}_t] \\ &= e^{-r(T-t)} \int_{\mathbb{R}^d} \varphi(f(T-t, S(t), \sigma z)) \frac{1}{(2\pi(T-t))^{d/2}} \exp \left\{ -\frac{\|z\|^2}{2(T-t)} \right\} dz \\ &= V(T-t, S(t)), \end{aligned}$$

where

$$V(t, p) := \begin{cases} e^{-rt} \int_{\mathbb{R}^d} \varphi(h(t, s, \sigma z; r)) \frac{e^{-\|z\|^2/2t}}{(2\pi t)^{d/2}} dz; & t > 0, s \in \mathbb{R}_+^d \\ \varphi(s); & t = 0, s \in \mathbb{R}_+^d \end{cases}.$$

In particular, the price $h(0)$ of the claim B is given, in terms of the function V , by

$$h(0) = X_B(0) = V(T, S(0)).$$

Moreover, function V is the unique solution to the Cauchy problem (by the Feynman–Kac theorem)

$$\frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d a_{ij} x_i x_j \frac{\partial^2 V}{\partial x_i \partial x_j} + \sum_{i=1}^d r \left(x_i \frac{\partial V}{\partial x_i} - V \right) = \frac{\partial V}{\partial t},$$

with the initial condition $V(0, x) = \varphi(x)$. Applying Itô's rule, we obtain

$$dV(T-t, S(t)) = rV(T-t, S(t)) + \sum_{i=1}^d \sum_{j=1}^d \sigma_{ij} S_i(t) \frac{\partial V}{\partial x_i}(T-t, S_i(t)) dW_0^{(j)}(t).$$

Comparing this with (2.9), we get that the hedging portfolio is given by

$$\pi_i(t) V(T-t, S(t)) = S_i(t) \frac{\partial V}{\partial x_i}(T-t, S(t)), \quad i = 1, \dots, d.$$

It should be noted that none of the above depends on the vector $b(\cdot)$ of return rates.

If, for example, we have $d = 1$ and in the case $\varphi(s) = (s - k)^+$ of a *European call option*, with $\sigma = \sigma_{11} > 0$, exercise price $k > 0$, $N(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-u^2/2} du$

and $d_{\pm}(t, s) := \frac{1}{\sigma\sqrt{t}} \left[\log\left(\frac{s}{k}\right) + (r \pm \frac{\sigma^2}{2})t \right]$, we have the famous *Black and Scholes (1973) formula*

$$V(t, s) = \begin{cases} sN(d_+(t, s)) - ke^{-rt}N(d_-(t, s)); & t > 0, s \in (0, \infty) \\ (s - k)^+; & t = 0, s \in (0, \infty) \end{cases}.$$

4 Portfolio constraints

We fix throughout a nonempty, closed, convex set K in $\mathbb{R}^d$, and denote by

$$\delta(x) := \sup_{\pi \in K} \{-\pi'x\} \quad (4.1)$$

the support function of the set $-K$. This is a closed, positively homogeneous, proper convex function on $\mathbb{R}^d$ (Rockafellar (1970), p. 114). It is finite on its *effective domain*

$$\tilde{K} := \{x \in \mathbb{R}^d / F\delta(x) < \infty\} \quad (4.2)$$

which is a convex cone (called the “barrier cone” of $-K$). For the rest of the paper we assume the following mild conditions.

Assumption 4.1 The closed convex set $K \subset \mathbb{R}^d$ contains the origin; in other words, the agent is allowed not to invest in stocks at all. In particular, $\delta(\cdot) \geq 0$ on $\tilde{K}$. Moreover, the set K is such that $\delta(\cdot)$ is continuous on the barrier cone $\tilde{K}$ of (4.2).

The role of the closed, convex set K that we just introduced is to model reasonable constraints on portfolio choice. One may, for instance, consider the following examples.

- (i) *Unconstrained case:* $K = \mathbb{R}^d$. Then $\tilde{K} = \{0\}$, and $\delta \equiv 0$ on $\tilde{K}$.
- (ii) *Prohibition of short-selling:* $K = [0, \infty)^d$. Then $\tilde{K} = K$, and $\delta \equiv 0$ on $\tilde{K}$.
- (iii) *Incomplete Market:* $K = \{\pi \in \mathbb{R}^d; \pi_i = 0, \forall i = m+1, \dots, d\}$ for some fixed $m \in \{1, \dots, d-1\}$. Then $\tilde{K} = \{x \in \mathbb{R}^d; x_i = 0, \forall i = 1, \dots, m\}$ and $\delta \equiv 0$ on $\tilde{K}$.
- (iv) *K is a closed, convex cone in $\mathbb{R}^d$.* Then $\tilde{K} = \{x \in \mathbb{R}^d; \pi'x \geq 0, \forall \pi \in K\}$ is the polar cone of $-K$, and $\delta \equiv 0$ on $\tilde{K}$. This case obviously generalizes (i)–(iii).
- (v) *Prohibition of borrowing:* $K = \{\pi \in \mathbb{R}^d; \sum_{i=1}^d \pi_i \leq 1\}$. Then $\tilde{K} = \{x \in \mathbb{R}^d; x_1 = \dots = x_d \leq 0\}$, and $\delta(x) = -x_1$ on $\tilde{K}$.
- (vi) *Rectangular constraints:* $K = \times_{i=1}^d I_i$, $I_i = [\alpha_i, \beta_i]$ for some fixed numbers $-\infty \leq \alpha_i \leq 0 \leq \beta_i \leq \infty$, with the understanding that the interval I_i is open to the right (left) if $\beta_i = \infty$ (respectively, if $\alpha_i = -\infty$). Then $\delta(x) = \sum_{i=1}^d (\beta_i x_i^- - \alpha_i x_i^+)$ and $\tilde{K} = \mathbb{R}^d$ if all the α_i 's, β_i 's are real. In general,

$\tilde{K} = \{x \in \mathbb{R}^d; x_i \geq 0, \forall i \in \mathcal{S}_+ \text{ and } x_j \leq 0, \forall j \in \mathcal{S}_-\}$ where $\mathcal{S}_+ := \{i = 1, \dots, d/\beta_i = \infty\}$, $\mathcal{S}_- := \{i = 1, \dots, d/\alpha_i = -\infty\}$.

We consider now only portfolios that take values in the given, convex, closed set $K \subset \mathbb{R}^d$, i.e., we replace the set of admissible policies $\mathcal{A}_0(x)$ with

$$\mathcal{A}'(x) := \{(\pi, c) \in \mathcal{A}_0(x); \pi(t, \omega) \in K \text{ for } \ell \otimes \mathbf{P}\text{-a.e. } (t, \omega)\}.$$

Here, ℓ stands for Lebesgue measure on $[0, T]$.

Denote by $\mathcal{D}$ the set of all *bounded* progressively measurable processes $v(\cdot)$ taking values in $\tilde{K}$ a.e. on $\Omega \times [0, T]$. In analogy with (2.2)–(2.5), introduce

$$\theta_v(t) := \sigma^{-1}(t)[v(t) + b(t) - r(t)\mathbf{1}], \quad 0 \leq t \leq T, \quad (4.3)$$

$$Z_v(t) := \exp \left[- \int_0^t \theta'_v(s) dW(s) - \frac{1}{2} \int_0^t \|\theta_v(s)\|^2 ds \right], \quad 0 \leq t \leq T, \quad (4.4)$$

$$P_v(\Lambda) := E[Z_v(T)1_\Lambda], \quad \Lambda \in \mathcal{F}_T \quad (4.5)$$

$$W_v(t) := W(t) + \int_0^t \theta_v(s) ds, \quad 0 \leq t \leq T, \quad (4.6)$$

a P^v -Brownian motion. Also denote

$$\gamma_v(t) := e^{-\int_0^t [r(u) + \delta(v(u))] du} \quad (4.7)$$

and

$$H_v(t) := \gamma_v(t) Z_v(t). \quad (4.8)$$

Proposition 4.2 *The process*

$$M_v(t) := H_v(t)X(t) + \int_0^t H_v(s) [X(s)(\delta(v_s) + v'(s)\pi(s))ds + dc(s)]$$

is a P -supermartingale for every $v \in \mathcal{D}$ and $(\pi, c) \in \mathcal{A}'(x)$. In particular,

$$\sup_{v \in \mathcal{D}} E \left[H_v(T)X(T) + \int_0^T H_v(s)X(s)\{\delta(v_s) + \pi'(s)v(s)\}ds \right] \leq x. \quad (4.9)$$

Proof Itô's rule implies

$$M_v(t) = x + \int_0^t H_v(s)X(s) [\pi'(s)\sigma(s) - \theta'_v(s)] dW(s).$$

In particular, the process on the right-hand side is a nonnegative local martingale, hence a supermartingale. $\square$

In general, there are several interpretations for the processes $v \in \mathcal{D}$: they are stochastic “Lagrange multipliers” associated with the portfolio constraints; in economics jargon, they correspond to the shadow prices relevant to the incompleteness of the market introduced by constraints. The number $h_v(0) := E^v[\gamma_v(T)B] = E[H_v(T)B]$ is the unconstrained hedging price for B in an *auxiliary market* $\mathcal{M}_v$; this market consists of a bank-account with interest rate $r^{(v)}(t) := r(t) + \delta(v(t))$ and d stocks, with the same volatility matrix $\{\sigma_{ij}(t)\}_{1 \leq i, j \leq d}$ as before and return rates $b_i^{(v)}(t) := b_i(t) + v_i(t) + \delta(v(t))$, $1 \leq i \leq d$, for any given $v \in \mathcal{D}$. We shall show that the price for superreplicating B with a constrained portfolio in the market $\mathcal{M}$ is given by the supremum of the unconstrained hedging prices $h_v(0)$ in these auxiliary markets $\mathcal{M}_v$, $v \in \mathcal{D}$.

5 Superreplication under portfolio constraints

Consider the minimal cost of superreplication of the claim B in the market with constraints:

$$h(0) := \begin{cases} \inf\{x > 0; \exists (\pi, c) \in \mathcal{A}'(x), \text{ s.t. } X^{x, \pi, c}(T) \geq B \text{ a.s.}\} \\ \infty, \text{ if the above set is empty} \end{cases}.$$

Let us denote by $\mathcal{S}$ the set of all $\{\mathcal{F}_t\}$ -stopping times τ with values in $[0, T]$, and by $\mathcal{S}_{\rho, \sigma}$ the subset of $\mathcal{S}$ consisting of stopping times τ s.t. $\rho \leq \tau \leq \sigma$, for any two $\rho \in \mathcal{S}, \sigma \in \mathcal{S}$ such that $\rho \leq \sigma$, a.s. For every $\tau \in \mathcal{S}$ consider also the $\mathcal{F}_\tau$ -measurable random variable

$$V(\tau) := \text{ess sup}_{v \in \mathcal{D}} E^v[B\gamma_0(T) \exp\{-\int_\tau^T \delta(v(s))ds\} | \mathcal{F}_\tau]. \quad (5.1)$$

We will show that $h(0) = V(0)$. We first need

Proposition 5.1 *If $V(0) = \sup_{v \in \mathcal{D}} E^v[\gamma_v(T)B] < \infty$, then the family of random variables $\{V(\tau)\}_{\tau \in \mathcal{S}}$ satisfies the equation of Dynamic Programming*

$$V(\tau) = \text{ess sup}_{v \in \mathcal{D}_{\tau, \theta}} E^v[V(\theta) \exp\{-\int_\tau^\theta \delta(v(u))du\} | \mathcal{F}_\tau]; \quad \forall \theta \in \mathcal{S}_{\tau, T}, \quad (5.2)$$

where $\mathcal{D}_{\tau, \theta}$ is the restriction of $\mathcal{D}$ to the stochastic interval $[\tau, \theta]$.

Proposition 5.2 *The process $V = \{V(t), \mathcal{F}_t; 0 \leq t \leq T\}$ can be considered in its RCLL modification and, for every $v \in \mathcal{D}$,*

$$\left\{ \begin{array}{l} Q_v(t) := V(t)e^{-\int_0^t \delta(v(u))du}, \mathcal{F}_t; \quad 0 \leq t \leq T \\ \text{is a } \mathbf{P}^v\text{-supermartingale with RCLL paths} \end{array} \right\}. \quad (5.3)$$

Furthermore, V is the smallest adapted, RCLL process that satisfies (5.3) as well as

$$V(T) = B\gamma_0(T), \quad \text{a.s.} \quad (5.4)$$

Proof of Proposition 5.1 Let us start by observing that, for any $\theta \in \mathcal{S}$, the random variable

$$\begin{aligned} J_v(\theta) &:= E^v[V(T)e^{-\int_\theta^T \delta(v(s))ds} | \mathcal{F}_\theta] \\ &= \frac{E[Z_v(\theta)Z_v(\theta, T)V(T)e^{-\int_\theta^T \delta(v(s))ds} | \mathcal{F}_\theta]}{E[Z_v(\theta)Z_v(\theta, T) | \mathcal{F}_\theta]} \\ &= E[Z_v(\theta, T)V(T)e^{-\int_\theta^T \delta(v(s))ds} | \mathcal{F}_\theta] \end{aligned}$$

depends only on the restriction of v to $[\![\theta, T]\!]$ (we have used the notation $Z_v(\theta, T) = Z_v(T)/Z_v(\theta)$). It is also easy to check that the family of random variables $\{J_v(\theta)\}_{v \in \mathcal{D}}$ is directed upwards; indeed, for any $\mu \in \mathcal{D}$, $v \in \mathcal{D}$ and with $A = \{(t, \omega); J_\mu(t, \omega) \geq J_v(t, \omega)\}$ the process $\lambda := \mu \mathbf{1}_A + v \mathbf{1}_{A^c}$ belongs to $\mathcal{D}$ and we have a.s. $J_\lambda(\theta) = \min\{J_\mu(\theta), J_v(\theta)\}$; then from Neveu (1975), p. 121, there exists a sequence $\{v_k\}_{k \in \mathbb{N}} \subseteq \mathcal{D}$ such that $\{J_{v_k}(\theta)\}_{k \in \mathbb{N}}$ is increasing and

$$(i) \quad V(\theta) = \lim_{k \rightarrow \infty} \uparrow J_{v_k}(\theta), \quad \text{a.s.}$$

Returning to the proof itself, let us observe that

$$\begin{aligned} V(\tau) &= \text{ess sup}_{v \in \mathcal{D}_{\tau, T}} E^v[e^{-\int_\tau^\theta \delta(v(s))ds} E^v\{V(T)e^{-\int_\theta^T \delta(v(s))ds} | \mathcal{F}_\theta\} | \mathcal{F}_\tau] \\ &\leq \text{ess sup}_{v \in \mathcal{D}_{\tau, T}} E^v[e^{-\int_\tau^\theta \delta(v(s))ds} V(\theta) | \mathcal{F}_\tau], \quad \text{a.s.} \end{aligned}$$

To establish the opposite inequality, it certainly suffices to pick $\mu \in \mathcal{D}$ and show that

$$(ii) \quad V(\tau) \geq E^\mu[V(\theta)e^{-\int_\tau^\theta \delta(\mu(s))ds} | \mathcal{F}_\tau]$$

holds almost surely.

Let us denote by $M_{\tau, \theta}$ the class of processes $v \in \mathcal{D}$ which agree with μ on $[\![\tau, \theta]\!]$. We have

$$\begin{aligned} V(\tau) &\geq \text{ess sup}_{v \in M_{\tau, \theta}} E^v[e^{-\int_\tau^\theta \delta(v(s))ds - \int_\theta^T \delta(v(s))ds} V(T) | \mathcal{F}_\tau] \\ &= \text{ess sup}_{v \in M_{\tau, \theta}} E^v[e^{-\int_\tau^\theta \delta(v(s))ds} E^v\{e^{-\int_\theta^T \delta(v(s))ds} V(T) | \mathcal{F}_\theta\} | \mathcal{F}_\tau]. \end{aligned}$$

Thus, for every $v \in M_{\tau, \theta}$, we have

$$\begin{aligned}
 V(\tau) &\geq E^v[e^{-\int_{\tau}^{\theta} \delta(v(s))ds} J_v(\theta) | \mathcal{F}_{\tau}] \\
 &= \frac{E[Z_v(\tau) Z_v(\tau, \theta) E\{Z_v(\theta, T) | \mathcal{F}_{\theta}\} e^{-\int_{\tau}^{\theta} \delta(v(s))ds} J_v(\theta) | \mathcal{F}_{\tau}]}{E[Z_v(\tau) Z_v(\tau, \theta) E\{Z_v(\theta, T) | \mathcal{F}_{\theta}\} | \mathcal{F}_{\tau}]} \\
 &= E[Z_v(\tau, \theta) e^{-\int_{\tau}^{\theta} \delta(v(s))ds} J_v(\theta) | \mathcal{F}_{\tau}] \\
 &= E[Z_{\mu}(\tau, \theta) e^{-\int_{\tau}^{\theta} \delta(\mu(s))ds} J_v(\theta) | \mathcal{F}_{\tau}] \\
 &= \dots = E^{\mu}[e^{-\int_{\tau}^{\theta} \delta(\mu(s))ds} J_v(\theta) | \mathcal{F}_{\tau}].
 \end{aligned}$$

Now clearly we may take $\{v_k\}_{k \in \mathbb{N}} \subseteq M_{\tau, \theta}$ in (i), as $J_v(\theta)$ depends only on the restriction of v on $[\![\theta, T]\!]$; and from the above,

$$\begin{aligned}
 V(\tau) &\geq \lim_{k \rightarrow \infty} \uparrow E^{\mu}[e^{-\int_{\tau}^{\theta} \delta(\mu(s))ds} J_{v_k}(\theta) | \mathcal{F}_{\tau}] \\
 &= E^{\mu}[e^{-\int_{\tau}^{\theta} \delta(\mu(s))ds} \lim_{k \rightarrow \infty} \uparrow J_{v_k}(\theta) | \mathcal{F}_{\tau}] \\
 &= E^{\mu}[e^{-\int_{\tau}^{\theta} \delta(\mu(s))ds} V(\theta) | \mathcal{F}_{\tau}], \text{ a.s.}
 \end{aligned}$$

by monotone convergence.

It is an immediate consequence of this proposition that

$$(iii) \quad V(\tau) e^{-\int_0^{\tau} \delta(v(u))du} \geq E^v[V(\theta) e^{-\int_0^{\theta} \delta(v(u))du} | \mathcal{F}_{\tau}], \text{ a.s.}$$

holds for any given $\tau \in \mathcal{S}$, $\theta \in \mathcal{S}_{\tau, T}$ and $v \in \mathcal{D}$. □

Proof of Proposition 5.2 Let us consider the positive, adapted process $\{V(t, \omega), \mathcal{F}_t; t \in [0, T] \cap \mathcal{Q}\}$ for $\omega \in \Omega$. From (iii), the process

$$\{V(t, \omega) e^{-\int_0^t \delta(v(s, \omega))ds}, \mathcal{F}_t; t \in [0, T] \cap \mathcal{Q}\} \text{ for } \omega \in \Omega$$

is a $\mathbf{P}^v$ -supermartingale on $[0, T] \cap \mathcal{Q}$, where $\mathcal{Q}$ is the set of rational numbers, and thus has a.s. finite limits from the right and from the left (recall Proposition 1.3.14 in Karatzas and Shreve (1991), as well as the right-continuity of the filtration $\{\mathcal{F}_t\}$). Therefore,

$$\begin{aligned}
 V(t+, \omega) &:= \begin{cases} \lim_{\substack{s \downarrow t \\ s \in \mathcal{Q}}} V(s, \omega); & 0 \leq t < T \\ V(T, \omega); & t = T \end{cases} \\
 V(t-, \omega) &:= \begin{cases} \lim_{\substack{s \uparrow t \\ s \in \mathcal{Q}}} V(s, \omega); & 0 < t \leq T \\ V(0); & t = 0 \end{cases}
 \end{aligned}$$

are well defined and finite for every $\omega \in \Omega^*$, $P(\Omega^*) = 1$, and the resulting processes are adapted. Furthermore (loc. cit.), $\{V(t+) e^{-\int_0^t \delta(v(s))ds}, \mathcal{F}_t; 0 \leq t \leq T\}$

is an RCLL, $\mathbf{P}^v$ -supermartingale, for all $v \in \mathcal{D}$; in particular,

$$V(t+) \geq E^v[V(T)e^{-\int_t^T \delta(v(s))ds} | \mathcal{F}_t], \text{ a.s.}$$

holds for every $v \in \mathcal{D}$, whence $V(t+) \geq V(t)$ a.s. On the other hand, from Fatou's lemma we have for any $v \in \mathcal{D}$:

$$\begin{aligned} V(t+) &= E^v \left[\lim_{n \rightarrow \infty} V\left(t + \frac{1}{n}\right) e^{-\int_t^{t+1/n} \delta(v(u))du} | \mathcal{F}_t \right] \\ &\leq \lim_{n \rightarrow \infty} E^v \left[V\left(t + \frac{1}{n}\right) e^{-\int_t^{t+1/n} \delta(v(u))du} | \mathcal{F}_t \right] \leq V(t), \text{ a.s.} \end{aligned}$$

and thus $\{V(t+), \mathcal{F}_t; 0 \leq t \leq T\}, \{V(t), \mathcal{F}_t; 0 \leq t \leq T\}$ are modifications of one another.

The remaining claims are immediate. $\square$

Theorem 5.3 *For an arbitrary contingent claim B , we have $h(0) = V(0)$. Furthermore, if $V(0) < \infty$, there exists a pair $(\hat{\pi}, \hat{c}) \in \mathcal{A}'(V(0))$ such that $X^{V(0), \hat{\pi}, \hat{c}}(T) = B$, a.s.*

Proof Proposition 4.2 implies $x \geq E^v[\gamma_v(T)B]$ for every $v \in \mathcal{D}$, hence $h(0) \geq V(0)$.

We now show the more difficult part: $h(0) \leq V(0)$. Clearly, we may assume $V(0) < \infty$. From (5.3), the martingale representation theorem and the Doob–Meyer decomposition, we have for every $v \in \mathcal{D}$:

$$Q_v(t) = V(0) + \int_0^t \psi'_v(s) dW_v(s) - A_v(t), \quad 0 \leq t \leq T, \quad (5.5)$$

where $\psi_v(\cdot)$ is an $\mathbb{R}^d$ -valued, $\{\mathcal{F}_t\}$ -progressively measurable and a.s. square-integrable process and $A_v(\cdot)$ is adapted with increasing, RCLL paths and $A_v(0) = 0$, $EA_v(T) < \infty$ a.s. The idea then is to consider the positive, adapted, RCLL process

$$\hat{X}(t) := \frac{V(t)}{\gamma_0(t)} = \frac{Q_v(t)}{\gamma_v(t)}, \quad 0 \leq t \leq T \quad (\forall v \in \mathcal{D}) \quad (5.6)$$

with $\hat{X}(0) = V(0)$, $\hat{X}(T) = B$ a.s., and to find a pair $(\hat{\pi}, \hat{c}) \in \mathcal{A}'(V(0))$ such that $\hat{X}(\cdot) = X^{V(0), \hat{\pi}, \hat{c}}(\cdot)$. This will prove that $h(0) \leq V(0)$.

In order to do this, let us observe that for any $\mu \in \mathcal{D}$, $v \in \mathcal{D}$ we have from (5.3)

$$Q_\mu(t) = Q_v(t) \exp \left[\int_0^t \{\delta(v(s)) - \delta(\mu(s))\} ds \right],$$

and from (5.5):

$$\begin{aligned}
 dQ_\mu(t) &= \exp\left[\int_0^t \{\delta(v(s)) - \delta(\mu(s))\}ds\right] \cdot [Q_v(t)\{\delta(v(t)) - \delta(\mu(t))\}dt \\
 &\quad + \psi'_v(t)dW_v(t) - dA_v(t)] \\
 &= \exp\left[\int_0^t \{\delta(v(s)) - \delta(\mu(s))\}ds\right] \cdot [\hat{X}(t)\gamma_v(t)\{\delta(v(t)) - \delta(\mu(t))\}dt \\
 &\quad - dA_v(t) + \psi'_v(t)\sigma^{-1}(t)(v(t) - \mu(t))dt + \psi'_v(t)dW_\mu(t)]. \quad (5.7)
 \end{aligned}$$

Comparing this decomposition with

$$dQ_\mu(t) = \psi'_\mu(t)dW_\mu(t) - dA_\mu(t), \quad (5.8)$$

we conclude that

$$\psi'_v(t) e^{\int_0^t \delta(v(s))ds} = \psi'_\mu(t) e^{\int_0^t \delta(\mu(s))ds}$$

and hence that this expression is independent of $v \in \mathcal{D}$:

$$\psi'_v(t) e^{\int_0^t \delta(v(s))ds} = \hat{X}(t)\gamma_0(t)\hat{\pi}'(t)\sigma(t); \quad \forall 0 \leq t \leq T, v \in \mathcal{D} \quad (5.9)$$

for some adapted, $\mathbb{R}^d$ -valued, a.s. square-integrable process $\hat{\pi}$ (we do not know yet that $\hat{\pi}$ takes values in K). If $X(t) = 0$, then $X(s) = 0$ for all $s \geq t$, and we can set, for example, $\pi(s) = 0$, $s \geq t$ (in fact, one can show that $\int_0^T \mathbf{1}_{\{\hat{X}(t)=0\}} \|\psi_v(t)\|^2 dt = 0$, a.s.; see Karatzas and Kou (1996)).

Similarly, we conclude from (5.7), (5.9) and (5.8):

$$\begin{aligned}
 e^{\int_0^t \delta(v(s))ds} dA_v(t) - \gamma_0(t)\hat{X}(t)[\delta(v(t)) + \hat{\pi}'(t)v(t)]dt \\
 = e^{\int_0^t \delta(\mu(s))ds} dA_\mu(t) - \gamma_0(t)\hat{X}(t)[\delta(\mu(t)) + \hat{\pi}'(t)\mu(t)]dt
 \end{aligned}$$

and hence this expression is also independent of $v \in \mathcal{D}$:

$$\hat{c}(t) := \int_0^t \gamma_v^{-1}(s)dA_v(s) - \int_0^t \hat{X}(s)[\delta(v(s)) + v'(s)\hat{\pi}(s)]ds, \quad (5.10)$$

for every $0 \leq t \leq T$, $v \in \mathcal{D}$. Setting $v \equiv 0$, we obtain $\hat{c}(t) = \int_0^t \gamma_0^{-1}(s)dA_0(s)$, $0 \leq t \leq T$ and hence

$$\left\{ \begin{array}{l} \hat{c}(\cdot) \text{ is an increasing, adapted, RCLL process} \\ \text{with } \hat{c}(0) = 0 \text{ and } \hat{c}(T) < \infty, \text{ a.s.} \end{array} \right\}. \quad (5.11)$$

Next, we claim that

$$\delta(v) + v'\hat{\pi}(t, \omega) \geq 0, \quad \ell \otimes \mathbf{P}\text{-a.e.} \quad (5.12)$$

holds for every $v \in \tilde{K}$. Then Theorem 13.1 of Rockafellar (1970) (together with continuity of $\delta(\cdot)$ and closedness of K) leads to the fact that

$$\hat{\pi}(t, \omega) \in K \quad \text{holds} \quad \ell \otimes \mathbf{P}\text{-a.e. on } [0, T] \times \Omega.$$

In order to verify (5.12), notice that from (5.10) we obtain

$$\int_0^t \gamma_v^{-1}(s) A_v(s) ds = \hat{c}(t) + \int_0^t \hat{X}(s) \{\delta(v_s) + v'_s \hat{\pi}_s\} ds; \quad 0 \leq t \leq T, \quad v \in \mathcal{D}.$$

Fix $v \in \tilde{K}$ and define the set $F_v := \{(t, \omega) \in [0, T] \times \Omega; \delta(v) + v' \hat{\pi}(t, \omega) < 0\}$. Let $\mu(t) := [v 1_{F_v^c} + n v 1_{F_v}]$, $n \in \mathbb{N}$; then $\mu \in \mathcal{D}$, and assuming that (5.12) does not hold, we get for n large enough

$$\begin{aligned} E \left[\int_0^T \gamma_\mu^{-1}(s) A_\mu(s) ds \right] &= E \left[\hat{c}(T) + \int_0^T \hat{X}(t) 1_{F_v^c} \{\delta(v) + v' \hat{\pi}(t)\} dt \right] \\ &\quad + n E \left[\int_0^T \hat{X}(t) 1_{F_v} \{\delta(v) + v' \hat{\pi}(t)\} dt \right] < 0, \end{aligned}$$

a contradiction.

Now we can put together (5.5)–(5.10) to deduce

$$\begin{aligned} d(\gamma_v(t) \hat{X}(t)) &= dQ_v(t) = \psi'_v(t) dW_v(t) - dA_v(t) \\ &= \gamma_v(t) [-d\hat{c}(t) - \hat{X}(t) \{\delta(v(t)) + v'(t) \hat{\pi}(t)\} dt \\ &\quad + \hat{X}(t) \hat{\pi}'(t) \sigma(t) dW_v(t)], \end{aligned} \tag{5.13}$$

for any given $v \in \mathcal{D}$. As a consequence, the process

$$\begin{aligned} \hat{M}_v(t) &:= \gamma_v(t) \hat{X}(t) + \int_0^t \gamma_v(s) d\hat{c}(s) + \int_0^t \gamma_v(s) \hat{X}(s) [\delta(v(s)) + v'(s) \hat{\pi}(s)] ds \\ &= V(0) + \int_0^t \gamma_v(s) \hat{X}(s) \hat{\pi}'(s) \sigma(s) dW_v(s), \quad 0 \leq t \leq T \end{aligned} \tag{5.14}$$

is a nonnegative, $\mathbf{P}^v$ -local martingale, hence supermartingale. In particular, for $v \equiv 0$, (5.13) gives:

$$\begin{aligned} d(\gamma_0(t) \hat{X}(t)) &= -\gamma_0(t) d\hat{c}(t) + \gamma_0(t) \hat{X}(t) \hat{\pi}'(t) \sigma(t) dW_0(t), \\ \hat{X}(0) &= V(0), \quad \hat{X}(T) = B, \end{aligned}$$

which is equation (2.9) for the process $\hat{X}(\cdot)$ of (5.6). This shows $\hat{X}(\cdot) \equiv X^{V(0), \hat{\pi}, \hat{c}}(\cdot)$, and hence $h(0) \leq V(0) < \infty$. $\square$

Definition 5.4 We say that claim B is K -hedgeable if its minimal cost of super-replication is finite, $V(0) < \infty$; we say it is K -attainable if there exists a portfolio

process π with values in K such that $(\pi, 0) \in \mathcal{A}'(V(0))$ and $X^{V(0), \pi, 0}(T) = B$, a.s.

Theorem 5.5 *For a given K -hedgeable contingent claim B , and any given $\lambda \in \mathcal{D}$, the conditions*

$$\{Q_\lambda(t) = V(t)e^{-\int_0^t \delta(\lambda(u))du}, \mathcal{F}_t; 0 \leq t \leq T\} \text{ is a } \mathbf{P}^\lambda\text{-martingale} \quad (5.15)$$

$$\lambda \text{ achieves the supremum in } V(0) = \sup_{v \in \mathcal{D}} E^v[B\gamma_v(T)] \quad (5.16)$$

$$\left\{ \begin{array}{l} B \text{ is } K\text{-attainable (by a portfolio } \pi), \text{ and the} \\ \text{corresponding } \gamma_\lambda(\cdot)X^{V(0), \pi, 0}(\cdot) \text{ is a } \mathbf{P}^\lambda\text{-martingale} \end{array} \right\} \quad (5.17)$$

are equivalent, and imply

$$\hat{c}(t, \omega) = 0, \quad \delta(\lambda(t, \omega)) + \lambda'(t, \omega)\hat{\pi}(t, \omega) = 0; \quad \ell \otimes P\text{-a.e.} \quad (5.18)$$

for the pair $(\hat{\pi}, \hat{c}) \in \mathcal{A}'(V(0))$ of Theorem 5.3.

Proof The $\mathbf{P}^\lambda$ -supermartingale $Q_\lambda(\cdot)$ is a $\mathbf{P}^\lambda$ -martingale, if and only if $Q_\lambda(0) = E^\lambda Q_\lambda(T) \Leftrightarrow V(0) = E^\lambda[B\gamma_\lambda(T)] \Leftrightarrow (5.16)$.

On the other hand, (5.15) implies $A_\lambda(\cdot) \equiv 0$, and so from (5.10): $\hat{c}(t) = -\int_0^t \hat{X}(s)[\delta(\lambda(s)) + \lambda'(s)\hat{\pi}(s)]ds$. Now (5.18) follows from the increase of $\hat{c}(\cdot)$ and the nonnegativity of $\delta(\lambda) + \lambda'\hat{\pi}$, since $\hat{\pi}$ takes values in K .

From (5.16) (and its consequences (5.15), (5.18)), the process $\hat{X}(\cdot)$ of (5.6) and (5.13) coincides with $X^{V(0), \hat{\pi}, 0}(\cdot)$, and we have: $\hat{X}(T) = B$ almost surely, $\gamma_\lambda(\cdot)\hat{X}(\cdot)$ is a $\mathbf{P}^\lambda$ -martingale; thus (5.17) is satisfied with $\pi \equiv \hat{\pi}$. On the other hand, suppose that (5.17) holds; then $V(0) = E^\lambda[B\gamma_\lambda(T)]$, so (5.16) holds. $\square$

Theorem 5.6 *Let B be a K -hedgeable contingent claim. Suppose that, for any $v \in \mathcal{D}$ with $\delta(v) + v'\hat{\pi} \equiv 0$,*

$$Q_v(\cdot) \text{ in (5.3) is of class } DL[0, T], \text{ under } \mathbf{P}^v. \quad (5.19)$$

Then, for any given $\lambda \in \mathcal{D}$, the conditions (5.15), (5.16), (5.18) are equivalent, and imply

$$\left\{ \begin{array}{l} B \text{ is } K\text{-attainable (by a portfolio } \pi), \text{ and the} \\ \text{corresponding } \gamma_0(\cdot)X^{V(0), \pi, 0}(\cdot) \text{ is a } \mathbf{P}^0\text{-martingale} \end{array} \right\}. \quad (5.20)$$

Proof We have already shown the implications $(5.15) \Leftrightarrow (5.16) \Rightarrow (5.18)$. To prove that these three conditions are actually equivalent under (5.19), suppose that (5.18) holds; then from (5.10): $A_\lambda(\cdot) \equiv 0$, whence the $\mathbf{P}^\lambda$ -local martingale $Q_\lambda(\cdot)$

is actually a $\mathbf{P}^\lambda$ -martingale (from (5.5) and the assumption (5.19)); thus (5.15) is satisfied.

Clearly then, if (5.15), (5.16), (5.18) are satisfied for some $\lambda \in \mathcal{D}$, they are satisfied for $\lambda \equiv 0$ as well; and from Theorem 5.5, we know then that (5.20) (i.e., (5.17) with $\lambda \equiv 0$) holds. $\square$

Remark 5.7

- (i) Loosely speaking, Theorems 5.5, 5.6 say that *the supremum in (5.16) is attained if and only if it is attained by $\lambda \equiv 0$, if and only if the Black–Scholes (unconstrained) portfolio happens to satisfy constraints.*
- (ii) It can be shown that the conditions $V(0) < \infty$ and (5.19) are satisfied (the latter, in fact, for every $v \in \mathcal{D}$) in the case of the simple European call option $B = (S_1(T) - k)^+$, provided

$$\text{the function } x \mapsto \delta(x) + x_1 \text{ is bounded from below on } \tilde{K}. \quad (5.21)$$

The same is true for any contingent claim B that satisfies $B \leq \alpha S_1(T)$ a.s., for some $\alpha \in (0, \infty)$. Note that the condition (5.21) is indeed satisfied, if the convex set K contains both the origin and the point $(1, 0, \dots, 0)$ (and thus also the line-segment adjoining these points); for then $x_1 + \delta(x) \geq x_1 + \sup_{0 \leq \alpha \leq 1} (-\alpha x_1) = x_1^+ \geq 0$, $\forall x \in \tilde{K}$.

We would like now to have a method for calculating the price $h(0)$. In order to do that, we assume constant market coefficients r, b, σ and consider only the claims of the form $B = b(S(T))$, for a given, lower-semicontinuous function b . Similarly as in the no-constraints case, the minimal hedging process will be given as $X(t) = V(t, S(t))$, for some function $V(t, s)$, depending on the constraints. Introduce also, for a given process $v(\cdot)$ in $\mathbb{R}^d$, the auxiliary, shadow economy vector of stock prices $S^v(\cdot)$ by

$$dS_i^v(t) = S_i^v(t) \left[r dt + \sum_{j=1}^d \sigma_{ij} dW_v^{(j)}(t) \right]$$

and notice that its distribution under measure P_v is the same as the one of $S(\cdot)$ under P_0 . From Theorem 5.3 we know that

$$V(t, s) = \sup_{v \in \mathcal{D}} E^v \left[b(S(T)) e^{-\int_t^T (r + \delta(v(s))) ds} \mid S(t) = s \right]. \quad (5.22)$$

We will show that this complex looking stochastic control problem has a simple solution. First, we modify the value of the claim by considering the following

function:

$$\hat{b}(s) = \sup_{v \in \tilde{K}} b(se^{-v})e^{-\delta(v)}.$$

Here, $se^{-v} = (s_1e^{-v_1}, \dots, s_de^{-v_d})'$, and we use the same notation for the componentwise product of two vectors throughout.

Theorem 5.8 *The minimal K -hedging price function $V(t, s)$ of the claim $b(S(T))$ is the Black–Scholes cost function for replicating $\hat{b}(S(T))$. In particular, under technical assumptions, it is the solution to the PDE*

$$V_t + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d a_{ij} s_i s_j V_{s_i s_j} + r \left(\sum_{i=1}^d s_i V_{s_i} - V \right) = 0, \quad (5.23)$$

with the terminal condition

$$V(T, s) = \hat{b}(s), \quad s \in \mathbb{R}_+^d, \quad (5.24)$$

and the corresponding hedging strategy π satisfies the constraints. Under technical assumptions, it is given by

$$\pi_i(t) = s_i(t) V_{s_i}(t, s(t)) / V(t, s(t)), \quad i = 1, \dots, d. \quad (5.25)$$

Proof (a) We first show that hedging $b(S(T))$ under constraints is no more expensive than hedging $\hat{b}(S(T))$ without constraints. Let $v \in \mathcal{D}$ and observe that, from the properties of the support function and the cone property of $\tilde{K}$,

$$(i) \quad \hat{\hat{b}} = \hat{b}$$

$$(ii) \quad \int_t^T \delta(v_s) ds \geq \delta \left(\int_t^T v_s ds \right),$$

$$(iii) \quad \int_t^T v_s ds \text{ is an element of } \tilde{K},$$

where $\int_t^T v(s) ds := (\int_t^T v_1(s) ds, \dots, \int_t^T v_d(s) ds)'$. Moreover, we have

$$(iv) \quad S_i^v(t) = S_i(t) e^{\int_0^t v_i(s) ds},$$

because the processes on the left-hand side and the right-hand side satisfy the same linear SDE. Then, for every $v \in \mathcal{D}$ we have

$$\begin{aligned} E^v[\hat{b}(S(T))e^{-\int_0^T (r+\delta(v(s)))ds}] &\leq E^v[\hat{b}(S^v(T))e^{-\int_0^T v(s)ds}e^{-\delta(\int_0^T v(s)ds)}e^{-rT}] \\ &\leq E^v[\sup_{v \in \tilde{K}} \hat{b}(S^v(T))e^{-\delta(v)}e^{-rT}] \\ &= E^v[\hat{\hat{b}}(S^v(T))e^{-rT}] = E^0[\hat{b}(S(T))e^{-rT}]. \end{aligned} \quad (5.26)$$

Similarly for conditional expectations of (5.22), hence $V(t, s)$ is no larger than the Black–Scholes price process of the claim $\hat{b}(S(T))$.

(b) To conclude we have to show that to superreplicate $b(S(T))$ we have to hedge at least $\hat{b}(S(T))$. It is sufficient to prove that the left limit of $V(t, s)$ at $t = T$ is larger than $\hat{b}(s)$. For this, let $\{v^k\}$ be the maximizing sequence in the cone $\tilde{K}$ attaining $\hat{b}(s)$, i.e., such that $b(se^{-v^k})e^{-\delta(v^k)}$ converges to $\hat{b}(s)$ as k goes to infinity. Then, using (for fixed $t < T$) constant deterministic controls $v_k(t) = v^k/(T - t)$ in (5.22), we get

$$V(t, s) \geq E^0[b(S(T)e^{-v^k})e^{-\delta(v^k)}e^{-r(T-t)} \mid S(t) = s],$$

hence

$$\lim_{t \rightarrow T} V(t, s) \geq b(se^{-v^k})e^{-\delta(v^k)}$$

and letting k go to infinity, we finish the proof. Here is a sketch of a PDE proof for part (a) in the proof above. Let V be the solution to (5.23), (5.24). For a given $v \in \tilde{K}$, consider the function $W_v = (sV_s)'v + \delta(v)V$, where V_s is the vector of partial derivatives of V with respect to s_i , $i = 1, \dots, d$. By Theorem 13.1 in Rockafellar (1970), to prove that portfolio π of (5.25) takes values in K , it is sufficient to prove that W_v is nonnegative, for all $v \in \tilde{K}$. It is not difficult to see (assuming enough smoothness) that W_v solves PDE (5.23), too. Moreover, it is also straightforward to check that $W_v(s, T) \geq 0$. So, by the maximum principle, $W_v \geq 0$ everywhere. $\square$

Example 5.9 We restrict ourselves to the case of only one stock, $d = 1$, and to constraints of the type

$$K = [-l, u], \quad (5.27)$$

with $0 \leq l, u \leq +\infty$, with the understanding that the interval K is open to the right (left) if $u = +\infty$ (respectively, if $l = +\infty$). It is straightforward to see that

$$\delta(v) = lv^+ + uv^-,$$

and $\tilde{K} = \mathbb{R}$ if both l and u are finite. In general,

$$\tilde{K} = \{x \in \mathbb{R} : x \geq 0 \text{ if } u = +\infty, x \leq 0 \text{ if } l = +\infty\}.$$

For the European call $b(s) = (s - k)^+$, one easily gets that $\hat{b}(s) \equiv \infty$, if $u < 1$, $\hat{b}(s) = s$ if $u = 1$ (no-borrowing) and $\hat{b}(s) = b(s)$ if $u = \infty$ (short-selling constraints don't matter for the call option). For $1 < u < \infty$ we have (by ordinary

calculus)

$$\hat{b}(s) = \begin{cases} s - k; & s \geq \frac{ku}{u-1} \\ \frac{k}{u-1} \left(\frac{(u-1)s}{ku} \right)^u; & s < \frac{ku}{u-1}. \end{cases}$$

For the European put $b(s) = (k - s)^+$, one gets $\hat{b} = b$ if $l = \infty$ (borrowing constraints don't matter), $\hat{b} \equiv k$ if $l = 0$ (no short-selling), and otherwise

$$\hat{b}(s) = \begin{cases} k - s; & s \leq \frac{kl}{l+1} \\ \frac{k}{l+1} \left(\frac{ku}{(l+1)s} \right)^l; & s > \frac{kl}{l+1}. \end{cases}$$

Numerical results on hedging these (and other) options under the above constraints can be found in Broadie, Cvitanić and Soner (1998).

6 The case of concave drift

In this section we consider the case of an agent whose drift is a concave function of his trading strategy. The most prominent example is the case in which the borrowing rate R is larger than the lending rate r . Moreover, it also includes examples of a “large investor” who can influence the drift of the asset prices by trading in the market (see Cuoco and Cvitanić 1998).

We assume that the wealth process $X(t)$ satisfies the stochastic differential equation

$$dX(t) = X(t)g(t, \pi_t)dt + X(t)\pi'(t)\sigma(t)dW(t) - dc(t), \quad X(0) = x > 0, \quad (6.1)$$

where the function $g(t, \cdot)$ is concave for all $t \in [0, T]$, and uniformly (with respect to t) Lipschitz:

$$|g(t, x) - g(t, y)| \leq k\|x - y\|, \quad \forall t \in [0, T]; \quad x, y \in \mathbb{R}^d,$$

for some $0 < k < \infty$. Moreover, we assume $g(\cdot, 0) \equiv 0$.

In analogy with the case of constraints we define the convex conjugate function $\tilde{g}$ of g by

$$\tilde{g}(t, v) := \sup_{\pi \in \mathbb{R}^d} \{g(t, \pi) + \pi'v\}, \quad (6.2)$$

on its *effective domain* $\mathcal{D}_t := \{v : \tilde{g}(v, t) < \infty\}$. Introduce also the class $\mathcal{D}$ of processes $v(t)$ taking values in $\mathcal{D}_t$, for all t . It is clear that under the above assumptions $\mathcal{D}$ is not empty. We also assume, for simplicity, that the function $\tilde{g}(t, \cdot)$ is bounded on its effective domain, uniformly in t .

For a given $\{\mathcal{F}_t\}$ -progressively measurable process $v(\cdot)$ with values in $\mathbb{R}^d$ we introduce

$$\gamma_v(t, u) := \exp \left\{ - \int_t^u \tilde{g}(s, v_s) ds \right\}, \quad \gamma_v(t) := \gamma_v(0, t),$$

$$dZ_v(t) := -\sigma^{-1}(t)v(t)Z_v(t)dW(t), \quad Z_v(0) = 1, \quad H_v(t) := Z_v(t)\gamma_v(t). \quad (6.3)$$

For every $v \in \mathcal{D}$ we have (by Itô's rule)

$$\begin{aligned} H_v(t)X(t) &+ \int_0^t H_v(s) [X(s)(\tilde{g}(s, v_s) - g(s, \pi_s) - \pi'(s)v(s))ds + dc(s)] \\ &= x + \int_0^t H_v(s)X(s) [\pi'(s)\sigma(s) + \sigma^{-1}(s)v(s)] dW(s). \end{aligned} \quad (6.4)$$

In particular, the process on the right-hand side is a nonnegative local martingale, hence a supermartingale. Therefore we get the following *necessary condition for π to be admissible*:

$$\sup_{v \in \mathcal{D}} E \left[H_v(T)X(T) + \int_0^T H_v(s)X(s) \{ \tilde{g}(s, v_s) - g(s, \pi_s) - \pi'(s)v(s) \} ds \right] \leq x. \quad (6.5)$$

The supermartingale property excludes arbitrage opportunities from this market: if $x = 0$, then necessarily $X(t) = 0, \forall 0 \leq t \leq T$, almost surely.

Next, for a given $v \in \mathcal{D}$, introduce the process

$$W_v(t) := W(t) - \int_0^t \sigma^{-1}(s)v(s)ds,$$

as well as the measure

$$\mathbf{P}_v(A) := E[Z_v(T)1_A] = E^v[1_A], \quad A \in \mathcal{F}_T.$$

It can be shown under our assumptions that the sets $\mathcal{D}_t$ are uniformly bounded. Therefore, if $v \in \mathcal{D}$, then $Z_v(\cdot)$ is a martingale. Thus, for every $v \in \mathcal{D}$, the measure $\mathbf{P}_v$ is a probability measure and the process $W_v(\cdot)$ is a $\mathbf{P}_v$ -Brownian motion, by Girsanov's theorem.

Given a contingent claim B , consider, for every stopping time τ , the $\mathcal{F}_\tau$ -measurable random variable

$$V(\tau) := \operatorname{ess\,sup}_{v \in \mathcal{D}} E^v[B\gamma_v(\tau, T)|\mathcal{F}_\tau].$$

The proof of the following theorem is similar to the corresponding theorem in the case of constraints.

Theorem 6.1 *For an arbitrary contingent claim B , we have $h(0) = V(0)$. Furthermore, there exists a pair $(\hat{\pi}, \hat{c}) \in \mathcal{A}_0(V(0))$ such that $X^{V(0), \hat{\pi}, \hat{c}}(\cdot) = V(\cdot)$.*

The theorem gives the minimal hedging price for a claim B ; in fact, it is easy to see (using the same supermartingale argument as before) that the process $V(\cdot)$ is the minimal wealth process that hedges B . There remains the question of whether consumption is necessary. We show that, in fact, $\hat{c}(\cdot) \equiv 0$.

Theorem 6.2 *Every contingent claim B is attainable, that is, the process $\hat{c}(\cdot)$ from Theorem 6.1 is a zero-process.*

Proof Let $\{\nu_n; n \in \mathbb{N}\}$ be a maximizing sequence for achieving $V(0)$, i.e., $\lim_{n \rightarrow \infty} E^{\nu_n}[B\gamma_{\nu_n}(T)] = V(0)$. Similarly to (6.5), one can get

$$\sup_{\nu \in \mathcal{D}} E^\nu \left[\gamma_\nu(T)V(T) + \int_0^T \gamma_\nu(t) d\hat{c}(t) \right] \leq V(0).$$

Since $V(T) = B$, this implies $\lim_{n \rightarrow \infty} E^{\nu_n} \int_0^T \gamma_{\nu_n}(t) d\hat{c}(t) = 0$ and, since the processes $\gamma_{\nu_n}(\cdot)$ are bounded away from zero (uniformly in n), $\lim_{n \rightarrow \infty} E[Z_{\nu_n}(T)\hat{c}(T)] = 0$. Using weak compactness arguments as in Cvitanić and Karatzas (1993, Theorem 9.1) we can show that there exists $\nu \in \mathcal{D}$ such that $\lim_{n \rightarrow \infty} E[Z_{\nu_n}\hat{c}(T)] = E[Z_\nu(T)\hat{c}(T)] = 0$ (along a subsequence). It follows that $\hat{c}(\cdot) \equiv 0$. $\square$

The theorems above also follow from the general theory of Backward Stochastic Differential Equations, as presented in El Karoui, Peng and Quenez (1997).

Example 6.3 *Different borrowing and lending rates.* We have studied so far a model in which one is allowed to borrow money, at an interest rate $R(\cdot)$ equal to the bank rate $r(\cdot)$. In this section we consider the more general case of a financial market $\mathcal{M}^*$ in which $R(\cdot) \geq r(\cdot)$, without constraints on portfolio choice. We assume that the progressively measurable process $R(\cdot)$ is also bounded.

In this market $\mathcal{M}^*$ it is not reasonable to borrow money and to invest money in the bank at the same time. Therefore, we restrict ourselves to policies for which the relative amount borrowed at time t is equal to $\left(1 - \sum_{i=1}^d \pi_i(t)\right)^-$. Then, the wealth process $X = X^{x, \pi, c}$ corresponding to initial capital $x > 0$ and portfolio/consumption pair (π, c) satisfies

$$\begin{aligned} dX(t) &= r(t)X(t)dt - dc(t) \\ &+ X(t) \left[\pi'(t)\sigma(t)dW_0(t) - (R(t) - r(t)) \left(1 - \sum_{i=1}^d \pi_i(t)\right)^- dt \right]. \end{aligned}$$

We get $\tilde{g}(v(t)) = r(t) - v_1(t)$ for $v \in \mathcal{D}$, where

$$\mathcal{D} := \{v; \ v \text{ a progressively measurable, } \mathbb{R}^d\text{-valued process with}$$

$$r - R \leq v_1 = \dots = v_d \leq 0, \ \ell \otimes \mathbf{P}\text{-a.e.}\}.$$

We also have

$$\begin{aligned} \tilde{g}(v(t)) - g(t, \pi(t)) - \pi'(t)v(t) &= [R(t) - r(t) + v_1(t)] \left(1 - \sum_{i=1}^d \pi_i(t)\right)^- \\ &\quad - v_1(t) \left(1 - \sum_{i=1}^d \pi_i(t)\right)^+, \end{aligned}$$

for $0 \leq t \leq T$. It can be shown, in analogy to the case of constraints, that the optimal dual process $\hat{\lambda}(\cdot) \in \mathcal{D}$ can be taken as the one that attains zero in this equation, namely

$$\hat{\lambda}(t) = \hat{\lambda}_1(t)\mathbf{1}, \quad \hat{\lambda}_1(t) := [r(t) - R(t)] \mathbf{1}_{\{\sum_{i=1}^d \hat{\pi}_i(t) > 1\}}.$$

Assume now constant coefficients, and observe that the stock price processes vector satisfies the equations

$$\begin{aligned} dS_i(t) &= S_i(t)[b_i(t)dt + \sum_{j=1}^d \sigma_{ij}dW^j(t)] \\ &= S_i(t)[(r - v_1(t))dt + \sum_{j=1}^d \sigma_{ij}dW_v^j(t)], \quad 1 \leq i \leq d, \end{aligned}$$

for every $v \in \mathcal{D}$. Consider now a contingent claim of the form $B = \varphi(S(T))$, for a given continuous function $\varphi : \mathbb{R}_+^d \rightarrow [0, \infty)$ that satisfies a polynomial growth condition, as well as the *value function*

$$Q(t, s) := \sup_{v \in \mathcal{D}} E^v[\varphi(S(T))e^{-\int_t^T (r - v_1(s))ds} | S(t) = s]$$

on $[0, T] \times \mathbb{R}_+^d$. Clearly, the processes $\hat{X}$, V are given as

$$\hat{X}(t) = Q(t, S(t)), \quad V(t) = e^{-rt} \hat{X}(t); \quad 0 \leq t \leq T,$$

where Q solves the semilinear parabolic partial differential equation of Hamilton–Jacobi–Bellman (HJB) type,

$$\frac{\partial Q}{\partial t} + \frac{1}{2} \sum_i \sum_j a_{ij} s_i s_j \frac{\partial^2 Q}{\partial s_i \partial s_j} + \max_{r - R \leq v_1 \leq 0} \left[(r - v_1) \left\{ \sum_i s_i \frac{\partial Q}{\partial s_i} - Q \right\} \right] = 0,$$

for $0 \leq t < T$, $s \in \mathbb{R}_+^d$,

$$Q(T, s) = \varphi(s); \quad s \in \mathbb{R}_+^d$$

(see Ladyženskaja, Solonnikov and Ural'tseva (1968) for the basic theory of such equations, and Fleming and Rishel (1975), Fleming and Soner (1993) for the connections with stochastic control). The maximization in the HJB equation is achieved by $v_1^* = (r - R)1_{\{\sum_i s_i \frac{\partial Q}{\partial s_i} \geq Q\}}$; the portfolio $\hat{\pi}(\cdot)$ and the process $\hat{\lambda}_1(\cdot)$ are then given, respectively, by

$$\hat{\pi}_i(t) = \frac{S_i(t) \cdot \frac{\partial}{\partial p_i} Q(t, S(t))}{Q(t, S(t))}, \quad i = 1, \dots, d$$

and

$$\hat{\lambda}_1(t) = (r - R)1_{\{\sum_i \hat{\pi}_i(t) \geq 1\}}.$$

The HJB PDE becomes

$$\frac{\partial Q}{\partial t} + \frac{1}{2} \sum_i \sum_j s_i s_j a_{ij} \frac{\partial^2 Q}{\partial s_i \partial s_j} + R \left(\sum_i s_i \frac{\partial Q}{\partial s_i} - Q \right)^+ - r \left(\sum_i s_i \frac{\partial Q}{\partial s_i} - Q \right)^- = 0.$$

Suppose now that the function φ satisfies $\sum_i s_i \frac{\partial \varphi(s)}{\partial s_i} \geq \varphi(s)$, $\forall s \in \mathbb{R}_+^d$. Then the solution Q also satisfies this inequality:

$$\sum_i s_i \frac{\partial Q(t, s)}{\partial s_i} \geq Q(t, s), \quad 0 \leq t \leq T$$

for all $s \in \mathbb{R}_+^d$ and is given as the solution to the Black–Scholes equation with r replaced with R :

$$\begin{aligned} \frac{\partial Q}{\partial t} + \frac{1}{2} \sum_i \sum_j s_i s_j a_{ij} \frac{\partial^2 Q}{\partial s_i \partial s_j} + R \left(\sum_i s_i \frac{\partial Q}{\partial s_i} - Q \right) &= 0; \quad t < T, s > 0 \\ Q(T, s) &= \varphi(s); \quad s > 0 \end{aligned}$$

In this case the seller's hedging portfolio $\hat{\pi}(\cdot)$ always borrows: $\sum_{i=1}^d \hat{\pi}_i(t) \geq 1$, $0 \leq t \leq T$, and it was to be expected that all he has to do is use R as the interest rate. Note, however, that this price may be too high for the buyer of the option.

7 Utility functions

A function $U : (0, \infty) \rightarrow \mathbb{R}$ will be called a *utility function* if it is strictly increasing, strictly concave, of class C^1 , and satisfies

$$U'(0+) := \lim_{x \downarrow 0} U'(x) = \infty, \quad U'(\infty) := \lim_{x \rightarrow \infty} U'(x) = 0.$$

We shall denote by I the (continuous, strictly decreasing) inverse of the function U' ; this function maps $(0, \infty)$ onto itself, and satisfies $I(0+) = \infty, I(\infty) = 0$. We also introduce the Legendre–Fenchel transform

$$\tilde{U}(y) := \max_{x>0} [U(x) - xy] = U(I(y)) - yI(y), \quad 0 < y < \infty$$

of $-U(-x)$; this function $\tilde{U}$ is strictly decreasing and strictly convex, and satisfies

$$\tilde{U}'(y) = -I(y), \quad 0 < y < \infty,$$

$$U(x) = \min_{y>0} [\tilde{U}(y) + xy] = \tilde{U}(U'(x)) + xU'(x), \quad 0 < x < \infty.$$

It is now readily checked that

$$U(I(y)) \geq U(x) + y[I(y) - x],$$

$$\tilde{U}(U'(x)) + x[U'(x) - y] \leq \tilde{U}(y)$$

are valid for all $x > 0, y > 0$. It is also easy to see that

$$\tilde{U}(\infty) = U(0+), \quad \tilde{U}(0+) = U(\infty)$$

hold; see Karatzas *et al.* (1991), Lemma 4.2.

For some of the results that follow, we will need to impose the following conditions on our utility functions:

$$c \mapsto cU'(c) \text{ is nondecreasing on } (0, \infty); \quad (7.1)$$

$$\text{for some } \alpha \in (0, 1), \gamma \in (1, \infty) \text{ we have : } \alpha U'(x) \geq U'(\gamma x), \quad \forall x \in (0, \infty). \quad (7.2)$$

Condition (7.1) is equivalent to

$$y \mapsto yI(y) \text{ is nonincreasing on } (0, \infty),$$

and implies that

$$x \mapsto \tilde{U}(e^x) \text{ is convex on } \mathbb{R}.$$

(If U is of class C^2 , then condition (7.1) amounts to the statement that $-cU''(c)/U'(c)$, the so-called “Arrow–Pratt measure of relative risk–aversion”, does not exceed 1. For the general treatment under the weakest possible conditions on the utility function see Kramkov and Schachermayer 1998.)

Similarly, condition (7.2) is equivalent to having

$$I(\alpha y) \leq \gamma I(y), \quad \forall y \in (0, \infty) \quad \text{for some } \alpha \in (0, 1), \quad \gamma > 1.$$

Iterating this, we obtain the apparently stronger statement

$$\forall \alpha \in (0, 1), \exists \gamma \in (1, \infty) \text{ such that } I(\alpha y) \leq \gamma I(y), \quad \forall y \in (0, \infty).$$

8 Portfolio optimization under constraints

In this section we consider the optimization problem of maximizing utility from terminal wealth for an investor subject to the portfolio constraints given by the set K , i.e., we want to maximize

$$J(x; \pi) := EU(X^{x,\pi}(T)),$$

over the class $\mathcal{A}'_0$ of constrained portfolios π for which $(\pi, 0) \in \mathcal{A}'(x)$ that satisfy

$$EU^-(X^{x,\pi}(T)) < \infty.$$

The value function of this problem will be denoted by

$$V(x) := \sup_{\pi \in \mathcal{A}'_0(x)} J(x; \pi), \quad x \in (0, \infty). \quad (8.1)$$

We assume that $V(x) < \infty$, $\forall x \in (0, \infty)$. It is fairly straightforward that the function $V(\cdot)$ is increasing and concave on $(0, \infty)$ and that this assumption is satisfied if the function U is nonnegative and satisfies the growth condition

$$0 \leq U(x) \leq \kappa(1 + x^\alpha); \quad \forall x \in (0, \infty) \quad (8.2)$$

for some constants $\kappa \in (0, \infty)$ and $\alpha \in (0, 1)$ – see Karatzas *et al.* (1991) for details.

Recall the notation

$$H_v(t) = \gamma_v(t)Z_v(t)$$

of (4.8). We introduce the function

$$\mathcal{X}_v(y) := E[H_v(T)I(yH_v(T))], \quad 0 < y < \infty,$$

and the class $\mathcal{H}$ of $\tilde{K}$ -valued, progressively measurable processes $v(\cdot)$ such that $E \int_0^T \|v(t)\|^2 dt + E \int_0^T \delta(v(t)) dt < \infty$. Consider the subclass $\mathcal{D}'$ of $\mathcal{H}$ given by

$$\mathcal{D}' := \{v \in \mathcal{H}; \mathcal{X}_v(y) < \infty, \quad \forall y \in (0, \infty)\}.$$

For every $v \in \mathcal{D}'$, the function $\mathcal{X}_v(\cdot)$ is continuous and strictly decreasing, with $\mathcal{X}_v(0+) = \infty$ and $\mathcal{X}_v(\infty) = 0$; we denote its inverse by $\mathcal{Y}_v(\cdot)$.

Next, we prove a crucial lemma, which provides sufficient conditions for optimality in the problem of (8.1). The duality approach of the lemma and subsequent analysis was implicitly used in Pliska (1986), Karatzas, Lehoczky and Shreve (1987), and Cox and Huang (1989) in the case of no constraints, and explicitly in He and Pearson (1991), Karatzas *et al.* (1991), Xu and Shreve (1992), and Cvitanić and Karatzas (1993) for various types of constraints.

Lemma 8.1 For any given $x > 0$, $y > 0$ and $\pi \in \mathcal{A}'(x)$, we have

$$EU(X^{x,\pi}(T)) \leq E\tilde{U}(yH_v(T)) + yx, \quad \forall v \in \mathcal{H}. \quad (8.3)$$

In particular, if $\hat{\pi} \in \mathcal{A}'(x)$ is such that equality holds in (8.3), for some $\lambda \in \mathcal{H}$ and $\hat{y} > 0$, then $\hat{\pi}$ is optimal for our (primal) optimization problem, while λ is optimal for the dual problem

$$\tilde{V}(\hat{y}) := \inf_{v \in \mathcal{H}} E\tilde{U}(\hat{y}H_v(T)) =: \inf_{v \in \mathcal{H}} \tilde{J}(\hat{y}; v). \quad (8.4)$$

Furthermore, equality holds in (8.3) if

$$X^{x,\pi}(T) = I(yH_v(T)) \text{ a.s.}, \quad (8.5)$$

$$\delta(v_t) = -v'(t)\pi(t) \text{ a.e.}, \quad (8.6)$$

$$E[H_v(T)X^{x,\pi}(T)] = x \quad (8.7)$$

(the latter being equivalent to $v \in \mathcal{D}'$ and $y = \mathcal{Y}_v(x)$, if (8.5) holds).

Proof By definitions of $\tilde{U}$, δ we get

$$U(X(T)) \leq \tilde{U}(yH_v(T)) + yH_v(T)X(T) + \int_0^T H_v(t)X(t)[\delta(v_t) + v'(t)\pi(t)]dt.$$

The upper bound of (8.3) follows from Proposition 4.2 (also valid for $v(\cdot) \in \mathcal{H}$); condition (8.5) follows from the definition of $\tilde{U}(\cdot)$, conditions (8.6) and (8.7) correspond to $H_v(\cdot)X(\cdot)$ being a martingale, not only a supermartingale. $\square$

Remark 8.2 Lemma 8.1 suggests the following strategy for solving the optimization problem:

- (i) show that the dual problem (8.4) has an optimal solution $\lambda_y \in \mathcal{D}'$ for all $y > 0$;
- (ii) using Theorem 5.3, find the minimal hedging price $h_y(0)$ and a corresponding portfolio $\hat{\pi}_y$ for hedging $B_{\lambda_y} := I(yH_{\lambda_y}(T))$;
- (iii) prove (8.6) for the pair $(\hat{\pi}_y, \lambda_y)$;
- (iv) show that, for every $x > 0$, you can find $\hat{y} = y_x > 0$ such that $x = h_{\hat{y}}(0) = E[H_{\lambda_{\hat{y}}}(T)I(\hat{y}H_{\lambda_{\hat{y}}}(T))]$.

Then (i)–(iv) would imply that $\hat{\pi}_{\hat{y}}$ is the optimal portfolio process for the utility maximization problem of an investor starting with initial capital equal to x .

To verify that step (i) can be accomplished, we impose the following condition:

$$\forall y \in (0, \infty), \quad \exists v \in \mathcal{H} \text{ such that } \tilde{J}(y; v) := E\tilde{U}(yH_v(T)) < \infty. \quad (8.8)$$

We also impose the assumption

$$U(0+) > -\infty, \quad U(\infty) = \infty. \quad (8.9)$$

Under the condition (8.2), the requirement (8.8) is satisfied. Indeed, we get

$$0 \leq \tilde{U}(y) \leq \tilde{\kappa}(1 + y^{-\rho}); \quad \forall y \in (0, \infty)$$

for some $\tilde{\kappa} \in (0, \infty)$ and $\rho = \alpha/(1 - \alpha)$.

Even though the log function does not satisfy (8.9), we solve that case directly in examples below.

Theorem 8.3 *Assume that (7.1), (7.2), (8.8) and (8.9) are satisfied. Then condition (i) of Remark 8.2 is true, i.e. the dual problem admits a solution in the set $\mathcal{D}'$, for every $y > 0$.*

The fact that the dual problem admits a solution under the conditions of Theorem 8.3 follows almost immediately (by standard weak compactness arguments) from Proposition 8.4 below. The details, as well as a relatively straightforward proof of Proposition 8.4, can be found in Cvitanic and Karatzas (1992). Denote by $\mathcal{H}'$ the Hilbert space of progressively measurable processes v with norm $\|v\| = E \int_0^T v^2(s) ds < \infty$.

Proposition 8.4 *Under the assumptions of Theorem 8.3, the functional $\tilde{J}(y; \cdot) : \mathcal{H}' \rightarrow \mathbb{R} \cup \{+\infty\}$ of (8.4) is (i) convex, (ii) coercive: $\lim_{\|v\| \rightarrow \infty} \tilde{J}(y; v) = \infty$, and (iii) lower-semicontinuous: for every $v \in \mathcal{H}'$ and $\{v_n\}_{n \in \mathbb{N}} \subseteq \mathcal{H}'$ with $\|v_n - v\| \rightarrow 0$ as $n \rightarrow \infty$, we have*

$$\tilde{J}(y; v) \leq \lim_{n \rightarrow \infty} \tilde{J}(y; v_n).$$

We move now to step (ii) of Remark 8.2. We have the following useful fact:

Lemma 8.5 *For every $v \in \mathcal{H}$, $0 < y < \infty$, we have*

$$E[H_v(T)B_{\lambda_y}] \leq E[H_{\lambda_y}(T)B_{\lambda_y}]. \quad (8.10)$$

In fact, (8.10) is equivalent to λ_y being optimal for the dual problem, but we do not need that result here; its proof is quite lengthy and technical (see Cvitanic and Karatzas (1992), Theorem 10.1). We are going to provide a simpler proof for Lemma 8.5, but under the additional assumption that

$$E[H_{\lambda_y}(T)I(yH_v(T))] < \infty, \quad \forall v \in \mathcal{H}, y > 0. \quad (8.11)$$

Proof of Lemma 8.5 Fix $\varepsilon \in (0, 1)$, $v \in \mathcal{H}$ and define (suppressing dependence on t)

$$G_\varepsilon := (1 - \varepsilon)H_{\lambda_y} + \varepsilon H_v, \quad \mu_\varepsilon := G_\varepsilon^{-1}((1 - \varepsilon)H_{\lambda_y}\lambda_y + \varepsilon H_v v),$$

$$\tilde{\mu}_\varepsilon := G_\varepsilon^{-1}((1 - \varepsilon)H_{\lambda_y}\delta(\lambda_y) + \varepsilon H_v\delta(v)).$$

Then $\mu_\varepsilon \in \mathcal{H}$, because of the convexity of $\tilde{K}$. Moreover, we have

$$dG_\varepsilon = (\theta + \sigma^{-1}\mu_\varepsilon)G_\varepsilon dW - \tilde{\mu}_\varepsilon G_\varepsilon dt,$$

and convexity of δ implies $\delta(\mu_\varepsilon) \leq \tilde{\mu}_\varepsilon$, and therefore, comparing the solutions to the respective (linear) SDEs, we get

$$G_\varepsilon(\cdot) \leq H_{\mu_\varepsilon}(\cdot), \text{ a.s.}$$

Since λ_y is optimal and $\tilde{U}$ is decreasing, this implies

$$\varepsilon^{-1}(E[\tilde{U}(yH_{\lambda_y}(T)) - \tilde{U}(yG_\varepsilon(T))]) \leq 0. \quad (8.12)$$

Next, recall that $I = -\tilde{U}'$ and denote by V_ε the random variable inside the expectation operator in (8.12). Fix $\omega \in \Omega$, and assume, suppressing the dependence on ω and T , that $H_v \geq H_{\lambda_y}$. Then $\varepsilon^{-1}V_\varepsilon = I(F)y(H_v - H_{\lambda_y})$, where $yH_{\lambda_y} \leq F \leq yH_{\lambda_y} + \varepsilon y(H_v - H_{\lambda_y})$. Since I is decreasing we get $\varepsilon^{-1}V_\varepsilon \geq yI(yH_v)(H_v - H_{\lambda_y})$. We get the same result when assuming $H_v \leq H_{\lambda_y}$. This and assumption (8.11) imply that we can use Fatou's lemma when taking the limit as $\varepsilon \downarrow 0$ in (8.12), which gives us (8.10). $\square$

Now, given $y > 0$ and the optimal λ_y for the dual problem, let π_y be the portfolio of Theorem 5.3 for hedging the claim $B_{\lambda_y} = I(yH_{\lambda_y}(T))$. Lemma 8.5 implies that, in the notation of Section 5,

$$h_y(0) = V_y(0) = E[H_{\lambda_y}(T)I(yH_{\lambda_y}(T))] = \text{initial capital for portfolio } \pi_y,$$

so (8.7) is satisfied for $x = h_y(0)$. It also implies, by (5.18), that (8.6) holds for the pair (π_y, λ_y) . Therefore we have completed both steps (ii) and (iii). Step (iv) is a corollary of the following result.

Proposition 8.6 *Under the assumptions of Theorem 8.3, for any given $x > 0$, there exists $\hat{y} > 0$ that achieves $\inf_{y>0}[\tilde{V}(y) + xy]$ and satisfies*

$$x = \mathcal{X}_{\lambda_{\hat{y}}}(y_x).$$

For the (straightforward) proof see Cvitanić and Karatzas (1992), Proposition 12.2. We now put together the results of this section:

Theorem 8.7 *Under the assumptions of Theorem 8.3, for any given $x > 0$ there exists an optimal portfolio process $\hat{\pi}$ for the utility maximization problem (8.1). Process $\hat{\pi}$ is equal to the portfolio of Theorem 5.3 for minimally hedging the claim $I(\hat{y}H_{\lambda_{\hat{y}}}(T))$, where $\hat{y}$ is given by Proposition 8.6 and $\lambda_{\hat{y}}$ is the optimal process for the dual problem (8.4).*

9 Examples

Example 9.1 *Logarithmic utility.* If $U(x) = \log x$, we have $I(y) = 1/y$, $\tilde{U}(y) = -(1 + \log y)$ and

$$\mathcal{X}_v(y) = \frac{1}{y}, \quad \mathcal{Y}_v(x) = \frac{1}{x},$$

and therefore the optimal terminal wealth is

$$X_\lambda(T) = x \frac{1}{H_\lambda(T)} \quad (9.1)$$

for $\lambda \in \mathcal{H}$ optimal. (In particular $\mathcal{D}' = \mathcal{H}$ in this case.) Therefore,

$$E[\tilde{U}(\mathcal{Y}_\lambda(x)H_v(T))] = -1 - \log \frac{1}{x} + E\left(\log \frac{1}{H_v(T)}\right).$$

But

$$E\left(\log \frac{1}{H_v(T)}\right) = E \int_0^T \left[r(s) + \delta(v(s)) + \frac{1}{2} \|\theta(s) + \sigma^{-1}(s)v(s)\|^2 \right] ds,$$

and thus the dual problem amounts to a point-wise minimization of the convex function $\delta(x) + \frac{1}{2} \|\theta(t) + \sigma^{-1}(t)x\|^2$ over $x \in \tilde{K}$, for every $t \in [0, T]$:

$$\lambda(t) = \arg \min_{x \in \tilde{K}} [2\delta(x) + \|\theta(t) + \sigma^{-1}(t)x\|^2].$$

Furthermore, (9.1) gives

$$H_\lambda(t)X_\lambda(t) = x; \quad 0 \leq t \leq T,$$

and using Itô's rule to get the SDE for $H_\lambda X_\lambda$ we get, by equating the integrand in the stochastic integral term to zero, $\sigma'(t)\hat{\pi}(t) = \theta_\lambda(t)$, $\ell \otimes P$ -a.e.

We conclude that the optimal portfolio is given by

$$\hat{\pi}(t) = (\sigma(t)\sigma'(t))^{-1}[\lambda(t) + b(t) - r(t)\mathbf{1}].$$

Example 9.2 (Constraints on borrowing) From the point of view of applications, an interesting example is the one in which the total proportion $\sum_{i=1}^d \pi_i(t)$ of wealth invested in stocks is bounded from above by some real constant $a > 0$. For

example, if we take $a = 1$, we exclude borrowing; with $a \in (1, 2)$, we allow borrowing up to a fraction $1 - a$ of wealth. If we take $a = 1/2$, we have to invest at least half of the wealth in the bank.

To illustrate what happens in this situation, let again $U(x) = \log x$, and, for the sake of simplicity, $d = 2$, $\sigma =$ unit matrix, and the constraints on the portfolio be given by

$$K = \{x \in \mathbb{R}^2; x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq a\}$$

for some $a \in (0, 1]$ (obviously, we also exclude short-selling with this K). We have here $\delta(x) \equiv a \max\{x_1^-, x_2^-\}$, and thus $\tilde{K} = \mathbb{R}^2$. By some elementary calculus and/or by inspection, and omitting the dependence on t , we can see that the optimal dual process λ that minimizes $\frac{1}{2}\|\theta_t + \nu_t\|^2 + \delta(\nu_t)$, and the optimal portfolio $\pi_t = \theta_t + \lambda_t$, are given respectively by

$$\lambda = -\theta; \pi = (0, 0)' \text{ if } \theta_1, \theta_2 \leq 0$$

(do not invest in stocks if the interest rate is larger than the stocks return rates),

$$\begin{aligned} \lambda &= (0, -\theta_2)'; \pi = (\theta_1, 0)' \text{ if } \theta_1 \geq 0, \theta_2 \leq 0, a \geq \theta_1, \\ \lambda &= (a - \theta_1, -\theta_2)'; \pi = (a, 0)' \text{ if } \theta_1 \geq 0, \theta_2 \leq 0, a < \theta_1, \\ \lambda &= (-\theta_1, 0)'; \pi = (0, \theta_2)' \text{ if } \theta_1 \leq 0, \theta_2 \geq 0, a \geq \theta_2, \\ \lambda &= (-\theta_1, a - \theta_2)'; \pi = (0, a)' \text{ if } \theta_1 \leq 0, \theta_2 \geq 0, a < \theta_2, \end{aligned}$$

(do not invest in the stock whose rate is less than the interest rate, invest $X \min\{a, \theta_i\}$ in the i -th stock whose rate is larger than the interest rate),

$$\lambda = (0, 0)'; \pi = \theta \text{ if } \theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 \leq a$$

(invest $\theta_i X$ in the respective stocks – as in the no constraints case – whenever the optimal portfolio of the no-constraints case happens to take values in K),

$$\begin{aligned} \lambda &= (a - \theta_1, -\theta_2)'; \pi = (a, 0)' \text{ if } \theta_1, \theta_2 \geq 0, a \leq \theta_1 - \theta_2, \\ \lambda &= (-\theta_1, a - \theta_2)'; \pi = (0, a)' \text{ if } \theta_1, \theta_2 \geq 0, a \leq \theta_2 - \theta_1, \end{aligned}$$

(with both $\theta_1, \theta_2 \geq 0$ and $\theta_1 + \theta_2 > a$ do not invest in the stock whose rate is smaller, invest aX in the other one if the absolute value of the difference of the stocks rates is larger than a),

$$\lambda_1 = \lambda_2 = \frac{a - \theta_1 - \theta_2}{2}; \pi_1 = \frac{a + \theta_1 - \theta_2}{2}, \pi_2 = \frac{a + \theta_2 - \theta_1}{2}$$

if $\theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 > a > |\theta_1 - \theta_2|$ (if none of the previous conditions is satisfied, invest the amount $\frac{a}{2}X$ in the stocks, corrected by the difference of their rates).

Let us consider now the case where the coefficients $r(\cdot)$, $b(\cdot)$, $\sigma(\cdot)$ of the market model are deterministic functions on $[0, T]$, which we shall take for simplicity to be continuous. Then there is a formal HJB (Hamilton–Jacobi–Bellman) equation associated with the dual optimization problem, specifically,

$$Q_t + \inf_{x \in \tilde{K}} \left[\frac{1}{2} y^2 Q_{yy} \|\theta(t) + \sigma^{-1}(t)x\|^2 - y Q_y \delta(x) \right] - y Q_y r(t) = 0, \quad (9.2)$$

in $[0, T) \times (0, \infty)$;

$$Q(T, y) = \tilde{U}(y); \quad y \in (0, \infty).$$

If there exists a classical solution $Q \in C^{1,2}([0, T) \times (0, \infty))$ of this equation, that satisfies appropriate growth conditions, then standard verification theorems in stochastic control (e.g. Fleming and Soner (1993)) lead to the representation

$$\tilde{V}(y) = Q(0, y), \quad 0 < y < \infty$$

for the dual value function.

Example 9.3 (Cone constraints) Suppose that $\delta \equiv 0$ on $\tilde{K}$. Then

$$\lambda(t) = \arg \min_{x \in \tilde{K}} \|\theta(t) + \sigma^{-1}(t)x\|^2$$

is *deterministic*, the same for all $y \in (0, \infty)$, and the equation (9.2) becomes

$$Q_t + \frac{1}{2} \|\theta_\lambda(t)\|^2 y^2 Q_{yy} - r(t) y Q_y + \tilde{U}_1(t, y) = 0; \quad \text{in } [0, T) \times (0, \infty).$$

Example 9.4 (Power utility) Consider the case $U(x) = x^\alpha/\alpha$, $x \in (0, \infty)$ for some $\alpha \in (0, 1)$. Then $\tilde{U}(y) = \frac{1}{\rho} y^{-\rho}$, $0 < y < \infty$ with $\rho := \alpha/(1 - \alpha)$. Again, the process $\lambda(\cdot)$ is *deterministic*, i.e.

$$\lambda(t) = \arg \min_{x \in \tilde{K}} [\|\theta(t) + \sigma^{-1}(t)x\|^2 + 2(1 - \alpha)\delta(x)],$$

and is the same for all $y \in (0, \infty)$. In this case one finds

$$\pi_\lambda(t) = \frac{1}{1 - \alpha} (\sigma(t)\sigma'(t))^{-1} [b(t) - r(t)\mathbf{1} + \lambda(t)].$$

Example 9.5 (Different interest rates for borrowing and lending) We consider the market with different interest rates for borrowing, R , and lending, r , $R(\cdot) \geq r(\cdot)$. The methodology of the previous section can still be used in the context of the models introduced in Section 6, of which the different interest rates case is just one example. We are looking for an optimal process $\lambda_y \in \mathcal{H}$ for the corresponding dual problem, in which the function $\delta(\cdot)$ is replaced by the function $\tilde{g}(\cdot)$ (see Cvitanić (1997) for details), and, for any given $x \in (0, \infty)$, for an optimal portfolio $\hat{\pi}$ for the

original primal control problem. In the case of logarithmic utility $U(x) = \log x$, we see that $\lambda(t) = \lambda_1(t)\mathbf{1}$, where

$$\lambda_1(t) = \arg \min_{r(t) - R(t) \leq x \leq 0} (-2x + \|\theta(t) + \sigma^{-1}(t)\mathbf{1}x\|^2).$$

With $A(t) := \text{tr}[(\sigma^{-1}(t))'(\sigma^{-1}(t))]$, $B(t) := \theta'(t)\sigma^{-1}(t)\mathbf{1}$, this minimization is achieved as follows:

$$\lambda_1(t) = \left\{ \begin{array}{ll} \frac{1 - B(t)}{A(t)}; & \text{if } 0 < B(t) - 1 < A(t)(R(t) - r(t)) \\ 0; & \text{if } B(t) \leq 1 \\ r(t) - R(t); & \text{if } B(t) - 1 \geq A(t)(R(t) - r(t)) \end{array} \right\}.$$

The optimal portfolio is then computed as

$$\hat{\pi}_t = \left\{ \begin{array}{ll} (\sigma_t \sigma_t')^{-1} \left[b_t - \left(r_t + \frac{B_t - 1}{A_t} \right) \mathbf{1} \right]; & 0 < B_t - 1 \leq A_t(R_t - r_t) \\ (\sigma_t \sigma_t')^{-1} [b_t - r_t \mathbf{1}]; & B_t \leq 1 \\ (\sigma_t \sigma_t')^{-1} [b_t - R_t \mathbf{1}]; & B_t - 1 \geq A_t(R_t - r_t) \end{array} \right.$$

In the case $U(x) = x^\alpha/\alpha$, for some $\alpha \in (0, 1)$, we get $\lambda(t) = \lambda_1(t)\mathbf{1}$ with

$$\begin{aligned} \lambda_1(t) &= \arg \min_{r(t) - R(t) \leq x \leq 0} [-2(1 - \alpha)x + \|\theta(t) + \sigma^{-1}(t)\mathbf{1}x\|^2] \\ &= \left\{ \begin{array}{ll} \frac{1 - \alpha - B(t)}{A(t)}; & \text{if } 0 < B(t) - 1 + \alpha < A(t)(R(t) - r(t)) \\ 0; & \text{if } B(t) \leq 1 - \alpha \\ r(t) - R(t); & \text{if } B(t) - 1 + \alpha \geq A(t)(R(t) - r(t)). \end{array} \right\}. \end{aligned}$$

The optimal portfolio is given as

$$\hat{\pi}_t = \left\{ \begin{array}{ll} \frac{(\sigma_t \sigma_t')^{-1}}{A_t} \left[b_t - \left(r_t + \frac{B_t - 1 + \alpha}{A_t} \right) \mathbf{1} \right]; & 0 < B_t - 1 + \alpha < A_t(R_t - r_t) \\ \frac{(\sigma_t \sigma_t')^{-1}}{1 - \alpha} [b_t - r_t \mathbf{1}]; & B_t \leq 1 - \alpha \\ \frac{(\sigma_t \sigma_t')^{-1}}{1 - \alpha} [b_t - R_t \mathbf{1}]; & B_t - 1 + \alpha \geq A_t(R_t - r_t). \end{array} \right.$$

10 Utility based pricing

How to choose a price of a contingent claim B in the no-arbitrage pricing interval $[\tilde{h}(0), h(0)]$ in the case of incomplete markets, i.e., when the interval is nondegenerate (consists of more than just the Black–Scholes price)? (Here, $\tilde{h}(0)$ is the

maximal price at which the buyer of the option would still be able to hedge away all the risk.) There have been many attempts to provide a satisfactory answer to this question. We describe one suggested by Davis (1997), as presented in Karatzas and Kou (1996), to which we refer for the proofs of the results presented below. The approach is based on the following “zero marginal rate of substitution” principle: given the agent’s utility function U and initial wealth x , the “utility based price” $\hat{p}$ is the one that makes the agent neutral with respect to diversion of a small amount of funds into the contingent claim at time zero, while maximizing the utility from total wealth at the exercise time T . It can be shown that

$$\hat{p} = E[H_{\lambda_x}(T)B], \quad (10.1)$$

where λ_x is the associated optimal dual process. In particular, this price can be calculated in the context of examples of the previous section, and *does not depend on* U and x , in the case of cone constraints ($\delta \equiv 0$) and constant coefficients (Example 9.3). It can also be shown that, in this case, it gives rise to the probability measure P_{λ_x} which minimizes the relative entropy with respect to the original measure P , among all measures P_v , $v \in \mathcal{D}$.

We describe now more precisely what we mean by “utility based price”. For a given $-x < \delta < x$ and price p of the claim, we introduce the value function

$$Q(\delta, p, x) := \sup_{\pi \in \mathcal{A}'(x-\delta)} EU\left(X^{x-\delta}(T) + \frac{\delta}{p}B\right). \quad (10.2)$$

In other words, the agent acquires δ/p units of the claim B at price p at time zero, and maximizes his/her terminal wealth at time T . Davis (1997) suggests the use of the price $\hat{p}$ for which

$$\left. \frac{\partial Q}{\partial \delta}(\delta, \hat{p}, x) \right|_{\delta=0} = 0,$$

so that this diversion of funds has a neutral effect on the expected utility. Since the derivative of Q need not exist, we have the following:

Definition 10.1 For a given $x > 0$, we call $\hat{p}$ a *weak solution* of (10.2) if, for every function $\varphi : (-x, x) \mapsto \mathbb{R}$ of class C^1 which satisfies

$$\varphi(\delta) \geq Q(\delta, \hat{p}, x), \quad \forall \delta \in (-x, x), \quad \varphi(0) = Q(0, p, x) = V(x),$$

we have $\varphi'(0) = 0$. If it is unique, then we call it the *utility based price* of B .

Theorem 10.2 *Under the conditions of Theorem 8.7, the utility based price of B is given as in (10.1).*

11 The transaction costs model

In the remaining sections we consider a financial market with proportional transaction costs. More precisely, the market consists of one riskless asset, a *bank-account* with price $B(\cdot)$ given by

$$dB(t) = B(t)r(t)dt, \quad B(0) = 1,$$

and of one risky asset, *stock*, with price-per-share $S(\cdot)$ governed by the stochastic equation

$$dS(t) = S(t)[b(t)dt + \sigma(t)dW(t)], \quad S(0) = s \in (0, \infty),$$

for $t \in [0, T]$. Here, $W = \{W(t), 0 \leq t \leq T\}$ is a standard, one-dimensional Brownian motion on a complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$, endowed with a filtration $\{\mathcal{F}_t\}$, the augmentation of the filtration generated by $W(\cdot)$. The coefficients of the model $r(\cdot)$, $b(\cdot)$ and $\sigma(\cdot) > 0$ are assumed to be bounded and $\mathbf{F}$ -progressively measurable processes; furthermore, $\sigma(\cdot)$ is also assumed to be bounded away from zero (uniformly in (t, ω)).

Now, a *trading strategy* is a pair (L, M) of $\mathbf{F}$ -adapted processes on $[0, T]$, with left-continuous, nondecreasing paths and $L(0) = M(0) = 0$; $L(t)$ (respectively, $M(t)$) represents the total amount of funds transferred from bank-account to stock (respectively, from stock to bank-account) by time t . Given *proportional transaction costs* $0 < \lambda, \mu < 1$ for such transfers, and initial holdings x, y in bank and stock, respectively, the portfolio holdings $X(\cdot) = X^{x,L,M}(\cdot)$, $Y(\cdot) = Y^{y,L,M}(\cdot)$ corresponding to a given trading strategy (L, M) , evolve according to the equations:

$$X(t) = x - (1 + \lambda)L(t) + (1 - \mu)M(t) + \int_0^t X(u)r(u)du, \quad 0 \leq t \leq T \quad (11.1)$$

$$Y(t) = y + L(t) - M(t) + \int_0^t Y(u)[b(u)du + \sigma(u)dW(u)], \quad 0 \leq t \leq T. \quad (11.2)$$

Definition 11.1 A *contingent claim* is a pair (C_0, C_1) of $\mathcal{F}_T$ -measurable random variables. We say that a trading strategy (L, M) *hedges the claim* (C_0, C_1) *starting with* (x, y) as initial holdings, if $X(\cdot), Y(\cdot)$ of (11.1), (11.2) satisfy

$$X(T) + (1 - \mu)Y(T) \geq C_0 + (1 - \mu)C_1 \quad (11.3)$$

$$X(T) + (1 + \lambda)Y(T) \geq C_0 + (1 + \lambda)C_1. \quad (11.4)$$

Interpretation: Here C_0 (respectively, C_1) is understood as a target-position in the bank-account (resp., the stock) at the terminal time $t = T$: for example

$$C_0 = -k\mathbf{1}_{\{S(T) > k\}}, \quad C_1 = S(T)\mathbf{1}_{\{S(T) > k\}}$$

in the case of a European call-option; and

$$C_0 = k\mathbf{1}_{\{S(T) < k\}}, \quad C_1 = -S(T)\mathbf{1}_{\{S(T) < k\}}$$

for a European put-option (both with exercise price $k \geq 0$).

“Hedging”, in the sense of (11.3) and (11.4), simply means that one is able to cover these positions at $t = T$. Indeed, assume that we have both $Y(T) \geq C_1$ and (11.3), in the form

$$X(T) + (1 - \mu)[Y(T) - C_1] \geq C_0;$$

then (11.4) holds too, and the agent can cover the position in the bank-account as well, by transferring the amount $Y(T) - C_1 \geq 0$ to it. Similarly for the case $Y(T) < C_1$.

The equations (11.1), (11.2) can be written in the equivalent form

$$d\left(\frac{X(t)}{B(t)}\right) = \left(\frac{1}{B(t)}\right) [(1 - \mu)dM(t) - (1 + \lambda)dL(t)], \quad X(0) = x \quad (11.5)$$

$$d\left(\frac{Y(t)}{S(t)}\right) = \left(\frac{1}{S(t)}\right) [dL(t) - dM(t)], \quad Y(0) = y \quad (11.6)$$

in terms of “number-of-shares” (rather than amounts) held.

12 State-price densities

Consider the class $\mathcal{D}$ of pairs of strictly positive $\mathbf{F}$ -martingales $(Z_0(\cdot), Z_1(\cdot))$ with

$$Z_0(0) = 1, \quad z := Z_1(0) \in [s(1 - \mu), s(1 + \lambda)]$$

and

$$1 - \mu \leq R(t) := \frac{Z_1(t)}{Z_0(t)P(t)} \leq 1 + \lambda, \quad \forall 0 \leq t \leq T, \quad (12.1)$$

where

$$P(t) := \frac{S(t)}{B(t)} = s + \int_0^t P(u)[(b(u) - r(u))du + \sigma(u)dW(u)], \quad 0 \leq t \leq T \quad (12.2)$$

is the *discounted stock price*.

The martingales $Z_0(\cdot), Z_1(\cdot)$ are the feasible *state-price densities* for holdings in bank and stock, respectively, in this market with transaction costs; as such, they reflect the “constraints” or “frictions” inherent in this market, in the form of condition (12.1). From the martingale representation theorem there exist $\mathbf{F}$ -progressively measurable processes $\theta_0(\cdot), \theta_1(\cdot)$ with $\int_0^T (\theta_0^2(t) + \theta_1^2(t))dt < \infty$ a.s. and

$$Z_i(t) = Z_i(0) \exp \left\{ \int_0^t \theta_i(s)dW(s) - \frac{1}{2} \int_0^t \theta_i^2(s)ds \right\}, \quad i = 0, 1; \quad (12.3)$$

thus, the process $R(\cdot)$ of (12.1) has the dynamics

$$dR(t) = R(t)[\sigma^2(t) + r(t) - b(t) - (\theta_1(t) - \theta_0(t))(\sigma(t) + \theta_0(t))]dt + R(t)(\theta_1(t) - \sigma(t) - \theta_0(t))dW(t), \quad R(0) = z/s. \quad (12.4)$$

Remark 12.1 A rather “special” pair $(Z_0^*(\cdot), Z_1^*(\cdot)) \in \mathcal{D}$ is obtained, if we take in (12.3) the processes $(\theta_0(\cdot), \theta_1(\cdot))$ to be given as

$$\theta_0^*(t) := \frac{r(t) - b(t)}{\sigma(t)}, \quad \theta_1^*(t) := \sigma(t) + \theta_0^*(t), \quad 0 \leq t \leq T, \quad (12.5)$$

and let $Z_0^*(0) = 1$, $s(1 - \mu) \leq Z_1^*(0) = z \leq s(1 + \lambda)$. Because then, from (12.4), $R^*(\cdot) := Z_1^*(\cdot)/(Z_0^*(\cdot)P(\cdot)) \equiv z/s$; in fact, the pair of (12.5) and $z = s$ provide the *only* member $(Z_0^*(\cdot), Z_1^*(\cdot))$ of $\mathcal{D}$, if $\lambda = \mu = 0$. Notice that the processes $\theta_0^*(\cdot)$, $\theta_1^*(\cdot)$ of (12.5) are *bounded*.

Let us observe also that

$$\begin{aligned} Z_0(t) \frac{X(t)}{B(t)} + Z_1(t) \frac{Y(t)}{S(t)} + \int_0^t \frac{Z_0(s)}{B(s)} [(1 + \lambda) - R(s)] dL(s) \\ + \int_0^t \frac{Z_0(s)}{B(s)} [R(s) - (1 - \mu)] dM(s) \\ = x + \frac{yz}{s} + \int_0^t \frac{Z_0(s)}{B(s)} [X(s)\theta_0(s) + R(s)Y(s)\theta_1(s)] dW(s), \\ t \in [0, T] \end{aligned} \quad (12.6)$$

is a **P**-local martingale, for any $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$ and any trading strategy (L, M) ; this follows directly from (11.5), (11.6), (12.3) and the product rule. Equivalently, (12.6) can be re-written as

$$\begin{aligned} \frac{X(t) + R(t)Y(t)}{B(t)} + \int_0^t \frac{(1 + \lambda) - R(s)}{B(s)} dL(s) + \int_0^t \frac{R(s) - (1 - \mu)}{B(s)} dM(s) \\ = x + \frac{yz}{s} + \int_0^t \frac{R(s)Y(s)}{B(s)} (\theta_1(s) - \theta_0(s)) dW_0(s), \end{aligned} \quad (12.7)$$

where

$$W_0(t) := W(t) - \int_0^t \theta_0(s) ds, \quad 0 \leq t \leq T \quad (12.8)$$

is a Brownian motion under the equivalent probability measure

$$\mathbf{P}_0(A) := E[Z_0(T)\mathbf{1}_A], \quad A \in \mathcal{F}_T. \quad (12.9)$$

We shall denote by $Z_0^*(\cdot)$, $W_0^*(\cdot)$ and $\mathbf{P}_0^*$ the processes and probability measure, respectively, corresponding to the process $\theta_0^*(\cdot)$ of (12.5), via the equations (12.3)

(with $Z_0^*(0) = 1$), (12.8) and (12.9). With this notation, (12.2) becomes $dP(t) = P(t)\sigma(t)dW_0^*(t)$, $P(0) = s$.

Definition 12.2 Let $\mathcal{D}_\infty$ be the class of positive martingales $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$, for which the random variable

$$\frac{Z_0(T)}{Z_0^*(T)}, \quad \text{and thus also } \frac{Z_1(T)}{Z_0^*(T)P(T)},$$

is essentially bounded.

Definition 12.3 We say that a given trading strategy (L, M) is *admissible* for (x, y) , and write $(L, M) \in \mathcal{A}(x, y)$, if

$$\frac{X(\cdot) + R(\cdot)Y(\cdot)}{B(\cdot)} \text{ is a } \mathbf{P}_0\text{-supermartingale, } \forall (Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}_\infty. \quad (12.10)$$

Consider, for example, a trading strategy (L, M) that satisfies the no-bankruptcy conditions

$$X(t) + (1 + \lambda)Y(t) \geq 0 \text{ and } X(t) + (1 - \mu)Y(t) \geq 0, \quad \forall 0 \leq t \leq T.$$

Then $X(\cdot) + R(\cdot)Y(\cdot) \geq 0$ for every $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$ (recall (12.1), and note Remark 12.4 below); this means that the $\mathbf{P}_0$ -local martingale of (12.7) is nonnegative, hence a $\mathbf{P}_0$ -supermartingale. But the second and the third terms

$$\int_0^\cdot \frac{1 + \lambda - R(s)}{B(s)} dL(s), \quad \int_0^\cdot \frac{R(s) - (1 - \mu)}{B(s)} dM(s)$$

in (12.7) are increasing processes, thus the first term $(X(\cdot) + R(\cdot)Y(\cdot))/B(\cdot)$ is also a $\mathbf{P}_0$ -supermartingale, for every pair $(Z_0(\cdot), Z_1(\cdot))$ in $\mathcal{D}$. The condition (12.10) is actually weaker, in that it requires this property only for pairs in $\mathcal{D}_\infty$. This provides a motivation for Definition 12.3, specifically, *to allow for as wide a class of trading strategies as possible, and still exclude arbitrage opportunities*. This is usually done by imposing a lower bound on the wealth process; however, that excludes simple strategies of the form “trade only once, by buying a fixed number of shares of the stock at a specified time t ”, which may require (unbounded) borrowing. We will need to use such strategies in the sequel.

Remark 12.4 Here is a trivial (but useful) observation: if $x + (1 - \mu)y \geq a + (1 - \mu)b$ and $x + (1 + \lambda)y \geq a + (1 + \lambda)b$, then $x + ry \geq a + rb$, $\forall 1 - \mu \leq r \leq 1 + \lambda$.

13 The minimal superreplication price

Suppose that we are given an initial holding $y \in \mathbb{R}$ in the stock, and want to hedge a given contingent claim (C_0, C_1) with strategies which are admissible (in the sense

of Definitions 11.1, 12.2. What is the smallest amount of holdings in the bank

$$h(C_0, C_1; y) := \inf\{x \in \mathbb{R} / \exists (L, M) \in \mathcal{A}(x, y) \text{ and } (L, M) \text{ hedges } (C_0, C_1)\} \quad (13.1)$$

that allows us to do this? We call $h(C_0, C_1; y)$ the *superreplication price* of the contingent claim (C_0, C_1) for initial holding y in the stock, and with the convention that $h(C_0, C_1; y) = \infty$ if the set in (13.1) is empty.

Suppose this is not the case, and let $x \in \mathbb{R}$ belong to the set of (13.1); then for any $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}_\infty$ we have from (12.10), the Definition 11.1 of hedging, and Remark 12.4:

$$\begin{aligned} x + \frac{y}{s} E Z_1(T) &= x + \frac{y}{s} z \geq E_0 \left[\frac{X(T) + R(T)Y(T)}{B(T)} \right] \\ &\geq E_0 \left[\frac{C_0 + R(T)C_1}{B(T)} \right] = E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) \right], \end{aligned}$$

so that $x \geq E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right]$. Therefore

$$h(C_0, C_1; y) \geq \sup_{\mathcal{D}_\infty} E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right], \quad (13.2)$$

and this inequality is clearly also valid if $h(C_0, C_1; y) = \infty$.

Lemma 13.1 *If the contingent claim (C_0, C_1) is bounded from below, in the sense*

$$C_0 + (1 + \lambda)C_1 \geq -K \text{ and } C_0 + (1 - \mu)C_1 \geq -K, \text{ for some } 0 \leq K < \infty, \quad (13.3)$$

then

$$\begin{aligned} \sup_{\mathcal{D}_\infty} E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right] \\ = \sup_{\mathcal{D}} E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right]. \end{aligned}$$

Proof Start with arbitrary $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$ and define the sequence of stopping times $\{\tau_n\} \uparrow T$ by

$$\tau_n := \inf \left\{ t \in [0, T] / \frac{Z_0(t)}{Z_0^*(t)} \geq n \right\} \wedge T, \quad n \in \mathbb{N}.$$

Consider also, for $i = 0, 1$ and in the notation of (12.5):

$$\theta_i^{(n)}(t) := \begin{cases} \theta_i(t), & 0 \leq t < \tau_n \\ \theta_i^*(t), & \tau_n \leq t \leq T \end{cases}$$

and

$$Z_i^{(n)}(t) = z_i \exp \left\{ \int_0^t \theta_i^{(n)}(s) dW(s) - \frac{1}{2} \int_0^t (\theta_i^{(n)}(s))^2 ds \right\}$$

with $z_0 = 1$, $z_1 = Z_1(0) = E Z_1(T)$. Then, for every $n \in \mathbb{N}$, both $Z_0^{(n)}(\cdot)$ and $Z_1^{(n)}(\cdot)$ are positive martingales, $R^{(n)}(\cdot) = Z_1^{(n)}(\cdot)/(Z_0^{(n)}(\cdot)P(\cdot)) = R(\cdot \wedge \tau_n)$ takes values in $[1 - \mu, 1 + \lambda]$ (by (12.1) and Remark 12.1), and $Z_0^{(n)}(\cdot)/Z_0^*(\cdot)$ is bounded by n (in fact, constant on $[\tau_n, T]$). Therefore, $(Z_0^{(n)}(\cdot), Z_1^{(n)}(\cdot)) \in \mathcal{D}_\infty$. Now let κ denote an upper bound on $K/B(T)$, and observe, from Remark 12.4, (13.3) and Fatou's lemma:

$$\begin{aligned} & E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right] + \frac{y}{s} Z_1(0) + \kappa \\ &= E \left[Z_0(T) \left\{ \frac{C_0 + R(T)C_1}{B(T)} + \kappa \right\} \right] \\ &= E \left[\lim_n Z_0^{(n)}(T) \left\{ \frac{C_0 + R^{(n)}(T)C_1}{B(T)} + \kappa \right\} \right] \\ &\leq \lim_n E \left[Z_0^{(n)}(T) \left\{ \frac{C_0 + R^{(n)}(T)C_1}{B(T)} + \kappa \right\} \right] \\ &= \lim_n E \left[\frac{Z_0^{(n)}(T)}{B(T)} (C_0 + R^{(n)}(T)C_1) - \frac{y}{s} Z_1^{(n)}(T) \right] + \frac{y}{s} Z_1(0) + \kappa. \end{aligned}$$

This shows that the left-hand side dominates the right-hand side in the statement of the lemma; the reverse inequality is obvious. $\square$

Remark 13.2 Formally taking $y = 0$ in the above, we deduce

$$E_0 \left(\frac{C_0 + R(T)C_1}{B(T)} \right) \leq \lim_{n \rightarrow \infty} E_0^{(n)} \left(\frac{C_0 + R^{(n)}(T)C_1}{B(T)} \right), \quad (13.4)$$

where $E_0, E_0^{(n)}$ denote expectations with respect to the probability measures $\mathbf{P}_0$ of (12.9) and $\mathbf{P}_0^{(n)}(\cdot) = E[Z_0^{(n)}(T)\mathbf{1}_\cdot]$, respectively.

Here is the main result of this section.

Theorem 13.3 Under the conditions (13.3) and

$$E_0^*(C_0^2 + C_1^2) < \infty, \quad (13.5)$$

we have

$$h(C_0, C_1; y) = \sup_{\mathcal{D}} E \left[\frac{Z_0(T)}{B(T)} (C_0 + R(T)C_1) - \frac{y}{s} Z_1(T) \right].$$

In (13.5), E_0^* denotes expectation with respect to the probability measure $\mathbf{P}_0^*$. The conditions (13.3), (13.5) are both easily verified for a European call or put. In fact, one can show that if a pair of admissible terminal holdings $(X(T), Y(T))$ hedges a pair $(\tilde{C}_0, \tilde{C}_1)$ satisfying (13.5) (for example, $(\tilde{C}_0, \tilde{C}_1) \equiv (0, 0)$), then necessarily the pair $(X(T), Y(T))$ also satisfies (13.5) – and so does any other pair of random variables (C_0, C_1) which are bounded from below and are hedged by $(X(T), Y(T))$. In particular, any strategy which satisfies the “no-bankruptcy” condition of hedging $(0, 0)$, necessarily results in a square-integrable final wealth. In this sense, the condition (13.5) is consistent with the standard “no-bankruptcy” condition, hence not very restrictive (this, however, is not necessarily the case if there are no transaction costs).

Proof In view of Lemma 13.1 and the inequality (13.2), it suffices to show

$$h(C_0, C_1; y) \leq \sup_{\mathcal{D}} E \left[Z_0(T) \frac{C_0}{B(T)} + Z_1(T) \left(\frac{C_1}{S(T)} - \frac{y}{s} \right) \right] =: R. \quad (13.6)$$

For simplicity we take $s = 1$, $r(\cdot) \equiv 0$, thus $B(\cdot) \equiv 1$, for the remainder of the section; the reader will verify easily that this entails no loss of generality.

We start by taking an arbitrary $b < h(C_0, C_1; y)$ and considering the sets

$$A_0 := \{(U, V) \in (\mathbf{L}_2^*)^2 : \exists (L, M) \in \mathcal{A}(0, 0) \text{ that hedges } (U, V) \text{ starting with } x = 0, y = 0\} \quad (13.7)$$

$$A_1 := \{(C_0 - b, C_1 - yS(T))\},$$

where $\mathbf{L}_2^* = \mathbf{L}_2(\Omega, \mathcal{F}_T, \mathbf{P}_0^*)$. It is not hard to prove (see below) that

$$A_0 \text{ is a convex cone, and contains the origin } (0, 0), \text{ in } (\mathbf{L}_2^*)^2, \quad (13.8)$$

$$A_0 \cap A_1 = \emptyset. \quad (13.9)$$

It is, however, considerably harder to establish that

$$A_0 \text{ is closed in } (\mathbf{L}_2^*)^2. \quad (13.10)$$

The proof can be found in the appendix of Cvitanić and Karatzas (1996). From (13.8)–(13.10) and the Hahn–Banach theorem there exists a pair of random variables $(\rho_0^*, \rho_1^*) \in (\mathbf{L}_2^*)^2$, not equal to $(0, 0)$, such that

$$E_0^*[\rho_0^* V_0 + \rho_1^* V_1] = E[\rho_0 V_0 + \rho_1 V_1] \leq 0, \quad \forall (V_0, V_1) \in A_0 \quad (13.11)$$

$$E_0^*[\rho_0^*(C_0 - b) + \rho_1^*(C_1 - yS(T))] = E[\rho_0(C_0 - b) + \rho_1(C_1 - yS(T))] \geq 0, \quad (13.12)$$

where $\rho_i := \rho_i^* Z_0^*(T)$, $i = 0, 1$. It is also not hard to check (see below) that

$$(1 - \mu)E[\rho_0|\mathcal{F}_t] \leq \frac{E[\rho_1 S(T)|\mathcal{F}_t]}{S(t)} \leq (1 + \lambda)E[\rho_0|\mathcal{F}_t], \quad \forall 0 \leq t \leq T \quad (13.13)$$

$$\rho_1 \geq 0, \quad \rho_0 \geq 0 \text{ and } E[\rho_0] > 0, \quad E[\rho_1 S(T)] > 0. \quad (13.14)$$

In view of (13.14), we may take $E[\rho_0] = 1$, and then (13.12) gives

$$b \leq E[\rho_0 C_0 + \rho_1 (C_1 - yS(T))]. \quad (13.15)$$

Consider now arbitrary $0 < \varepsilon < 1$, $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$, and define

$$\tilde{Z}_0(t) := \varepsilon Z_0(t) + (1 - \varepsilon)E[\rho_0|\mathcal{F}_t], \quad \tilde{Z}_1(t) := \varepsilon Z_1(t) + (1 - \varepsilon)E[\rho_1 S(T)|\mathcal{F}_t],$$

for $0 \leq t \leq T$. Clearly these are positive martingales, and $\tilde{Z}_0(0) = 1$; on the other hand, multiplying in (13.13) by $1 - \varepsilon$, and in $(1 - \mu)Z_0(t) \leq Z_1(t)/S(t) \leq (1 + \lambda)Z_0(t)$, $0 \leq t \leq T$ by ε , and adding up, we obtain $(\tilde{Z}_0(\cdot), \tilde{Z}_1(\cdot)) \in \mathcal{D}$. Thus, in the notation of (13.6),

$$\begin{aligned} R &\geq E \left[\tilde{Z}_0(T)C_0 + \tilde{Z}_1(T) \left(\frac{C_1}{S(T)} - y \right) \right] \\ &= (1 - \varepsilon)E[\rho_0 C_0 + \rho_1 (C_1 - yS(T))] \\ &\quad + \varepsilon E \left[Z_0(T)C_0 + Z_1(T) \left(\frac{C_1}{S(T)} - y \right) \right] \\ &\geq b(1 - \varepsilon) + \varepsilon E \left[Z_0(T)C_0 + Z_1(T) \left(\frac{C_1}{S(T)} - y \right) \right] \end{aligned}$$

from (13.15); letting $\varepsilon \downarrow 0$ and then $b \uparrow h(C_0, C_1; y)$, we obtain (13.6), as required to complete the proof of Theorem 13.3. $\square$

Proof of (13.9) Suppose that $A_0 \cap A_1$ is not empty, i.e., that there exists $(L, M) \in \mathcal{A}(0, 0)$ such that, with $X(\cdot) = X^{0,L,M}(\cdot)$ and $Y(\cdot) = Y^{0,L,M}(\cdot)$, the process $X(\cdot) + R(\cdot)Y(\cdot)$ is a $\mathbf{P}_0$ -supermartingale for every $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}_\infty$, and we have:

$$X(T) + (1 - \mu)Y(T) \geq (C_0 - b) + (1 - \mu)(C_1 - yS(T)),$$

$$X(T) + (1 + \lambda)Y(T) \geq (C_0 - b) + (1 + \lambda)(C_1 - yS(T)).$$

But then, with

$$\tilde{X}(\cdot) := X^{b,L,M}(\cdot) = b + X(\cdot), \quad \tilde{Y}(\cdot) := Y^{y,L,M}(\cdot) = Y(\cdot) + yS(\cdot)$$

we have, from above, that $\tilde{X}(\cdot) + R(\cdot)\tilde{Y}(\cdot) = X(\cdot) + R(\cdot)Y(\cdot) + b + yZ_1(\cdot)/Z_0(\cdot)$ is a $\mathbf{P}_0$ -supermartingale for every $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}_\infty$, and that

$$\tilde{X}(T) + (1 - \mu)\tilde{Y}(T) \geq C_0 + (1 - \mu)C_1,$$

$$\tilde{X}(T) + (1 + \lambda)\tilde{Y}(T) \geq C_0 + (1 + \lambda)C_1.$$

In other words, (L, M) belongs to $\mathcal{A}(b, y)$ and hedges (C_0, C_1) starting with (b, y) – a contradiction to the definition (13.1), and to the fact that $h(C_0, C_1; y) > b$. $\square$

Proof of (13.13) and (13.14) Fix $t \in [0, T]$ and let ξ be an arbitrary bounded, non-negative, $\mathcal{F}_t$ -measurable random variable. Consider the strategy of starting with $(x, y) = (0, 0)$ and buying ξ shares of stock at time $s = t$, otherwise doing nothing (“buy-and-hold strategy”); more explicitly, $M^\xi(\cdot) \equiv 0$, $L^\xi(s) = \xi S(t)\mathbf{1}_{(t, T]}(s)$ and thus

$$\begin{aligned} X^\xi(s) &:= X^{0, L^\xi, M^\xi}(\cdot) = -\xi(1 + \lambda)S(t)\mathbf{1}_{(t, T]}(s), \\ Y^\xi(s) &:= Y^{0, L^\xi, M^\xi}(s) = \xi S(s)\mathbf{1}_{(t, T]}(s), \end{aligned}$$

for $0 \leq s \leq T$. Consequently, $Z_0(s)[X^\xi(s) + R(s)Y^\xi(s)] = \xi[Z_1(s) - (1 + \lambda)S(t)Z_0(s)]\mathbf{1}_{(t, T]}(s)$ is a **P**-supermartingale for every $(Z_0(\cdot), Z_1(\cdot)) \in \mathcal{D}$, since, for instance with $t < s \leq T$:

$$\begin{aligned} E[Z_0(s)(X_s^\xi + R_s Y_s^\xi) | \mathcal{F}_t] &= \xi(E[Z_1(s) | \mathcal{F}_t] - (1 + \lambda)S_t E[Z_0(s) | \mathcal{F}_t]) \\ &= \xi[Z_1(t) - (1 + \lambda)S(t)Z_0(t)] = \xi S(t)Z_0(t)[R(t) - (1 + \lambda)] \\ &\leq 0 = Z_0(t)[X_t^\xi + R(t)Y_t^\xi]. \end{aligned}$$

Therefore, $(L^\xi, M^\xi) \in \mathcal{A}(0, 0)$, thus $(X^\xi(T), Y^\xi(T))$ belongs to the set A_0 of (13.7), and, from (13.11):

$$\begin{aligned} 0 &\geq E[\rho_0 X^\xi(T) + \rho_1 Y^\xi(T)] = E[\xi(\rho_1 S(T) - (1 + \lambda)\rho_0 S(t))] \\ &= E[\xi(E[\rho_1 S(T) | \mathcal{F}_t] - (1 + \lambda)S(t)E[\rho_0 | \mathcal{F}_t])]. \end{aligned}$$

From the arbitrariness of $\xi \geq 0$, we deduce the inequality of the right-hand side in (13.13), and a dual argument gives the inequality of the left-hand side, for given $t \in [0, T]$. Now all three processes in (13.13) have continuous paths; consequently, (13.13) is valid for *all* $t \in [0, T]$.

Next, we notice that (13.13) with $t = T$ implies $(1 - \mu)\rho_0 \leq \rho_1 \leq (1 + \lambda)\rho_0$, so that ρ_0 , hence also ρ_1 , is nonnegative. Similarly, (13.13) with $t = 0$ implies $(1 - \mu)E[\rho_0] \leq E[\rho_1 S(T)] \leq (1 + \lambda)E[\rho_0]$, and therefore, since (ρ_0, ρ_1) is not equal to $(0, 0)$, $E[\rho_0] > 0$, hence also $E[\rho_1 S(T)] > 0$. This proves (13.14). $\square$

Remark 13.4 For the European call-option with $y = 0$, we have

$$h(C_0, C_1; 0) = \sup_{\mathcal{D}} E \left[Z_1(T)\mathbf{1}_{\{S(T) > k\}} - k \frac{Z_0(T)}{B(T)}\mathbf{1}_{\{S(T) > k\}} \right],$$

and therefore, $h(C_0, C_1, 0) \leq \sup_{\mathcal{D}} E[Z_1(T)] = \sup_{\mathcal{D}} Z_1(0) \leq (1 + \lambda)s$. The number $(1 + \lambda)s$ corresponds to the cost of the “buy-and-hold strategy”, of acquiring one share of the stock at $t = 0$, and holding on to it until $t = T$. Davis and

Clark (1994) conjectured that this hedging strategy is actually the least expensive superreplication strategy:

$$h(C_0, C_1, 0) = (1 + \lambda)s.$$

The conjecture was proved by Soner, Shreve and Cvitanić (1995) by analytic methods. Moreover, the following analogous result has been obtained in more general continuous-time models and for more general contingent claims by Levental and Skorohod (1997) (using probabilistic methods) and Cvitanić, Pham and Touzi (1998) (using Theorem 13.3): “the cheapest buy-and-hold strategy which dominates a given claim in a market with transaction costs is equal to its least expensive superreplication strategy”. However, the result is not always true, and, in particular, it does not hold for discrete-time models.

14 Utility maximization under transaction costs

Consider now a small investor who starts with initial capital $(x, 0)$, $x > 0$, and derives utility $U(X(T+))$ from his *terminal wealth*

$$X(T+) := X(T) + f(Y(T)) \geq 0, \text{ where } f(u) := \begin{cases} (1 + \lambda)u; & u \leq 0 \\ (1 - \mu)u; & u > 0 \end{cases}.$$

In other words, this agent liquidates at time T his position in the stock, incurs the appropriate transaction cost, and collects all the money in the bank-account. Denote by $\mathcal{A}^+(x)$ the set of terminal holdings $(X(T), Y(T))$ that hedge $(0, 0)$, so that, in particular, $X(T+) \geq 0$. The agent's optimization problem is to find an admissible pair $(\hat{L}, \hat{M}) \in \mathcal{A}^+(x)$ that *maximizes expected utility from terminal wealth*, i.e., attains the supremum

$$V(x) := \sup_{\mathcal{A}^+(x)} EU(X(T+)). \quad (14.1)$$

Here, $U : (0, \infty) \rightarrow \mathbb{R}$ is a strictly concave, strictly increasing, continuously differentiable utility function which satisfies $U'(0+) = \infty$, $U'(\infty) = 0$ and

Assumption 14.1 The utility function $U(x)$ has asymptotic elasticity strictly less than 1, i.e.

$$AE(U) := \limsup_{x \rightarrow \infty} \frac{xU'(x)}{U(x)} < 1. \quad (14.2)$$

It is shown in Kramkov and Schachermayer (1998) (henceforth [KS98]) that this condition is basically necessary and sufficient to ensure nice properties of value function $V(x)$ and the existence of an optimal solution.

We are again going to consider the dual problem. However, unlike the case of portfolio constraints, we have to go beyond the set of state-price densities for the dual problem, and we introduce the set

$$\mathcal{H} := \left\{ Z \in \mathbf{L}_+^0 / E \left[\frac{Z}{B(T)} (X(T) + f(Y(T))) \right] \leq x, \right. \\ \left. \forall (X(T), Y(T)) \in \mathcal{A}^+(x) \right\}. \quad (14.3)$$

(Here, $\mathbf{L}^0$ is the set of all random variables on $(\Omega, \mathcal{F}, \mathbf{P})$.) In particular, if $(Z_0(T), Z_1(T)) \in \mathcal{D}$, then $Z_0(T) \in \mathcal{H}$. For a given $z > 0$, the *auxiliary dual problem* associated with (14.1) is given by

$$\tilde{V}(z) := \inf_{Z \in \mathcal{H}} E \tilde{U}(zZ/B(T)). \quad (14.4)$$

More precisely, similarly as in Cvitanić and Karatzas (1996) (henceforth [CK96]), for every $z > 0$, $Z \in \mathcal{H}$ and $(X(T), Y(T)) \in \mathcal{A}^+(x)$ we have

$$EU(X(T+)) \leq E[\tilde{U}(zZ/B(T)) + X(T+)Z/B(T)] \leq E\tilde{U}(zZ/B(T)) + zx. \quad (14.5)$$

Consequently, we have

$$V(x) \leq \inf_{z>0} [\tilde{V}(z) + zx] =: \inf_{z>0} \gamma(z). \quad (14.6)$$

Remark 14.2 The duality approach used in the market with portfolio constraints suggests that we should look for pairs $(\hat{z}, \hat{Z}) \in (0, \infty) \times \mathcal{H}$ and $(\hat{X}(T+), 0) \in \mathcal{A}^+(x)$ such the inequalities in (14.5) and (14.6) become equalities. The pair $(\hat{X}(T+), 0)$ is then optimal for (14.1). It is easily seen that this is the case (i.e. that those inequalities become equalities) if and only if

$$(\hat{X}(T+), 0) = (I(\hat{z}\hat{Z}/B(T)), 0) \in \mathcal{A}^+(x), \quad E \left[\hat{Z} I \left(\frac{\hat{z}\hat{Z}}{B(T)} \right) \right] = x.$$

We first state our results and then provide the proofs.

Proposition 14.3 *For every $z > 0$ there exists $\hat{Z}_z \in \mathcal{H}$ that attains the infimum in (14.4).*

Proposition 14.4 *For every $x \in (0, \infty)$ there exists $\hat{z} \in (0, \infty)$ that attains the infimum of $\gamma(z)$ in (14.6).*

Denote $\hat{Z} := \hat{Z}_{\hat{z}}$ the optimal solution to (14.4) with $z = \hat{z}$ denoting the optimal solution to $\inf_{z>0} \gamma(z)$ of (14.6). The main result of this section is the following:

Theorem 14.5 *The pair $(\hat{C}_0, 0) := (I(\hat{z}\hat{Z}/B(T)), 0)$ belongs to the set $\mathcal{A}^+(x)$ of (nonnegative) terminal holdings that can be hedged starting with initial wealth $x > 0$ in the bank-account. Furthermore,*

$$E \left[U \left(I \left(\frac{\hat{z}\hat{Z}}{B(T)} \right) \right) \right] = V(x) = \inf_{z>0} [\tilde{V}(z) + zx] = \tilde{V}(\hat{z}) + x\hat{z}.$$

In particular, the strategy that hedges $(\hat{C}_0, 0)$ is optimal for the utility maximization problem (14.1).

Remark 14.6 Under Assumption 14.1, there exist $z_0 > 0$, $0 < \gamma, \mu < 1$ and $0 < c < \infty$ such that

$$zI(z) < \frac{\gamma}{1-\gamma} \tilde{U}(z) \quad \text{and} \quad \tilde{U}(\mu z) < c\tilde{U}(z), \quad \forall 0 < z < z_0; \quad (14.7)$$

see [KS98] Lemma 6.3 and Corollary 6.1 for details.

Proof of Proposition 14.3 We first observe that $\mathcal{H}$ is convex, closed under a.s.-convergence by Fatou's lemma, and bounded in $\mathbf{L}^1(\mathbf{P})$; the latter is seen by setting $(X(T), Y(T)) = (xB(T), 0)$ in (14.3), implying $E[Z] \leq 1$ for $Z \in \mathcal{H}$. Fix $z > 0$ and let $\{Z_n\}$ be a minimizing sequence for (14.4). By Komlós' theorem (see Schwartz (1986)), there exists a subsequence Z'_k such that

$$\tilde{Z}_k := \frac{1}{k} \sum_{i=1}^k Z'_i \rightarrow \hat{Z}_z \in \mathcal{H}$$

as $k \rightarrow \infty$, almost surely. As in Lemma 3.4 of [KS98], Fatou's lemma is applicable here, so that $\liminf_{k \rightarrow \infty} E\tilde{U}(z\tilde{Z}_k) \geq E\tilde{U}(z\hat{Z}_z)$. In conjunction with convexity of $\tilde{U}$ this easily implies that $\hat{Z}_z$ is optimal for (14.4). $\square$

For a given progressively measurable process $\theta(\cdot)$ introduce the local martingale

$$Z_\theta(t) := \exp \left\{ \int_0^t \theta(s) dW(s) - \frac{1}{2} \int_0^t \theta^2(s) ds \right\}, \quad 0 \leq t \leq T. \quad (14.8)$$

In this section we will use the notation $Z_0 := Z_{\theta_0^*}(T)$ for the risk-neutral density for the market without transaction costs, where, as before, $\theta^*(t) := (r(t) - b(t))/\sigma(t)$. We have $Z_0 \in \mathcal{H}$.

Lemma 14.7 *The value function $V(\cdot) : (0, \infty) \rightarrow \mathbb{R}$ is finite, decreasing and strictly convex.*

Proof It is straightforward to check that $\tilde{V}(\cdot)$ is decreasing and strictly convex. Next, since $r(\cdot)$ is bounded, we have $k^{-1} \leq B(T) \leq k$ for some $k > 0$. In

conjunction with Jensen's inequality, we obtain

$$E\tilde{U}(zZ/B(T)) \geq \tilde{U}(zkE[Z]) \geq \tilde{U}(zk), \quad (14.9)$$

hence $\tilde{V}(z) \geq \tilde{U}(zk) > -\infty$. On the other hand, Assumption 14.1 ensures the existence of $0 < \alpha < 1$, $z_1 > 0$ such that

$$\tilde{U}(\mu z_1) < \mu^{\alpha/(\alpha-1)} \tilde{U}(z_1) \quad \text{for all } 0 < \mu < 1;$$

see [KS98] Lemma 6.3 for the proof. We get, since $Z_0 \in \mathcal{H}$,

$$\begin{aligned} \tilde{V}(z) &\leq E\tilde{U}(zZ_0/B(T)) \\ &= E[\tilde{U}(zZ_0/B(T))\mathbf{1}_{\{zZ_0/B(T) > z_1\}}] + E[\tilde{U}(zZ_0/B(T))\mathbf{1}_{\{zZ_0/B(T) \leq z_1\}}] \\ &\leq |\tilde{U}(z_1)| + (z/z_1)^{\alpha/(\alpha-1)} |\tilde{U}(z_1)| \cdot E[(Z_0/B(T))^{\alpha/(\alpha-1)}] < \infty. \end{aligned}$$

□

Proof of Proposition 14.4 We have $\tilde{V}(0+) = \tilde{U}(0+)$, so $\lim_{z \downarrow 0} \gamma(z) = \tilde{U}(0+) = U(\infty)$. Therefore, if $U(\infty) = \infty$, the infimum $\gamma(z)$ on $[0, \infty)$ cannot be attained at $z = 0$. Suppose now that $U(\infty) < \infty$ and that the infimum is attained at $\hat{z} = 0$, i.e. $\inf_{z > 0} \gamma(z) = \tilde{U}(0+)$. Then we have

$$x \geq \frac{\tilde{U}(0+) - \tilde{U}(zH)}{z} \geq E[HI(zH)]$$

for all $H \in \mathcal{H}$ and $z > 0$. Letting $z \rightarrow 0$ we get $x \geq \infty$, a contradiction. Therefore, either the infimum of $\gamma(z)$ is attained at a (unique) number $\hat{z} = \hat{z}_x \in (0, \infty)$ or it is attained at $\hat{z} = \infty$. If the latter is the case, then there exists a sequence $z_n \rightarrow \infty$ such that for z_n large enough and a fixed $z < z_n$, we have (by (14.9))

$$x \leq \frac{\tilde{V}(z) - \tilde{V}(z_n)}{z_n - z} \leq \frac{\tilde{V}(z) - \tilde{U}(z_n k)}{z_n - z}.$$

Letting $z_n \rightarrow \infty$ we get $x \leq 0$ by de l'Hôpital's Rule, a contradiction. □

Lemma 14.8

$$\tilde{V}'(\hat{z}) = -E \left[\frac{\hat{Z}}{B(T)} I \left(\hat{z} \frac{\hat{Z}}{B(T)} \right) \right] = -x.$$

Proof Let $h(z) := E[\tilde{U}(z\hat{Z}/B(T))]$. Then $h(\cdot)$ is convex, $h(\cdot) \geq \tilde{V}(\cdot)$ and $h(\hat{z}) = \tilde{V}(\hat{z})$. These three facts easily imply $\Delta^- h(\hat{z}) \leq \Delta^- \tilde{V}(\hat{z}) \leq \Delta^+ \tilde{V}(\hat{z}) \leq \Delta^+ h(\hat{z})$, where $\Delta^\pm$ denotes the left and the right derivatives. Because of this, it is sufficient to prove the lemma with $\tilde{V}$ replaced by h . It is easy to show, by the monotone

convergence theorem, that

$$\Delta^+ h(\hat{z}) \leq -E \left[\frac{\hat{Z}}{B(T)} I \left(\hat{z} \frac{\hat{Z}}{B(T)} \right) \right]. \quad (14.10)$$

On the other hand,

$$\Delta^- h(\hat{z}) \geq \limsup_{\varepsilon \rightarrow 0^+} E \left[-\frac{\hat{Z}}{B(T)} I \left((\hat{z} - \varepsilon) \frac{\hat{Z}}{B(T)} \right) \right].$$

We claim that

$$\begin{aligned} \frac{\hat{Z}}{B(T)} I \left((\hat{z} - \varepsilon) \frac{\hat{Z}}{B(T)} \right) &= \frac{\hat{Z}}{B(T)} I \left((\hat{z} - \varepsilon) \frac{\hat{Z}}{B(T)} \right) \mathbf{1}_{\{\hat{z} \frac{\hat{Z}}{B(T)} \geq z_0\}} \\ &\quad + \frac{\hat{Z}}{B(T)} I \left((\hat{z} - \varepsilon) \frac{\hat{Z}}{B(T)} \right) \mathbf{1}_{\{\hat{z} \frac{\hat{Z}}{B(T)} < z_0\}} \end{aligned}$$

is uniformly integrable when ε is small enough, where z_0 is the number from (14.7). Indeed, the first term is dominated by $(\hat{Z}/B(T))I((\hat{z} - \varepsilon)/\hat{z})z_0$, which is uniformly integrable when ε is sufficiently small since $E[\hat{Z}/B(T)] \leq k \cdot E[\hat{Z}] \leq k$. It follows from (14.7) that the second term is dominated by

$$\frac{1}{\hat{z} - \varepsilon} \frac{\gamma}{1 - \gamma} \tilde{U} \left((\hat{z} - \varepsilon) \frac{\hat{Z}}{B(T)} \right),$$

which is in turn dominated by

$$\frac{1}{\hat{z} - \varepsilon} \frac{\gamma c}{1 - \gamma} \tilde{U} \left(\frac{\hat{z} \hat{Z}}{B(T)} \right)$$

when ε is small. The uniform integrability follows from $E \left| \tilde{U} \left(\hat{z} \hat{Z}/B(T) \right) \right| < \infty$.

Therefore, we can use the mean convergence criterion to get the inequality

$$\Delta^- h(\hat{z}) \geq -E \left[\frac{\hat{Z}}{B(T)} I \left(\hat{z} \frac{\hat{Z}}{B(T)} \right) \right].$$

Together with (14.10) we establish $h'(\hat{z}) = -E[(\hat{Z}/B(T))I(\hat{z}\hat{Z}/B(T))] = -x$. The latter equality follows from the fact that $\hat{z}$ attains $\inf_{z>0} [\tilde{V}(z) + xz]$. $\square$

Lemma 14.9 *We have*

$$\sup_{z \in \mathcal{H}} E \left[\frac{Z}{B(T)} I \left(\hat{z} \frac{\hat{Z}}{B(T)} \right) \right] = E \left[\frac{\hat{Z}}{B(T)} I \left(\hat{z} \frac{\hat{Z}}{B(T)} \right) \right] = x.$$

Proof For a given $Z \in \mathcal{H}$, $\varepsilon \in (0, 1)$, let $Z_\varepsilon := (1 - \varepsilon)\hat{Z} + \varepsilon Z \in \mathcal{H}$. By optimality of $\hat{Z}$ we get

$$\begin{aligned} 0 &\geq \frac{1}{\varepsilon} E \left[\tilde{U} \left(\frac{\hat{z}\hat{Z}}{B(T)} \right) - \tilde{U} \left(\frac{\hat{z}Z_\varepsilon}{B(T)} \right) \right] \geq -\frac{1}{\varepsilon} E \left[\frac{\hat{z}(\hat{Z} - Z_\varepsilon)}{B(T)} I \left(\frac{\hat{z}Z_\varepsilon}{B(T)} \right) \right] \\ &= E \left[\frac{\hat{z}(Z - \hat{Z})}{B(T)} I \left(\frac{\hat{z}Z_\varepsilon}{B(T)} \right) \right]. \end{aligned} \quad (14.11)$$

However, it follows that, as in the proof of Lemma 14.8,

$$\left(\frac{Z - \hat{Z}}{B(T)} I \left(\frac{\hat{z}Z_\varepsilon}{B(T)} \right) \right)^- \leq \frac{\hat{Z}}{B(T)} I \left(\frac{\hat{z}(1 - \varepsilon)\hat{Z}}{B(T)} \right)$$

is uniformly integrable. We can now use Fatou's lemma in (14.11), to get

$$E \left[\frac{Z - \hat{Z}}{B(T)} I \left(\frac{\hat{z}\hat{Z}}{B(T)} \right) \right] \leq 0,$$

which completes the proof. $\square$

Proof of Theorem 14.5 For fixed $x > 0$ define

$$\mathcal{C} := \{\xi \in \mathbf{L}_+^0 / xB(T)\xi \leq X(T) + f(Y(T)), \text{ for some } (X(T), Y(T)) \in \mathcal{A}^+(x)\}.$$

Denote by

$$\mathcal{C}^0 := \{Z \in \mathbf{L}_+^0 / E[Z\xi] \leq 1, \forall \xi \in \mathcal{C}\}$$

the polar of set $\mathcal{C}$. It is clear then that $\mathcal{H} = \mathcal{C}^0$. We also want to show $\mathcal{C} = \mathcal{H}^0 = \mathcal{C}^{00}$. By the bipolar theorem of Brannath and Schachermayer (1998), it is sufficient to show that $\mathcal{C}$ is convex, solid and closed under a.s.-convergence (a subset $\mathcal{C}$ of $\mathbf{L}_+^0$ is solid if $f \in \mathcal{C}$ and $0 \leq g \leq f$ imply $g \in \mathcal{C}$). It is obvious that $\mathcal{C}$ is convex and solid. On the other hand, from Theorem 13.3 we know that $\xi \in \mathcal{C}$ if and only if

$$E_0^*[(\xi B(T))^2] < \infty \text{ and } \sup_{Z \in \mathcal{H}} E[Z\xi] \leq 1.$$

This implies (by Fatou's lemma) that $\mathcal{C}$ is closed under a.s.-convergence, because the set $\{\xi B(T)\}_{\xi \in \mathcal{C}}$ is bounded in $\mathbf{L}^2(\mathbf{P}_0^*)$. Indeed, the latter follows from [CK96] (as remarked in Appendix B of that paper, this can be shown by setting $U_n = V_n = 0$ in the arguments of its Appendix A; see (A.8)–(A.11) on p. 156). We conclude that $\mathcal{C} = \mathcal{H}^0$. Now, Lemma 14.9 implies $I(\hat{z}\hat{Z}/B(T))/(xB(T)) \in \mathcal{H}^0 = \mathcal{C}$, hence $(I(\hat{z}\hat{Z}/B(T)), 0) \in \mathcal{A}^+(x)$. This, in conjunction with Lemma 14.9 and Remark 14.2, implies the remaining statements of the theorem. $\square$

Notice that, if $r(\cdot)$ is deterministic, then Jensen's inequality gives

$$\begin{aligned} E \left[\tilde{U} \left(z \frac{Z}{B(T)} \right) \right] &\geq \tilde{U} \left(\frac{z}{B(T)} E[Z] \right) \\ &\geq \tilde{U} \left(\frac{z}{B(T)} \right), \end{aligned} \quad (14.12)$$

for all $Z \in \mathcal{H}$. We will use this observation to find examples in which the optimal strategy $(\hat{L}, \hat{M})$ never trades.

Example 14.10 Let us assume that $r(\cdot)$ is deterministic. In this case we see from (14.12) that

$$\tilde{V}(z) \geq \tilde{U}(z/B(T)),$$

and the infimum is attained by taking $\hat{Z} \equiv 1$, if $1 \in \mathcal{H}$. A sufficient condition for this is $(1, \hat{Z}_1(\cdot)) \in \mathcal{D}$ for some positive martingale $\hat{Z}_1(\cdot)$ such that $1 - \mu \leq \hat{R}(\cdot) = \hat{Z}_1(\cdot)/P(\cdot) \leq 1 + \lambda$. In particular, one can set $\hat{Z}_1(0) = (1 + \lambda)s$ and $\hat{Z}_1(\cdot) = Z_{\hat{\theta}_1}(\cdot)$, where $\hat{\theta}_1(\cdot) \equiv \sigma(\cdot)$, in which case $(1, \hat{Z}_1(\cdot)) \in \mathcal{D}$ if and only if

$$0 \leq \int_0^t (b(s) - r(s)) ds \leq \log \frac{1 + \lambda}{1 - \mu}, \quad \forall 0 \leq t \leq T. \quad (14.13)$$

Furthermore,

$$\hat{X}(T+) = I(\hat{z}/B(T)) = xB(T).$$

This means that the *no-trading strategy* $\hat{L} \equiv 0$, $\hat{M} \equiv 0$ is optimal. Condition (14.13) is satisfied, for instance, if

$$r(\cdot) \leq b(\cdot) \leq r(\cdot) + \rho, \text{ for some } 0 \leq \rho \leq \frac{1}{T} \log \frac{1 + \lambda}{1 - \mu}. \quad (14.14)$$

If $b(\cdot) = r(\cdot)$ the result is not surprising – even without transaction costs, it is then optimal not to trade. However, if there are no transaction costs, in the case $b(\cdot) > r(\cdot)$ the optimal portfolio always invests a positive amount in the stock; the same is true even in the presence of transaction costs, if one is maximizing expected discounted utility from consumption over an infinite time-horizon, and if the market coefficients are constant – see Shreve and Soner (1994), Theorem 11.6. The situation here, on the finite time-horizon $[0, T]$, is quite different: if the excess rate of return $b(\cdot) - r(\cdot)$ is positive but small relative to the transaction costs, and/or if the time-horizon is small, in the sense of (14.14), then it is optimal not to trade.

Acknowledgements

This chapter is adapted from my lecture notes ‘Optimal Trading Under Constraints’, which appeared in *Financial Mathematics*, W.J. Runggaldier (ed.), Lecture Notes in Mathematics 1656, Springer, 1997. Some material also appeared in Cvitanić (1997).

References

- Avellaneda, M. and Parás, A. (1994) Dynamic hedging portfolios for derivative securities in the presence of large transaction costs. *Applied Math. Finance* **1**, 165–94.
- Barles, G. and Soner, H.M. (1998) Option pricing with transaction costs and a nonlinear Black–Scholes equation. *Finance and Stochastics* **4**, 369–98.
- Bensaid, B., Lesne, J., Pagès, H. and Scheinkman, J. (1992) Derivative asset pricing with transaction costs. *Math. Finance* **2** (2), 63–86.
- Bergman, Y.Z. (1995) Option pricing with differential interest rates. *Rev. Financial Studies* **8**, 475–500.
- Bismut, J.M. (1973) Conjugate convex functions in optimal stochastic control. *J. Math. Analysis and Applic.* **44**, 384–404.
- Bismut, J.M. (1975) Growth and optimal intertemporal allocations of risks. *J. Econ. Theory* **10**, 239–87.
- Black, F. and Scholes, M. (1973), The pricing of options and corporate liabilities. *J. Polit. Economy* **81**, 637–59.
- Boyle, P.P. and Vorst, T. (1992), Option replication in discrete time with transaction costs. *J. Finance* **47**, 272–93.
- Brannath, W. and Schachermayer, W. (1999), A bipolar theorem for subsets of $L^0_+(\Omega, \mathcal{F}, P)$. *Séminaire de Probabilités XXXIII*, 344–54.
- Broadie, M., Cvitanić, J. and Soner, H.M. (1998), On the cost of super-replication under portfolio constraints. *Rev. Financial Studies* **11**, 59–79.
- Constantinides, G.M. (1979), Multiperiod consumption and investment behavior with convex transaction costs. *Management Sci.* **25**, 1127–37.
- Constantinides, G.M. and Zariphopoulou, T. (1999), Bounds on prices of contingent claims in an intertemporal economy with proportional transaction costs and general preferences. *Finance and Stochastics* **3**, 345–70.
- Cox, J. and Huang, C.F. (1989), Optimal consumption and portfolio policies when asset prices follow a diffusion process. *J. Econ. Theory* **49**, 33–83.
- Cox, J. and Huang, C.F. (1991), A variational problem arising in financial economics. *J. Math. Economics* **20**, 465–87.
- Cuoco, D and Cvitanić, J. (1998), Optimal consumption choices for a large investor. *J. Econ. Dynamics and Control* **22**, 401–36.
- Cvitanić, J. (1997), Nonlinear financial markets: hedging and portfolio optimization. In *Mathematics of Derivative Securities*, M.H.A. Dempster and S. Pliska, eds., *Proc. of the Isaac Newton Institute*, Cambridge University Press.
- Cvitanić, J. and Karatzas, I. (1992), Convex duality in constrained portfolio optimization. *Ann. Appl. Probab.* **2**, 767–818.
- Cvitanić, J. and Karatzas, I. (1993), Hedging contingent claims with constrained portfolios. *Ann. Appl. Probab.* **3**, 652–81.
- Cvitanić, J. and Karatzas, I. (1996), Hedging and portfolio optimization under transaction costs: a martingale approach. *Mathematical Finance* **6**, 133–65.

- Cvitanic, J., Pham H. and Touzi N. (1998), A closed form solution to the problem of super-replication under transaction costs. *Finance and Stochastics* **3**, 35–54.
- Cvitanic, J. and Wang, H. (1999), On optimal terminal wealth under transaction costs. *J. Math. Economics*, to appear.
- Davis, M.H.A. (1997), Option pricing in incomplete markets. In *Mathematics of Derivative Securities*, M.A.H. Dempster and S. Pliska, eds., *Proc. of the Isaac Newton Institute*, Cambridge University Press.
- Davis, M.H.A. and Clark, J.M.C. (1994), A note on super-replicating strategies. *Phil. Trans. Royal Soc. London A* **347**, 485–94.
- Davis, M.H.A. and Norman, A. (1990), Portfolio selection with transaction costs. *Math. Operations Research* **15**, 676–713.
- Davis, M.H.A. and Panas, V.G. (1994), The writing price of a European contingent claim under proportional transaction costs. *Comp. Appl. Math.* **13**, 115–57.
- Davis, M.H.A., Panas, V.G. and Zariphopoulou, T. (1993), European option pricing with transaction costs. *SIAM J. Control and Optimization* **31**, 470–93.
- Davis, M.H.A. and Zariphopoulou, T. (1995), American options and transaction fees. In *Mathematical Finance*, M.H.A. Davis *et al.*, eds., *The IMA Volumes in Mathematics and its Applications* **65**, 47–62. Springer-Verlag.
- Edirisinghe, C., Naik, V. and Uppal, R. (1993), Optimal replication of options with transaction costs and trading restrictions. *J. Financial and Quantitative Analysis* **28**, 117–38.
- Ekeland, I. and Temam, R. (1976), *Convex Analysis and Variational Problems*. North-Holland, Amsterdam and Elsevier, New York.
- El Karoui, N., Peng, S. and Quenez, M.C. (1997), Backward stochastic differential equations in finance. *Math. Finance* **7**, 1–71.
- El Karoui, N. and Quenez, M.C. (1995), Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM J. Control and Optimization*, **33**, 29–66.
- Fleming, W.H. and Rishel, R.W. (1975), *Deterministic and Stochastic Optimal Control*. Springer-Verlag, New York.
- Fleming, W.H. and Soner, H.M. (1993), *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, New York.
- Fleming, W. and Zariphopoulou, T. (1991), An optimal investment/consumption model with borrowing. *Math. Oper. Res.* **16**, 802–22.
- Flesaker, B. and Hughston, L.P. (1994), Contingent claim replication in continuous time with transaction costs. *Proc. Derivative Securities Conference*, Cornell University.
- Foldes, L. (1978a) Martingale conditions for optimal saving – discrete time. *J. Math. Economics* **5**, 83–96.
- Foldes, L. (1978b) Optimal saving and risk in continuous time. *Rev. Economic Studies* **45**, 39–65.
- Föllmer, H. and Kramkov, D. (1997), Optional decomposition under constraints. *Prob. Theory and Related Fields* **109**, 1–25.
- Gilster, J.E. and Lee, W. (1984), The effect of transaction costs and different borrowing and lending rates on the option pricing model. *J. Finance* **43**, 1215–21.
- Grannan, E.R. and Swindle, G.H. (1996), Minimizing transaction costs of option hedging strategies. *Math. Finance* **6**, 239–53.
- Harrison, J.M. and Kreps, D.M. (1979), Martingales and arbitrage in multiperiod security markets. *J. Econ. Theory* **20**, 381–408.
- Harrison, J.M. and Pliska, S.R. (1981), Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Appl.* **11**, 215–260.
- Harrison, J.M. and Pliska, S.R. (1983), A stochastic calculus model of continuous time

- trading: complete markets. *Stochastic Processes and Appl.* **15**, 313–316.
- He, H. and Pearson, N. (1991), Consumption and portfolio policies with incomplete markets and short-sale constraints: the infinite-dimensional case. *J. Econ. Theory* **54**, 259–304.
- Hodges, S.D. and Neuberger, A. (1989), Optimal replication of contingent claims under transaction costs. *Review of Future Markets* **8**, 222–39.
- Hoggard, T., Whalley, A.E. and Wilmott, P. (1994), Hedging option portfolios in the presence of transaction costs. *Adv. in Futures and Options Research*, **7**, 21–35.
- Jouini, E. and Kallal, H. (1995a) Arbitrage in securities markets with short-sale constraints. *Math. Finance* **5**, 197–232.
- Jouini, E. and Kallal, H. (1995b) Martingales and arbitrage in securities markets with transaction costs. *J. Econ. Theory* **66**, 178–97.
- Kabanov, Yu.M. (1999), Hedging and liquidation under transaction costs in currency markets. *Finance and Stochastics* **3**, 237–48.
- Karatzas, I. and Kou, S-G. (1996), On the pricing of contingent claims under constraints. *Ann. Appl. Probab.*, **6**, 321–69.
- Karatzas, I., Lehoczky, J.P. and Shreve, S.E. (1987), Optimal portfolio and consumption decisions for a “small investor” on a finite horizon. *SIAM J. Control Optimization* **25**, 1557–86.
- Karatzas, I., Lehoczky, J.P., Shreve, S.E. and Xu, G.L. (1991), Martingale and duality methods for utility maximization in an incomplete market. *SIAM J. Control Optimization* **29**, 702–30.
- Karatzas, I. and Shreve, S.E. (1991), *Brownian Motion and Stochastic Calculus* (2nd edition), Springer-Verlag, New York.
- Karatzas, I. and Shreve, S.E. (1998), *Methods of Mathematical Finance*. Springer-Verlag, New York.
- Komlós, J. (1967), A generalization of a problem of Steinhaus. *Acta Math. Acad. Sci. Hungar.* **18**, 217–29.
- Korn, R. (1997), *Optimal Portfolios: Stochastic Models for Optimal Investment and Risk Management in Continuous Time*. World Scientific, Singapore.
- Kramkov, D. and Schachermayer, W. (1998), The asymptotic elasticity of utility functions and optimal investment in incomplete markets. *The Annals of Applied Probability* **9**.
- Kusuoka, S. (1995), Limit theorem on option replication with transaction costs. *Ann. Appl. Probab.* **5**, 198–221.
- Ladyženskaja, O.A., Solonnikov, V.A. and Ural'tseva, N.N. (1968), *Linear and Quasilinear Equations of Parabolic Type*. Translations of Mathematical Monographs, Vol. 23, American Math. Society, Providence, R.I.
- Leland, H.E. (1985), Option pricing and replication with transaction costs. *J. Finance* **40**, 1283–301.
- Levental, S. and Skorohod, A.V. (1997), On the possibility of hedging options in the presence of transactions costs. *Ann. Appl. Probab.* **7**, 410–43.
- Magill, M.J.P. and Constantinides, G.M. (1976), Portfolio selection with transaction costs. *J. Economic Theory* **13**, 264–71.
- Merton, R.C. (1969), Lifetime portfolio selection under uncertainty: the continuous-time case. *Rev. Econ. Statist.*, **51**, 247–57.
- Merton, R.C. (1971), Optimum consumption and portfolio rules in a continuous-time model. *J. Econom. Theory* **3**, 373–413. Erratum: *ibid* **6** (1973), 213–4.
- Merton, R.C. (1989), On the application of the continuous time theory of finance to financial intermediation and insurance, *The Geneva Papers on Risk and Insurance*, 225–261.

- Merton, R.C. (1990), *Continuous-Time Finance*. Basil Blackwell, Oxford and Cambridge.
- Morton, A.J. and Pliska, S.R. (1995), Optimal portfolio management with fixed transaction costs, *Math. Finance* **5**, 337–56.
- Neveu, J. (1975), *Discrete-Parameter Martingales*. North-Holland, Amsterdam.
- Pliska, S. (1986), A stochastic calculus model of continuous trading: optimal portfolios. *Math. Oper. Res.* **11**, 371–82.
- Pliska, S. (1997), *Introduction to Mathematical Finance. Discrete Time Models*. Blackwell, Oxford.
- Rockafellar, R.T. (1970), *Convex Analysis*. Princeton University Press, Princeton.
- Schwartz, M. (1986), New proofs of a theorem of Komlós. *Acta Math. Hung.* **47**, 181–5.
- Shreve, S.E. and Soner, H.M. (1994), Optimal investment and consumption with transaction costs, *Ann. Appl. Probab.* **4**, 609–92.
- Soner, H.M., Shreve, S.E. and Cvitanić, J. (1995), There is no nontrivial hedging portfolio for option pricing with transaction costs, *Ann. Appl. Probab.* **5**, 327–55.
- Taksar, M., Klass, M.J. and Assaf, D. (1988), A diffusion model for optimal portfolio selection in the presence of brokerage fees, *Math. Operations Research* **13**, 277–94.
- Xu, G.L. (1990), *A Duality Method for Optimal Consumption and Investment under Short-Selling Prohibition*. Doctoral Dissertation, Carnegie-Mellon University.
- Xu, G. and Shreve, S.E. (1992), A duality method for optimal consumption and investment under short-selling prohibition. I. General market coefficients. II. Constant market coefficients. *Ann. Appl. Probab.* **2**, 87–112, 314–28.
- Zariphopoulou, T. (1992), Investment/consumption model with transaction costs and Markov-chain parameters. *SIAM J. Control Optimization* **30**, 613–36.

Bayesian Adaptive Portfolio Optimization

Ioannis Karatzas and Xiaoliang Zhao

1 Introduction

This chapter is a contribution to the study of portfolio optimization problems in stochastic control and mathematical finance. Starting with initial capital $x_0 > 0$, an investor tries to maximize his expected utility from terminal wealth, by choosing portfolio strategies based on information about asset-prices in a financial market. The investor cannot observe directly the stock appreciation rates or the driving Brownian motion; he can only observe past and present stock-prices. We adopt a *Bayesian* approach, by assuming that the unknown “drift” (i.e., vector of stock appreciation rates) is an unobservable random variable, independent of the driving Brownian motion and with known probability distribution. We refer to this as the case of *partial observations*, in order to distinguish it from the case of *complete observations*, on which a large literature already exists.

The original utility maximization problems were introduced by Merton (1971) in the context of constant coefficients, and were treated by the Markovian methods of *continuous-time stochastic control*; see also Fleming & Rishel (1975), pp. 159–65, Fleming & Soner (1993), Chapter 4, and Karatzas & Shreve (1998), pp. 118–36. For general parameter-processes and complete markets, methodologies based on martingale theory and convex duality were developed by Pliska (1986), by Karatzas, Lehoczky & Shreve (1987), and by Cox & Huang (1989); they were extended to the setting of incomplete and/or general constrained markets by Karatzas, Lehoczky, Shreve & Xu (1991), He & Pearson (1991) and Cvitanić & Karatzas (1992). Chapters 3 and 6 of the monograph by Karatzas & Shreve (1998) contain a comprehensive account and overview of these developments.

Models with *partial observations* were studied by Detemple (1986), Dothan & Feldman (1986) and Gennotte (1986) in a linear Gaussian filtering setting. Karatzas & Xue (1991) introduced a Bayesian approach for the utility maximization problems, using filtering and martingale representation theory. Within this

framework, Lakner (1995, 1998) and Zohar (1999) solved the optimization problems via the martingale approach, Kuwana (1995) studied necessary and sufficient conditions for the *certainty-equivalence principle* to hold, and Karatzas (1997) studied the problem of maximizing the probability of reaching a given “goal” during some finite time-horizon. For an unobservable drift process driven by an independent Brownian motion, the optimization problem was studied by Rishel (1999) for utility functions of power-type. The special case of logarithmic utility function and normal prior distribution was studied by Browne & Whitt (1996) on an infinite horizon.

In this chapter we first use results from filtering theory, to reduce the optimization problem with *partial observations* to the case of a drift process which is adapted to the observation process; this way the well-developed martingale methods can be applied (Sections 2 and 3). We obtain explicit formulae for the *optimal portfolio process*, the *optimal wealth process* and the *value function* of the stochastic control problem. In Section 4, we use the standard framework of *stochastic control* and *dynamic programming* to treat this problem again, which leads us to generalized *parabolic Monge–Ampère-type equations*. Using the results of Sections 2 and 3, we solve these equations explicitly. In Section 5 we study the optimization problem for an “insider” investor who can observe both the drift vector and the driving Brownian motion. We compute in this framework the relative cost for the uncertainty associated with the prior distribution; for logarithmic utility functions, we show that this relative cost is asymptotically negligible as $T \rightarrow \infty$. We conclude in Sections 6 and 7 with a discussion of optimal strategies and value functions under *convex constraints on portfolio-proportions*, in the manner of Cvitanić & Karatzas (1992); such constraints include incomplete markets, prohibition or constraints on the short-selling of stocks, prohibition or constraints on borrowing, etcetera.

2 Formulation and financial interpretation

Let us start with a given complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and on it

- (i) an $\mathbb{R}^d$ -valued Brownian motion $W(\cdot) = \{W(t), \mathcal{F}^W(t); 0 \leq t < \infty\}$, as well as
- (ii) a random variable $\Theta : \Omega \rightarrow \mathbb{R}^d$, independent of the process $W(\cdot)$ under the probability measure $\mathbf{P}$, and with known distribution $\mu(A) = \mathbf{P}[\Theta \in A]$, $A \in \mathcal{B}(\mathbb{R}^d)$ that satisfies

$$\int_{\mathbb{R}^d} \|\vartheta\| \mu(d\vartheta) < \infty. \quad (2.1)$$

We shall denote by

$$Y(t) \triangleq W(t) + \Theta t, \quad 0 \leq t < \infty \quad (2.2)$$

the $\mathbf{P}$ -Brownian motion with drift Θ , by $\mathbf{F} = \{\mathcal{F}(t); 0 \leq t < \infty\}$ the $\mathbf{P}$ -augmentation of

$$\mathcal{F}^Y(t) \triangleq \sigma(Y(s); 0 \leq s \leq t), \quad (2.3)$$

the filtration generated by the process $Y(\cdot)$, and by $\mathbf{G} = \{\mathcal{G}(t); 0 \leq t < \infty\}$ the augmentation of the auxiliary, enlarged filtration

$$\mathcal{G}^{\Theta, W}(t) \triangleq \sigma(\Theta, W(s); 0 \leq s \leq t) = \sigma(\Theta) \vee \mathcal{F}^W(t) \quad (2.4)$$

generated by *both* the process $W(\cdot)$ and the random variable Θ . Clearly, $\mathcal{F}(t) \subseteq \mathcal{G}(t)$ for every $0 \leq t < \infty$.

Lemma 2.1 *$W(\cdot)$ is a $(\mathbf{G}, \mathbf{P})$ -Brownian motion, and the exponential process*

$$\begin{aligned} \Lambda(t) &\equiv \frac{1}{Z(t)} \triangleq \exp\left(-\Theta^* W(t) - \frac{1}{2}\|\Theta\|^2 t\right) = \exp\left(-\Theta^* Y(t) + \frac{1}{2}\|\Theta\|^2 t\right), \\ 0 \leq t < \infty \end{aligned} \quad (2.5)$$

is a $(\mathbf{G}, \mathbf{P})$ -martingale.

Thus, for any given $T \in (0, \infty)$, we can define

$$\tilde{\mathbf{P}}_T(A) \triangleq \mathbf{E}[\Lambda(T) \cdot 1_A], \quad A \in \mathcal{G}(T), \quad (2.6)$$

a probability measure equivalent to $\mathbf{P}$ on $\mathcal{G}(T)$.

Lemma 2.2 *Under the probability measure $\tilde{\mathbf{P}}_T$ of (2.6), the process*

$$Y(t) = W(t) + \Theta t, \quad 0 \leq t \leq T$$

is standard d -dimensional Brownian motion with respect to $\mathbf{G}$ (thus also with respect to $\mathbf{F}$) and is independent of the random variable Θ , whereas the exponential process

$$Z(t) = \exp\left[\Theta^* Y(t) - \frac{1}{2}\|\Theta\|^2 t\right], \quad 0 \leq t \leq T$$

is a martingale with respect to $\mathbf{G}$. Furthermore, we have

$$\mathbf{P}[\Theta \in A] = \tilde{\mathbf{P}}_T[\Theta \in A] = \mu(A), \quad \forall A \in \mathcal{B}(\mathbb{R}^d).$$

The proofs of Lemma 2.1 and Lemma 2.2 are deferred to the Appendix.

For a given initial position $x_0 > 0$, a constant $r \geq 0$, an invertible $(d \times d)$ -matrix $\sigma = \{\sigma_{ij}\}_{1 \leq i, j \leq d}$, and a given time-horizon $[0, T]$ with $T \in (0, \infty)$, consider the

space $\mathcal{A}(x_0) \equiv \mathcal{A}(x_0; 0, T)$ of $\mathbf{F}$ -progressively measurable processes $\pi : [0, T] \times \Omega \rightarrow \Re^d$ which satisfy

$$\int_0^T e^{-2rt} \|\pi(t)\|^2 dt < \infty, \quad (2.7)$$

$$0 \leq e^{-rt} X^{x_0, \pi}(t) \triangleq x_0 + \int_0^t e^{-rs} \pi^*(s) \sigma dY(s), \quad \forall 0 \leq t \leq T \quad (2.8)$$

$\mathbf{P}$ -almost surely. This is the class of our admissible control processes for the initial position x_0 .

Definition 2.3 A function $u : (0, \infty) \rightarrow \Re$ will be called a *utility function* if it is strictly increasing, strictly concave, of class C^2 , and satisfies

$$u'(0+) \triangleq \lim_{x \downarrow 0} u'(x) = \infty, \quad u'(\infty) \triangleq \lim_{x \rightarrow \infty} u'(x) = 0. \quad (2.9)$$

We can now state the stochastic control problem we are interested in, as follows.

Problem 2.4 For a given utility function $u(\cdot)$, initial position x_0 and finite time-horizon $[0, T]$, maximize the expected utility from $X(\cdot)$ of (2.8) at the terminal time T , over the class $\mathcal{A}(x_0)$. The value function of this problem will be denoted by

$$V(x_0) \triangleq \sup_{\pi(\cdot) \in \mathcal{A}(x_0)} \mathbf{E}u(X^{x_0, \pi}(T)). \quad (2.10)$$

Remark 2.5 We want to emphasize the *financial interpretation* of Problem 2.1. Suppose that a financial market $\mathcal{M}$ has one riskless asset (money market) with constant interest-rate $r \geq 0$ and price $S_0(t) = e^{-rt}$, as well as d risky assets (stocks). Assume that the prices-per-share $S(\cdot) = \{(S_1(t), \dots, S_d(t))^*; 0 \leq t < \infty\}$ of these risky assets are modelled by the equations

$$dS_i(t) = S_i(t) \left[B_i dt + \sum_{j=1}^d \sigma_{ij} dW_j(t) \right] = S_i(t) \left[r dt + \sum_{j=1}^d \sigma_{ij} dY_j(t) \right],$$

$$S_i(0) > 0, \quad i = 1, \dots, d.$$

Here $W(\cdot)$ is the driving Brownian motion under the probability measure $\mathbf{P}$, and

$$B = (B_1, \dots, B_d), \quad B_i \triangleq r + (\sigma \Theta)_i, \quad \text{for } i = 1, \dots, d$$

is the vector of “stock appreciation rates”. These unobservable rates are modelled by means of a random vector $\Theta \equiv \sigma^{-1}[B - r \cdot (1, \dots, 1)^*]$ which represents the “market-price of risk”; this random vector is independent of the Brownian motion $W(\cdot)$, and has a known distribution μ . We assume that we cannot observe either B

(equivalently, Θ) or $W(\cdot)$ directly, but that we *can* observe the stock-price process $S(\cdot)$. In other words, this process $S(\cdot)$ generates the “observation filtration” $\mathbf{F} = \{\mathcal{F}(t); 0 \leq t < \infty\}$, which coincides with the $\mathbf{P}$ -augmentation of the filtration

$$\mathcal{F}^Y(t) \triangleq \sigma(Y(u); 0 \leq u \leq t) = \sigma(S(u); 0 \leq u \leq t).$$

A small investor with initial capital $x_0 > 0$ and finite time-horizon $[0, T]$ chooses his “portfolio” $\pi(t) = (\pi_1(t), \dots, \pi_d(t))^*$ at time t based on the information $\mathcal{F}(t)$ from past and present stock-prices observed up to that time; here $\pi_i(t)$ represents the amount of money invested in the i th stock at time t . Thus, the wealth process $X(\cdot) \equiv X^{x_0, \pi}(\cdot)$ of this investor satisfies the linear stochastic differential equation

$$\begin{aligned} dX(t) &= \sum_{i=1}^d \pi_i(t) \cdot \frac{dS_i(t)}{S_i(t)} + \left(X(t) - \sum_{i=1}^d \pi_i(t) \right) \cdot \frac{dS_0(t)}{S_0(t)} \\ &= rX(t)dt + \pi^*(t)\sigma dY(t), \quad X(0) = x_0, \end{aligned} \quad (2.11)$$

on $[0, T]$, whose solution is given by $X^{x_0, \pi}(\cdot)$ of (2.8). We emphasize that a trading strategy $\pi(\cdot)$ is required to be $\mathbf{F}$ -adapted; in other words, investors indeed observe the security prices only, not the stock appreciation rates B or the driving Brownian motion $W(\cdot)$. For a given utility function $u(\cdot)$, the investor’s objective is to maximize his expected utility of wealth at the terminal time T . Now we are exactly in the setting of Problem 2.1.

Remark 2.6 More generally, the financial market model may allow for random, time-varying interest rate $r(\cdot)$ and volatility $\sigma(\cdot)$, that is,

$$dS_0(t) = S_0(t)r(t)dt, \quad S_0(0) = 1$$

for the riskless asset, and

$$dS_i(t) = S_i(t) \left[B_i dt + \sum_{j=1}^d \sigma_{ij}(t) dW_j(t) \right], \quad i = 1, \dots, d$$

for the prices-per-share of the risky assets. Here $\sigma(\cdot) = (\sigma_{ij}(\cdot))_{1 \leq i, j \leq d}$ is a bounded, $\mathbf{F}$ -progressively measurable process with values in the space of $(d \times d)$ -matrices with full-rank and bounded inverse, and $r(\cdot)$ is a measurable, $\mathbf{F}$ -adapted scalar process with $\int_0^T r(t)dt < \infty$ almost surely. One of the main results of this chapter, Theorem 3.1, can be easily extended to such a setting, provided that $\sigma(\cdot)$ is a smooth function of past and present stock-prices; more precisely, of the form $\sigma_{ij}(t) = \Sigma_{i,j}(t, S(\cdot))$, $0 \leq t \leq T$, $1 \leq i, j \leq d$ where $\Sigma_{i,j} : [0, T] \times C([0, T]; \mathbb{R}^d) \rightarrow \mathbb{R}$ is progressively measurable and Lipschitz continuous in the sup-norm on $C([0, T]; \mathbb{R}^d)$ (see Karatzas & Shreve (1991), Definition 3.5.15 and pp. 302–11).

Remark 2.7 Notice from (2.2) the *consistency property*

$$\lim_{t \rightarrow \infty} \frac{Y(t)}{t} = \Theta, \quad \mathbf{P}\text{-a.s.} \quad (2.12)$$

for the *maximum likelihood estimator* $(Y(t)/t)$ of Θ on $[0, T]$, given the observations $Y(s)$, $0 \leq s \leq t$. In particular, Θ is measurable with respect to the $\mathbf{P}$ -completion of the σ -algebra $\mathcal{F}(\infty) \triangleq \sigma\left(\bigcup_{0 \leq t < \infty} \mathcal{F}(t)\right)$.

3 Filtering and martingale methods

In this section we shall use the well-developed martingale methodology (e.g. Karatzas & Shreve (1998), Chapter 3), along with elementary filtering theory, to solve the optimization Problem 2.1. Let us start by introducing the $(\mathbf{F}, \tilde{\mathbf{P}}_T)$ -martingale

$$\begin{aligned} \hat{Z}(t) &\triangleq \tilde{\mathbf{E}}_T \left[\frac{d\mathbf{P}}{d\tilde{\mathbf{P}}_T} \middle| \mathcal{F}(t) \right] = \tilde{\mathbf{E}}_T [Z(T) | \mathcal{F}(t)] \\ &= \tilde{\mathbf{E}}_T [\tilde{\mathbf{E}}_T [Z(T) | \mathcal{G}(t)] | \mathcal{F}(t)] = \tilde{\mathbf{E}}_T [Z(t) | \mathcal{F}(t)] \\ &= \begin{cases} F(t, Y(t)); & 0 < t \leq T \\ 1; & t = 0 \end{cases} \end{aligned} \quad (3.1)$$

from Lemma 2.2, where

$$F(t, y) \triangleq \int_{\mathfrak{N}^d} \exp \left(\vartheta^* y - \frac{1}{2} \|\vartheta\|^2 t \right) \mu(d\vartheta), \quad (t, y) \in (0, \infty) \times \mathfrak{N}^d. \quad (3.2)$$

This function satisfies the backwards heat-equation $F_t + \frac{1}{2} \Delta F = 0$ (see Remark 4.1 on notation). At any given time $t \in [0, \infty)$, the “posterior distribution of Θ under $\mathbf{P}$, given the observations $\mathcal{F}(t)$ up to that time”, is given by the familiar Bayes rule of Lemma 3.5.3 in Karatzas & Shreve (1991), i.e.,

$$\mu_t(A) \triangleq \mathbf{P}[\Theta \in A | \mathcal{F}(t)] = \frac{\nu_t(A)}{\nu_t(\mathfrak{N}^d)}, \quad A \in \mathcal{B}(\mathfrak{N}^d), \quad (3.3)$$

in the terms of the random measure

$$\begin{aligned} \nu_t(A) &\triangleq \tilde{\mathbf{E}}_T [\mathbf{1}_A(\Theta) Z(T) | \mathcal{F}(t)] \\ &= \tilde{\mathbf{E}}_T [\mathbf{1}_A(\Theta) \tilde{\mathbf{E}}_T [Z(T) | \mathcal{G}(t)] | \mathcal{F}(t)] = \tilde{\mathbf{E}}_T [\mathbf{1}_A(\Theta) Z(t) | \mathcal{F}(t)] \\ &= \tilde{\mathbf{E}}_T [\mathbf{1}_A(\Theta) \exp(\Theta^* Y(t) - \|\Theta\|^2 t/2) | \mathcal{F}(t)] \\ &= \begin{cases} \int_A \exp(\vartheta^* y - \|\vartheta\|^2 t/2) \mu(d\vartheta) \big|_{y=Y(t)}; & 0 < t < \infty \\ \tilde{\mathbf{P}}_T[\Theta \in A] = \mu(A); & t = 0 \end{cases} \end{aligned} \quad (3.4)$$

with $t \leq T < \infty$. Clearly, $v_t(\mathfrak{N}^d) = \hat{Z}(t) = F(t, Y(t))$ for $t > 0$. The mean-vector of the conditional distribution $\mu_t(\cdot)$ in (3.3) is the $(\mathbf{F}, \mathbf{P})$ -martingale

$$\hat{\Theta}(t) \triangleq \int_{\mathfrak{N}^d} \vartheta \mu_t(d\vartheta) = \mathbf{E}[\Theta | \mathcal{F}(t)] = \begin{cases} G(t, Y(t)); & 0 < t < \infty \\ \int_{\mathfrak{N}^d} \vartheta \mu(d\vartheta); & t = 0 \end{cases}, \quad (3.5)$$

where we have set

$$G(t, y) \triangleq \left(\frac{\nabla F}{F} \right) (t, y), \quad (t, y) \in (0, \infty) \times \mathfrak{N}^d. \quad (3.6)$$

The random vector $\hat{\Theta}(t)$ is the *Bayes estimator* of Θ on the interval $[0, t]$ with respect to the prior distribution μ , given the observations $Y(s)$, $0 \leq s \leq t$. Now it is easy to check that the process

$$N(t) \triangleq Y(t) - \int_0^t \hat{\Theta}(s) ds = Y(t) - \int_0^t G(s, Y(s)) ds, \quad 0 \leq t < \infty \quad (3.7)$$

is an $(\mathbf{F}, \mathbf{P})$ -Brownian motion, the so-called *innovations process* of filtering theory (see Kallianpur (1980), Elliott (1982), Chapter 18, or Rogers & Williams (1987), pp. 322–9). On the other hand, from the Lévy martingale convergence theorem and in conjunction with Remark 2.3, we obtain the *consistency property for the Bayes estimator* $\hat{\Theta}(\cdot)$ of (3.5),

$$\lim_{t \rightarrow \infty} \hat{\Theta}(t) = \Theta, \quad \mathbf{P}\text{-a.s.} \quad (3.8)$$

An application of Itô's rule to the process $\hat{Z}(\cdot)$ of (3.1) and to its reciprocal $\hat{\Lambda}(\cdot) \triangleq 1/\hat{Z}(\cdot)$ gives

$$d\hat{Z}(t) = \hat{Z}(t)\hat{\Theta}^*(t)dY(t), \quad \hat{Z}(0) = 1 \quad (3.9)$$

$$d\hat{\Lambda}(t) = -\hat{\Lambda}(t)\hat{\Theta}^*(t)dN(t), \quad \hat{\Lambda}(0) = 1 \quad (3.10)$$

as well as

$$\begin{aligned} d(\hat{\Lambda}(t) \cdot e^{-rt} X^{x_0, \pi}(t)) &= \hat{\Lambda}(t)d(e^{-rt} X^{x_0, \pi}(t)) + e^{-rt} X^{x_0, \pi}(t)d\hat{\Lambda}(t) + d\langle e^{-rt} X^{x_0, \pi}, \hat{\Lambda} \rangle(t) \\ &= e^{-rt} [\hat{\Lambda}(t)\pi^*(t)\sigma dY(t) - \hat{\Lambda}(t)X^{x_0, \pi}(t)\hat{\Theta}^*(t)dN(t) - \hat{\Lambda}(t)\pi^*(t)\sigma \hat{\Theta}(t)dt] \\ &= e^{-rt} \hat{\Lambda}(t) [\sigma^* \pi(t) - X^{x_0, \pi}(t)\hat{B}(t)]^* dN(t). \end{aligned} \quad (3.11)$$

This shows that, on a given finite time-horizon $[0, T]$ and every $\pi(\cdot) \in \mathcal{A}(x_0)$, the process $e^{-r\cdot} \hat{\Lambda}(\cdot) X^{x_0, \pi}(\cdot)$ is a nonnegative $(\mathbf{F}, \mathbf{P})$ -local martingale, hence also a supermartingale; in particular

$$e^{-rT} \cdot \mathbf{E}[\hat{\Lambda}(T) X^{x_0, \pi}(T)] \leq x_0, \quad \forall \pi(\cdot) \in \mathcal{A}(x_0). \quad (3.12)$$

We can now use *convex duality* methods, to maximize the expected utility $\mathbf{E}u(X^{x_0, \pi}(T))$ of (2.10) subject to the constraint (3.12), as follows. Let us introduce the monotone decreasing function $I(\cdot)$ as the *inverse* of the marginal utility function $u'(\cdot)$, and the convex dual

$$\tilde{u}(k) \triangleq \max_{x>0} [u(x) - xk] = u(I(k)) - kI(k), \quad k > 0 \quad (3.13)$$

of $u(\cdot)$. From (3.12) and (3.13), we obtain

$$\begin{aligned} \mathbf{E}u(X^{x_0, \pi}(T)) &\leq \mathbf{E}[\tilde{u}(ke^{-rT} \hat{\Lambda}(T))] + ke^{-rT} \cdot \mathbf{E}[\hat{\Lambda}(T)X^{x_0, \pi}(T)] \\ &\leq \mathbf{E}[\tilde{u}(ke^{-rT} \hat{\Lambda}(T))] + x_0k \end{aligned} \quad (3.14)$$

for every $k > 0$, $\pi(\cdot) \in \mathcal{A}(x_0)$. Furthermore, (3.14) is valid as equality, if and only if both

$$X^{x_0, \pi}(T) = I(ke^{-rT} \hat{\Lambda}(T)), \quad \text{a.s.} \quad (3.15)$$

$$\mathbf{E}[\hat{\Lambda}(T)X^{x_0, \pi}(T)] = \tilde{\mathbf{E}}_T \left[I \left(\frac{ke^{-rT}}{F(T, Y(T))} \right) \right] = x_0e^{rT} \quad (3.16)$$

hold.

Assumption 3.1 Suppose that the function

$$L(k; s, y) \triangleq \left\{ \begin{array}{ll} e^{-rs} \int_{\mathbb{R}^d} I \left(\frac{ke^{-rT}}{F(T, y+z)} \right) \varphi_s(z) dz; & k > 0, s > 0, y \in \mathbb{R}^d \\ I \left(\frac{ke^{-rT}}{F(T, y)} \right); & k > 0, s = 0, y \in \mathbb{R}^d \end{array} \right\} \quad (3.17)$$

is finite for every $(k, s, y) \in (0, \infty) \times [0, T] \times \mathbb{R}^d$. We are using the notation

$$\varphi_s(z) \triangleq (2\pi s)^{-d/2} \cdot e^{-\|z\|^2/2s}; \quad z \in \mathbb{R}^d, s > 0 \quad (3.18)$$

for the Gaussian density function, and assume that $L(k; s, y)$ has finite first derivatives with respect to the arguments s, k and y . We also assume (for the results of Section 4) that $L(k; s, y)$ has finite second derivatives with respect to the arguments k and y on $(0, \infty) \times (0, T) \times \mathbb{R}^d$.

Under this assumption, the strictly decreasing function

$$\begin{aligned} k &\longmapsto e^{-rT} \cdot \tilde{\mathbf{E}}_T \left[I \left(\frac{ke^{-rT}}{F(T, Y(T))} \right) \right] \\ &= e^{-rT} \int_{\mathbb{R}^d} I \left(\frac{ke^{-rT}}{F(T, z)} \right) \varphi_T(z) dz = L(k; T, 0) \end{aligned} \quad (3.19)$$

is continuous, and maps $(0, \infty)$ onto itself. Thus, the equation $L(k; T, 0) = x_0$ of (3.16) is satisfied for a unique constant $k = \mathcal{K}(x_0) \in (0, \infty)$. By the martingale representation property of the Brownian filtration (e.g. Karatzas & Shreve (1991)), we obtain

$$\begin{aligned} e^{-rt} \hat{X}(t) &\triangleq e^{-rT} \cdot \tilde{\mathbf{E}}_T \left[I \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, Y(T))} \right) \middle| \mathcal{F}(t) \right] \\ &= x_0 + \int_0^t e^{-rs} \hat{\pi}^*(s) \sigma dY(s), \quad 0 \leq t \leq T \end{aligned} \quad (3.20)$$

for some $\mathbf{F}$ -progressively measurable process $\hat{\pi} : [0, T] \times \Omega \rightarrow \mathbb{R}^d$ that satisfies $\int_0^T e^{-2rs} \|\hat{\pi}(t)\|^2 dt < \infty$ almost surely (with respect to both $\mathbf{P}$ and $\tilde{\mathbf{P}}_T$). Furthermore, we have

$$X^{x_0, \hat{\pi}}(t) \equiv \hat{X}(t) = \mathcal{X}(T - t, Y(t)), \quad 0 \leq t \leq T, \quad (3.21)$$

where

$$\mathcal{X}(s, y) \triangleq \begin{cases} e^{-rs} \int_{\mathbb{R}^d} I \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, y + z)} \right) \varphi_s(z) dz; & 0 < s \leq T \\ I \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, y)} \right); & s = 0 \end{cases} = L(\mathcal{K}(x_0); s, y). \quad (3.22)$$

This function solves the Cauchy problem

$$\mathcal{X}_s = \frac{1}{2} \Delta \mathcal{X} - r \mathcal{X}; \quad s > 0, \quad y \in \mathbb{R}^d \quad (3.23)$$

$$\mathcal{X}(0, y) = I \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, y)} \right); \quad s = 0, \quad y \in \mathbb{R}^d \quad (3.24)$$

for the heat-equation with cooling at rate $r \geq 0$. Together with (2.11), the equations (3.21) and (3.23) lead to the expression

$$\hat{\pi}(t) = (\sigma^*)^{-1} \cdot \nabla \mathcal{X}(T - t, Y(t)), \quad 0 \leq t < T \quad (3.25)$$

for the optimal portfolio of (3.20). Finally, in conjunction with (3.22), (3.6) and Assumption 3.1, we have

$$\nabla \mathcal{X}(s, y) = -\mathcal{K}(x_0) e^{-r(T+s)} \int_{\mathbb{R}^d} \frac{G(T, y + z)}{F(T, y + z)} I' \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, y + z)} \right) \varphi_s(z) dz \quad (3.26)$$

for the gradient in the equation (3.25). We can now formalize all of this, as follows.

Theorem 3.2 *For any given $x_0 > 0$, the control process $\hat{\pi}(\cdot) \in \mathcal{A}(x_0)$ of (3.25) and (3.26) is optimal for Problem 2.1. Its corresponding wealth process $\hat{X}(\cdot)$ is given*

by (3.21) and (3.22), and the value function of Problem 2.1 is

$$\begin{aligned} V(x_0) &= \mathbf{E}u(\hat{X}(T)) = \mathbf{E} \left[(u \circ I) \left(\frac{\mathcal{K}(x_0)e^{-rT}}{F(T, Y(T))} \right) \right] = \tilde{\mathbf{E}}_T[\hat{Z}(T) \cdot u(\mathcal{X}(0, Y(T)))] \\ &= \int_{\mathfrak{N}^d} F(T, z) \cdot (u \circ I) \left(\frac{\mathcal{K}(x_0)e^{-rT}}{F(T, z)} \right) \cdot \varphi_T(z) dz. \end{aligned} \quad (3.27)$$

Example 3.3 *Logarithmic utility function* $u(x) = \log(x)$. In this case $I(k) = 1/k$, the function of (3.19) becomes

$$L(k; T, 0) = e^{-rT} \cdot \tilde{\mathbf{E}}_T \left(\frac{F(T, Y(T))}{ke^{-rT}} \right) = \frac{1}{k} \cdot \tilde{\mathbf{E}}_T \hat{Z}(T) = \frac{1}{k},$$

and thus $\mathcal{K}(x_0) = 1/x_0$. From (3.20) and (3.9), we have

$$\begin{aligned} e^{-rt} \hat{X}(t) &= x_0 \cdot \tilde{\mathbf{E}}_T[F(T, Y(T)) | \mathcal{F}(t)] = x_0 \cdot \tilde{\mathbf{E}}_T[\hat{Z}(T) | \mathcal{F}(t)] = x_0 \hat{Z}(t) \\ &= x_0 + \int_0^t x_0 \hat{Z}(s) \hat{\Theta}^*(s) dY(s) = x_0 + \int_0^t e^{-rs} \hat{X}(s) \hat{\Theta}^*(s) dY(s), \\ &0 \leq t \leq T. \end{aligned}$$

This gives us the optimal portfolio-weight process

$$\hat{p}(t) \triangleq \frac{\hat{\pi}(t)}{\hat{X}(t)} = (\sigma^*)^{-1} \hat{\Theta}(t) = (\sigma^*)^{-1} G(t, Y(t)),$$

and thus also the optimal portfolio process in the form

$$\hat{\pi}(t) = (\sigma^*)^{-1} \hat{X}(t) \hat{\Theta}(t) = x_0 e^{rt} (\sigma^*)^{-1} \nabla F(t, Y(t)), \quad 0 \leq t < T.$$

In particular, the functions of (3.22), (3.27) now become

$$\mathcal{X}(s, y) = x_0 e^{r(T-s)} \cdot F(T-s, y); \quad 0 \leq s \leq T, \quad y \in \mathfrak{N}^d,$$

$$V(x_0) = \log(x_0) + rT + \int_{\mathfrak{N}^d} F(T, z) \log(F(T, z)) \varphi_T(z) dz, \quad 0 < x_0 < \infty.$$

Remark 3.4 In the special case $\mu = \delta_\theta$ for some $\theta \in \mathfrak{N}^d$, we have

$$F^{(\theta)}(t, y) = \exp \left(\theta y - \frac{1}{2} \|\theta\|^2 t \right) \quad \text{and} \quad G^{(\theta)}(t, y) = \theta,$$

so that $\hat{p}^{(\theta)}(t) = \hat{\pi}^{(\theta)}(t) / \hat{X}^{(\theta)}(t) = (\sigma^*)^{-1} \theta$. On the other hand, for a general prior distribution μ on Θ , we have the *certainty-equivalence principle*

$$\hat{p}(t) = \hat{\pi}(t) / \hat{X}(t) = (\sigma^*)^{-1} \mathbf{E}[\Theta | \mathcal{F}(t)] = \hat{p}^{(\theta)}(t) \Big|_{\theta = \mathbf{E}[\Theta | \mathcal{F}(t)]}. \quad (3.28)$$

Specifically, in the case of a logarithmic utility function, the optimal portfolio-proportion is obtained by substituting, in the expression $\hat{p}^{(t)}(\cdot)$ for the optimal portfolio-proportion corresponding to the Dirac measure δ_θ , the Bayes estimate $\mathbb{E}[\Theta|\mathcal{F}(t)]$ for the unobserved variable Θ .

Example 3.5 *Utility function of power-type* $u(x) = x^\alpha/\alpha$, for $\alpha < 1$, $\alpha \neq 0$. In this case $u'(x) = x^{\alpha-1}$, $I(k) = k^{-\beta}$ with $\beta = 1/(1 - \alpha)$, and thus

$$\mathcal{K}(x_0)e^{-rT} = e^{\alpha rT} \left(\frac{\tilde{\mathbb{E}}_T(F(T, Y(T)))^\beta}{x_0 e^{rT}} \right)^{1/\beta}.$$

Substitution back into (3.22) gives

$$e^{-r(T-s)} \cdot \mathcal{X}(s, y) = \begin{cases} x_0 \frac{\int_{\mathbb{R}^d} (F(T, y+z))^\beta \varphi_s(z) dz}{\int_{\mathbb{R}^d} (F(T, z))^\beta \varphi_T(z) dz}; & s > 0, y \in \mathbb{R}^d \\ x_0 \frac{(F(T, y))^\beta}{\int_{\mathbb{R}^d} (F(T, z))^\beta \varphi_T(z) dz}; & s = 0, y \in \mathbb{R}^d \end{cases},$$

$$e^{-r(T-s)} \cdot \nabla \mathcal{X}(s, y) = \beta x_0 \frac{\int_{\mathbb{R}^d} \nabla F(T, y+z) (F(T, y+z))^{\beta-1} \varphi_s(z) dz}{\int_{\mathbb{R}^d} (F(T, z))^\beta \varphi_T(z) dz};$$

$$s > 0, y \in \mathbb{R}^d,$$

$$\left(\frac{\nabla \mathcal{X}}{\mathcal{X}} \right)(s, y) = \beta \frac{\int_{\mathbb{R}^d} \nabla F(T, y+z) (F(T, y+z))^{\beta-1} \varphi_s(z) dz}{\int_{\mathbb{R}^d} (F(T, y+z))^\beta \varphi_s(z) dz}; \quad s > 0, y \in \mathbb{R}^d,$$

and

$$\hat{p}(t) = \frac{\hat{\pi}(t)}{\hat{X}(t)} = (\sigma^*)^{-1} \cdot \left(\frac{\nabla \mathcal{X}}{\mathcal{X}} \right)(T-t, Y(t)), \quad 0 \leq t < T.$$

On the other hand, (3.27) leads to the expression

$$V(x_0) = \frac{(x_0 e^{rT})^\alpha}{\alpha} \left(\int_{\mathbb{R}^d} (F(T, z))^\beta \varphi_T(z) dz \right)^{1/\beta}$$

for the value function.

Remark 3.6 In the special case $\mu = \delta_\theta$, we have $\frac{\nabla \mathcal{X}}{\mathcal{X}}(s, y) = \beta\theta$. This shows that the *certainty-equivalence principle* of (3.28) fails for utility functions of power-type $u(x) = x^\alpha/\alpha$ with $\alpha < 1$, $\alpha \neq 0$, because for a nondegenerate prior distribution μ we have typically

$$\left(\frac{\nabla \mathcal{X}}{\mathcal{X}} \right)(s, y) \neq \beta G(s, y) = \beta \left(\frac{\nabla F}{F} \right)(s, y),$$

or equivalently

$$\left(\frac{\nabla F}{F}\right)(s, y) \neq \frac{\int_{\mathbb{R}^d} \nabla F(T, y + z) (F(T, y + z))^{\beta-1} \varphi_s(z) dz}{\int_{\mathbb{R}^d} (F(T, y + z))^{\beta} \varphi_s(z) dz}.$$

Remark 3.7 For general utility functions, Kuwana (1995) proved that logarithmic utilities are the only ones for which the *certainty-equivalence principle* holds. Karatzas (1997) studied this property for the *goal problem* of maximizing the probability $\mathbf{P}[X(T) = 1]$ of reaching the “goal” $x = 1$ during the finite time-horizon $[0, T]$. For a more general nonnegative $\mathcal{F}(T)$ -measurable random variable C , the generalized *goal problem* of maximizing the probability $\mathbf{P}[X(T) \geq C]$ was studied in Section 3 of Spivak (1998) via a duality approach.

4 Dynamic programming

In this section we shall place Problem 2.1 within the standard framework of *Stochastic Control* and *Dynamic Programming* as expounded, for instance, in Fleming & Rishel (1975), Chapter 6 or Fleming & Soner (1993), Chapter 4. We shall show that the Hamilton–Jacobi–Bellman (HJB) equation for this problem reduces to a parabolic Monge–Ampère-type equation (4.12) with specific initial, boundary and concavity conditions (4.9)–(4.11). Using the martingale-based results of the previous section, we shall solve this equation explicitly. In order to simplify notation somewhat, we shall take $r = 0$, $\sigma = I_d$ in this section.

More precisely, for a general *utility function* $u(\cdot)$ we introduce the stochastic control problem

$$U(s, x, y) \triangleq \sup_{\pi(\cdot) \in \mathcal{A}(x; T-s, T)} \mathbf{E}u(X(T)), \quad (s, x, y) \in [0, T] \times (0, \infty) \times \mathbb{R}^d \quad (4.1)$$

on the time-horizon $[T - s, T]$, subject to the dynamics

$$dY(t) = \tilde{G}(T - t, Y(t))dt + dN(t); \quad Y(T - s) = y, \quad (4.2)$$

$$dX(t) = \pi^*(t)[\tilde{G}(T - t, Y(t))dt + dN(t)]; \quad X(T - s) = x, \quad (4.3)$$

by analogy with (3.7) and (2.12), respectively. Here $N(\cdot)$ is the *innovations process* introduced in Section 3, an $(\mathbf{F}, \mathbf{P})$ -Brownian motion on $[T - s, T]$; and $\tilde{G}(T - t, \cdot) \equiv G(t, \cdot)$. We expect the value function $U(\cdot)$ of (4.1) to be of class $C^{1,2,2}$ on the strip $(0, T) \times (0, \infty) \times \mathbb{R}^d$, and to satisfy the *Hamilton–Jacobi–Bellman (HJB) equation of Dynamic Programming*

$$U_s = \frac{1}{2} \Delta U + \tilde{G}^* \cdot \nabla U + \max_{\pi \in \mathbb{R}^d} \left[\frac{\|\pi\|^2}{2} U_{xx} + \pi^* (\tilde{G} U_x + \nabla U_x) \right]$$

$$= \frac{1}{2} \Delta U - \frac{1}{U_{xx}} \|\tilde{G} U_{xx} + \nabla U_x\|^2 + \tilde{G}^* \cdot \nabla U \quad (4.4)$$

associated with the dynamics of (4.2), (4.3) on this strip. We also expect the function of (4.1) to inherit the concavity property

$$U_{xx} < 0, \quad \text{on } (0, T) \times (0, \infty) \times \mathfrak{R}^d, \quad (4.5)$$

of the utility function $u(\cdot)$, and to satisfy the *initial condition*

$$U(0, x, y) = u(x), \quad \text{for } (x, y) \in (0, \infty) \times \mathfrak{R}^d \quad (4.6)$$

and the *boundary condition*

$$U(s, 0+, y) = u(0+), \quad \text{for } 0 < s < T, \quad y \in \mathfrak{R}^d. \quad (4.7)$$

Remark 4.1 For any given function $\phi(t, x, y) : [0, T] \times \mathfrak{R} \times \mathfrak{R}^d \rightarrow \mathfrak{R}$, we denote by $\phi_t = \frac{\partial \phi}{\partial t}$ the time-derivative, by $\phi_x = \frac{\partial \phi}{\partial x}$ the derivative with respect to x , by $\nabla \phi = (\frac{\partial \phi}{\partial y_1}, \dots, \frac{\partial \phi}{\partial y_d})^*$ the gradient with respect to y , and by $\Delta \phi = \sum_{i=1}^d \frac{\partial^2 \phi}{\partial y_i^2}$ the Laplacian with respect to y .

The equation of (4.4) looks quite complicated; it can be simplified somewhat, by use of the transformation

$$Q(s, x, y) \triangleq \begin{cases} U(s, x, y) \cdot F(T - s, y); & 0 \leq s < T \\ \lim_{\sigma \uparrow T} Q(\sigma, x, y); & s = T \end{cases}, \quad (4.8)$$

into the *initial-boundary value problem*

$$Q(0, x, y) = u(x) \cdot F(T, y), \quad \text{for } (x, y) \in (0, \infty) \times \mathfrak{R}^d, \quad (4.9)$$

$$Q(s, 0+, y) = u(0+) \cdot F(T - s, y), \quad \text{for } 0 < s < T, \quad y \in \mathfrak{R}^d, \quad (4.10)$$

$$Q_{xx} < 0, \quad \text{on } (0, T) \times (0, \infty) \times \mathfrak{R}^d, \quad (4.11)$$

for the equation

$$\begin{aligned} Q_s &= \frac{1}{2} \Delta Q + \max_{\pi \in \mathfrak{R}^d} \left[\frac{1}{2} \|\pi\|^2 Q_{xx} + \pi^* \nabla Q_x \right] \\ &= \frac{1}{2} \left[\Delta Q - \frac{\|\nabla Q_x\|^2}{Q_{xx}} \right], \quad \text{on } (0, T) \times (0, \infty) \times \mathfrak{R}^d. \end{aligned} \quad (4.12)$$

Remark 4.2 The equation (4.12) is the HJB equation associated with the *stochastic control problem* of maximizing

$$\mathbf{E}u(X(T)) = \tilde{\mathbf{E}}_T[\hat{Z}(T)u(X(T))] = \tilde{\mathbf{E}}_T[F(T, Y(T)) \cdot u(X(T))],$$

subject to the dynamics

$$\begin{aligned} dX(t) &= \pi^*(t)d\zeta(t), & X(T-s) &= x, \\ dY(t) &= d\zeta(t), & Y(T-s) &= y, \end{aligned}$$

on the time interval $[T-s, T]$, where $\zeta(\cdot)$ is an $(\mathbf{F}, \tilde{\mathbf{P}}_T)$ -Brownian motion with values in $\mathfrak{R}^d$. In the case $d = 1$, the equation (4.12) takes the form

$$2Q_{xx}Q_s = Q_{xx}Q_{yy} - (Q_{xy})^2 \quad (4.13)$$

of a *parabolic-Monge-Ampère* type equation, already encountered in Karatzas (1997).

Once we have managed to solve the initial-boundary value problem of (4.8)–(4.12), we can expect to recover the value function of (2.10) in the form

$$V(x_0) = U(T, x_0, 0) = Q(T, x_0, 0), \quad (4.14)$$

and the optimal portfolio process of (3.20) as

$$\hat{\pi}(t) = -\left(\frac{\nabla Q_x}{Q_{xx}}\right)(T-t, \hat{X}(t), Y(t)), \quad 0 \leq t < T. \quad (4.15)$$

Remark 4.3 In conjunction with (3.21) and (3.25), this equation suggests that the solution $Q(s, x, y)$ of (4.9)–(4.12) should be related to the function $\mathcal{X}(s, y)$ of (3.22) via

$$\nabla \mathcal{X}(s, y) = -\left(\frac{\nabla Q_x}{Q_{xx}}\right)(s, \mathcal{X}(s, y), y), \quad \text{on } (0, T] \times \mathfrak{R}^d. \quad (4.16)$$

Let us consider now the value process

$$\begin{aligned} h(t) &\triangleq \mathbf{E}[u(\hat{X}(T))|\mathcal{F}(t)] = \mathbf{E}\left[(u \circ I)\left(\frac{\mathcal{K}(x_0)}{F(T, Y(T))}\right)\middle|\mathcal{F}(t)\right] \\ &= \frac{1}{F(t, Y(t))} \tilde{\mathbf{E}}_T\left[F(T, Y(T)) \cdot (u \circ I)\left(\frac{\mathcal{K}(x_0)}{F(T, Y(T))}\right)\middle|\mathcal{F}(t)\right] \\ &= \frac{\mathcal{H}(T-t, Y(t))}{F(t, Y(t))}, \quad 0 < t \leq T \end{aligned} \quad (4.17)$$

and $h(0) \triangleq \mathcal{H}(T, 0)$, where we have set

$$\mathcal{H}(s, y) \triangleq \left\{ \begin{array}{ll} \int_{\mathfrak{R}^d} F(T, y+z) \cdot (u \circ I)\left(\frac{\mathcal{K}(x_0)}{F(T, y+z)}\right) \varphi_s(z) dz; & 0 < s \leq T \\ F(T, y) \cdot (u \circ I)\left(\frac{\mathcal{K}(x_0)}{F(T, y)}\right); & s = 0 \end{array} \right\} \quad (4.18)$$

for $y \in \mathbb{R}^d$. This function satisfies the heat equation

$$\mathcal{H}_s = \frac{1}{2} \Delta \mathcal{H} \quad \text{on } (0, T) \times \mathbb{R}^d, \quad (4.19)$$

as well as

$$V(x_0) = \mathcal{H}(T, 0). \quad (4.20)$$

Now (4.20) and (4.14) imply that we should have $\mathcal{H}(T, 0) = Q(T, x_0, 0)$, which then suggests the even more general relation

$$\mathcal{H}(s, y) = Q(s, \mathcal{X}(s, y), y), \quad \text{for } (s, y) \in [0, T] \times \mathbb{R}^d. \quad (4.21)$$

This reduces to (4.20) for $s = T$, $y = 0$, since $\mathcal{X}(T, 0) = L(\mathcal{K}(x_0); T, 0) = x_0$.

Before establishing the solvability of the initial-boundary value problem (4.9)–(4.12) and the validity of the expressions (4.16) and (4.21), let us continue the discussion of Examples 3.1 and 3.2.

Example 4.4 *Logarithmic utility function* $u(x) = \log(x)$ and $r = 0$, $\sigma = I_d$. In this case, we have $u(I(k)) = \log 1/k$ and $\mathcal{K}(x_0) = 1/x_0 = \tilde{F}(s, y)/\mathcal{X}(s, y)$, where we have defined $\tilde{F}(s, y) \triangleq F(T - s, y)$; recall the computations of Example 3.1. Since $\hat{Z}(t) = F(t, Y(t))$ is an $(\mathbf{F}, \tilde{\mathbf{P}}_T)$ -martingale,

$$\begin{aligned} \tilde{F}(s, y) &= F(T - s, y) = \mathbf{E}^0[F(T, Y(T)) | Y(T - s) = y] \\ &= \int_{\mathbb{R}^d} F(T, y + z) \varphi_s(z) dz. \end{aligned} \quad (4.22)$$

Thus, the expression of (4.18) becomes

$$\begin{aligned} \mathcal{H}(s, y) &= \int_{\mathbb{R}^d} F(T, y + z) \log \left(\frac{\mathcal{X}(s, y)}{\tilde{F}(s, y)} \cdot F(T, y + z) \right) \varphi_s(z) dz \\ &= \log \left(\frac{\mathcal{X}(s, y)}{\tilde{F}(s, y)} \right) \int_{\mathbb{R}^d} F(T, y + z) \varphi_s(z) dz + \rho(s, y) \\ &= \tilde{F}(s, y) \log \left(\frac{\mathcal{X}(s, y)}{\tilde{F}(s, y)} \right) + \rho(s, y), \end{aligned} \quad (4.23)$$

where

$$\rho(s, y) \triangleq \int_{\mathbb{R}^d} F(T, \xi) \log F(T, \xi) \varphi_s(y - \xi) d\xi. \quad (4.24)$$

Note that both $\tilde{F}(s, y)$ and $\rho(s, y)$ solve the heat-equation $q_s = \frac{1}{2} \Delta q$. Now the expression of (4.23) leads, in conjunction with the Ansatz (4.21), to the *conjecture*

$$Q(s, x, y) = \tilde{F}(s, y) \log \left(\frac{x}{\tilde{F}(s, y)} \right) + \rho(s, y) \quad (4.25)$$

for the solution of the initial-boundary value problem of (4.9)–(4.12). Indeed, for the function Q of (4.25), we have $Q(0, x, y) = F(T, y) \log x$ for $s = 0$ (since $\rho(0, y) = F(T, y) \log F(T, y)$), and

$$\begin{aligned} Q_x(s, x, y) &= \frac{\tilde{F}(s, y)}{x}, \\ \nabla Q_x(s, x, y) &= \frac{\nabla \tilde{F}(s, y)}{x}, \\ Q_{xx}(s, x, y) &= \frac{-\tilde{F}(s, y)}{x^2} < 0 \end{aligned}$$

for $s > 0$. In particular, the requirements (4.9)–(4.11) are satisfied. We can also compute

$$Q_s(s, x, y) = \tilde{F}_s(s, y) \cdot \log x - \tilde{F}_s(s, y)(1 + \log \tilde{F}(s, y)) + \rho_s(s, y),$$

$$\nabla Q(s, x, y) = \nabla \tilde{F}(s, y) \cdot \log x - \nabla \tilde{F}(s, y)(1 + \log \tilde{F}(s, y)) + \nabla \rho(s, y),$$

and

$$\begin{aligned} \Delta Q(s, x, y) &= \Delta \tilde{F}(s, y) \cdot \log x - \frac{\|\nabla \tilde{F}(s, y)\|^2}{\tilde{F}(s, y)} - \Delta \tilde{F}(s, y)(1 + \log \tilde{F}(s, y)) \\ &\quad + \Delta \rho(s, y). \end{aligned}$$

Substituting these expressions into (4.12), we can see readily that this equation is satisfied. It is also straightforward to compute

$$-\left(\frac{\nabla Q_x}{Q_{xx}}\right)(s, x, y) = x \cdot \left(\frac{\nabla F}{F}\right)(T - s, y),$$

so that

$$-\left(\frac{\nabla Q_x}{Q_{xx}}\right)(s, \mathcal{X}(s, y), y) = \mathcal{X}(s, y) \cdot \left(\frac{\nabla F}{F}\right)(T - s, y) = \nabla \mathcal{X}(s, y)$$

and thus (4.16) is also satisfied.

Remark 4.5 Recall that for any two probability measures P and Q on a measurable space $(\Omega, \mathcal{F})$, the *relative entropy of P with respect to Q , conditional on a sub- σ -algebra $\mathcal{G}$ of $\mathcal{F}$* , is defined as

$$H_{\mathcal{G}}(P|Q) \triangleq \begin{cases} \mathbf{E}^P \left[\log \frac{dP}{dQ} | \mathcal{G} \right]; & \text{if } P \ll Q \text{ on } \mathcal{G} \\ \infty; & \text{otherwise} \end{cases}. \quad (4.26)$$

Now, for the probability measures $\mathbf{P}$ and $\tilde{\mathbf{P}}_T$, we can compute the relative entropy, conditional on the σ -algebra $\mathcal{F}(t)$, in the form

$$\begin{aligned}
 H_{\mathcal{F}(t)}(\mathbf{P}|\tilde{\mathbf{P}}_T) &= \mathbf{E} \left[\log \left(\frac{d\mathbf{P}}{d\tilde{\mathbf{P}}_T} \Big|_{\mathcal{F}(T)} \right) \Big| \mathcal{F}(t) \right] = \mathbf{E} \left[\log \hat{Z}(T) \Big| \mathcal{F}(t) \right] \\
 &= \frac{1}{\hat{Z}(t)} \tilde{\mathbf{E}}_T \left[\hat{Z}(T) \log \hat{Z}(T) \Big| \mathcal{F}(t) \right] \\
 &= \frac{1}{F(t, Y(t))} \tilde{\mathbf{E}}_T \left[(F \log F)(T, Y(t) + Y(T) - Y(t)) \Big| \mathcal{F}(t) \right] \\
 &= \frac{1}{F(t, y)} \int_{\mathbb{R}^d} (F \log F)(T, y + z) \cdot \varphi_s(z) dz \Big|_{s=T-t, y=Y(t)} \\
 &= \frac{\rho(s, y)}{\tilde{F}(s, y)} \Big|_{s=T-t, y=Y(t)}. \tag{4.27}
 \end{aligned}$$

This provides an interpretation of the last term in the expression

$$Q(s, x, y) = \tilde{F}(s, y) \left[\log \left(\frac{x}{\tilde{F}(s, y)} \right) + \frac{\rho(s, y)}{\tilde{F}(s, y)} \right]$$

of (4.25) for the value-function, in terms of conditional relative entropy.

Example 4.6 *Utility function of power-type* $u(x) = x^\alpha/\alpha$, for $\alpha < 1$, $\alpha \neq 0$ and $r = 0$, $\sigma = I_d$. In this case $u(I(k)) = (1/\alpha) \cdot k^{-\alpha/(1-\alpha)} = (1/\alpha) \cdot k^{-\alpha\beta}$, and we have

$$(\mathcal{K}(x_0))^\beta = \frac{\int_{\mathbb{R}^d} (F(T, z))^\beta \varphi_T(z) dz}{x_0} = \frac{1}{\mathcal{X}(s, y)} \int_{\mathbb{R}^d} (F(T, y + z))^\beta \varphi_s(z) dz$$

for all $s > 0$, from the computation of Example 3.2. Substituting this expression into (4.18), we obtain

$$\begin{aligned}
 \mathcal{H}(s, y) &= \frac{1}{\alpha} \int_{\mathbb{R}^d} F(T, y + z) \left(\frac{F(T, y + z)}{\mathcal{K}(x_0)} \right)^{\beta\alpha} \varphi_s(z) dz \\
 &= \frac{(\mathcal{X}(s, y))^\alpha}{\alpha} \left(\int_{\mathbb{R}^d} (F(T, y + z))^\beta \varphi_s(z) dz \right)^{\frac{1}{\beta}}.
 \end{aligned}$$

This suggests that the function

$$Q(s, x, y) = \frac{x^\alpha}{\alpha} \rho(s, y), \quad \rho(s, y) \triangleq \left(\int_{\mathbb{R}^d} (F(T, y + z))^\beta \varphi_s(z) dz \right)^{\frac{1}{\beta}} \tag{4.28}$$

solves the initial-boundary value problem (4.9)–(4.12), for the HJB equation (4.12). Substitution of the first expression of (4.28) into (4.12), leads to the equa-

tion

$$\rho_s = \frac{1}{2} \Delta \rho + \frac{\beta - 1}{2} \frac{\|\nabla \rho\|^2}{\rho} \quad (4.29)$$

that the function ρ of (4.28) must satisfy. To check this, observe that the function

$$v(s, y) \triangleq (\rho(s, y))^\beta = \int_{\mathbb{R}^d} (F(T, \xi))^\beta \varphi_s(y - \xi) d\xi$$

solves the heat-equation $v_s = \frac{1}{2} \Delta v$; and that

$$v_s = \beta \rho^{\beta-1} \rho_s, \quad \nabla v = \beta \rho^{\beta-1} \nabla \rho, \quad \Delta v = \beta \rho^{\beta-1} \Delta \rho + \beta(\beta - 1) \rho^{\beta-2} \|\nabla \rho\|^2.$$

Substituting these derivatives into the heat-equation for v , we arrive at the equation (4.29). The conditions (4.9)–(4.11) are rather straightforward to check, directly from (4.28).

Let us return now to the case of a general utility function $u(\cdot)$. We have

$$\mathcal{X}(s, y) = L(\mathcal{K}(x_0); s, y) \quad (4.30)$$

from (3.22) and (3.17). Under Assumption 3.1, for every $(s, y) \in [0, T] \times \mathbb{R}^d$ the mapping $L(\cdot) \equiv L(\cdot; s, y)$ of (3.17) is continuous and strictly decreasing with $L(0+) = \infty$, $L(\infty) = 0$. After denoting the (continuous, strictly decreasing) inverse of this mapping by $K(\cdot; s, y)$, we observe that $K(\cdot; 0, y) = F(T, y)u'(\cdot)$, that

$$\mathcal{K}(x_0) = K(\mathcal{X}(s, y); s, y) \quad (4.31)$$

holds for every $(s, y) \in [0, T] \times \mathbb{R}^d$ from (4.30), and that (4.18) yields

$$\mathcal{H}(s, y) = \int_{\mathbb{R}^d} F(T, y + z) \cdot (u \circ I) \left(\frac{K(\mathcal{X}(s, y); s, y)}{F(T, y + z)} \right) \cdot \varphi_s(z) dz \quad (4.32)$$

for $s > 0$. In conjunction with the Ansatz (4.21), this suggests the following result.

Theorem 4.7 *The function*

$$Q(s, x, y) \triangleq \begin{cases} \int_{\mathbb{R}^d} F(T, y + z) \cdot (u \circ I) \left(\frac{K(x; s, y)}{F(T, y + z)} \right) \cdot \varphi_s(z) dz; & s > 0, x > 0, y \in \mathbb{R}^d \\ F(T, y) \cdot u(x); & s = 0, x > 0, y \in \mathbb{R}^d \end{cases} \quad (4.33)$$

solves the initial-boundary value problem (4.9)–(4.12). Furthermore, this function satisfies the conditions of (4.14)–(4.16) and (4.21).

We defer to the Appendix the extensive computations required for the proof.

5 The cost of uncertainty

Let us suppose now that there is an “insider” investor, who *can* observe *both* the drift-vector Θ and the driving Brownian motion $W(\cdot)$, in the model $\mathcal{M}$ of (2.11) for the financial market. In other words, the trading strategies $\pi(\cdot)$ available to this investor are adapted to the *enlarged filtration* $\mathbf{G}$ of (2.4).

More formally, let us introduce a nonnegative, $\mathcal{G}(0) = \sigma(\Theta)$ -measurable random variable $X(0)$ with $\tilde{\mathbf{E}}_T X(0) = x_0$; this random variable will play the role of initial wealth for this “insider” investor at time $t = 0$. We denote by $\mathcal{A}_*(x_0)$ the class of the pairs $(X(0), \pi(\cdot))$, where $X(0)$ is as in the previous sentence and the $\mathbf{G}$ -progressively measurable process $\pi : [0, T] \times \Omega \rightarrow \mathfrak{R}^d$ satisfies the conditions (2.7) and

$$0 \leq e^{-rt} X^{x_0, \pi}(t) = X(0) + \int_0^t e^{-rs} \pi^*(s) \sigma dY(s), \quad \forall 0 \leq t \leq T \quad (2.8)'$$

almost surely. The objective of this “insider” investor is also to maximize the expected utility of his wealth at the terminal time $t = T$, so the optimization problem he faces has value function

$$V_*(x_0) \triangleq \sup_{(X(0), \pi(\cdot)) \in \mathcal{A}_*(x_0)} \mathbf{E}u(X^{x_0, \pi}(T)), \quad (5.1)$$

for $x_0 > 0$. For any $\pi(\cdot) \in \mathcal{A}(x_0)$, it is clear that $(x_0, \pi(\cdot)) \in \mathcal{A}_*(x_0)$, so

$$V(x_0) \leq V_*(x_0). \quad (5.2)$$

The martingale methodology of Section 3 can now be repeated, in fact with $Z(\cdot)$, $\Lambda(\cdot) = 1/Z(\cdot)$ of (2.5) replacing their “filtered” counterparts $\hat{Z}(\cdot)$, $\hat{\Lambda}(\cdot) = 1/\hat{Z}(\cdot)$ of (3.8) and (3.9). In particular, $e^{-r\cdot} \Lambda(\cdot) X^{x_0, \pi}(\cdot) = e^{-r\cdot} X^{x_0, \pi}(\cdot)/Z(\cdot)$ is now a nonnegative $(\mathbf{G}, \mathbf{P})$ -local martingale, hence also supermartingale, for every $\pi(\cdot) \in \mathcal{A}_*(x_0)$. By analogy with (3.20)–(3.27), we conclude that the value function of (5.1) takes the form

$$\begin{aligned} V_*(x_0) &= \mathbf{E} \left[(u \circ I) \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{Z(T)} \right) \right] = \mathbf{E} \left[(u \circ I) \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{\exp(\Theta^* W(T) + T \|\Theta\|^2/2)} \right) \right] \\ &= \int_{\mathfrak{N}^d} \int_{\mathfrak{N}^d} (u \circ I) \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{\exp(\vartheta^* w + T \|\vartheta\|^2/2)} \right) \varphi_T(w) dw \mu(d\vartheta), \end{aligned} \quad (5.3)$$

where $\mathcal{K}_*(\cdot)$ is the inverse of the mapping

$$k \longmapsto e^{-rT} \int_{\mathfrak{N}^d} \int_{\mathfrak{N}^d} I \left(\frac{k e^{-rT}}{\exp(\vartheta^* y - T \|\vartheta\|^2/2)} \right) \varphi_T(y) dy \mu(d\vartheta) \quad \text{on } (0, \infty), \quad (5.4)$$

under the assumption that the integral of (5.4) is finite on $(0, \infty)$. Therefore, the optimal wealth process $\check{X}(\cdot)$ is given by

$$e^{-rt} \check{X}(t) \triangleq e^{-rT} \cdot \tilde{\mathbf{E}}_T \left[I \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{Z(T)} \right) \middle| \mathcal{G}(t) \right] = e^{-rt} \mathcal{X}_*(T-t, Y(t); \Theta), \quad (5.5)$$

$$0 \leq t \leq T$$

with $\tilde{\mathbf{E}}_T[\check{X}(0)] = e^{-rT} \cdot \tilde{\mathbf{E}}_T \left[I \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{Z(T)} \right) \right] = x_0$ and

$$\mathcal{X}_*(s, y; \vartheta) \triangleq \begin{cases} e^{-rs} \int_{\mathbb{R}^d} I \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{\exp(\vartheta^* z - T \|\vartheta\|^2/2)} \right) \varphi_s(y-z) dz; & 0 < s \leq T \\ I \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{\exp(\vartheta^* y - T \|\vartheta\|^2/2)} \right); & s = 0 \end{cases}. \quad (5.6)$$

Under conditions analogous to those of Assumption 3.1, the function $(s, y) \mapsto \mathcal{X}_*(s, y; \vartheta)$ satisfies the heat-equation

$$\frac{\partial}{\partial s} \mathcal{X}_* = \frac{1}{2} \Delta \mathcal{X}_* - r \mathcal{X}_*, \quad \text{on } (0, T) \times (0, \infty)^d,$$

for every $\vartheta \in \mathbb{R}^d$. In conjunction with Lemma 2.1 and Itô's rule, this leads to the stochastic integral representation of (2.8)', $e^{-rt} \check{X}(t) = \check{X}(0) + \int_0^t e^{-rs} \tilde{\pi}^*(s) \sigma dY(s)$, $0 \leq t \leq T$ with $\check{X}(0) = \mathcal{X}_*(T, 0; B)$ and

$$\tilde{\pi}(t) = (\sigma^*)^{-1} \nabla \mathcal{X}_*(T-t, Y(t); \Theta), \quad 0 \leq t < T. \quad (5.7)$$

The resulting pair $(\check{X}(0), \tilde{\pi}(\cdot)) \in \mathcal{A}_*(x_0)$ then attains the supremum in (5.1).

Remark 5.1 With these assumptions and notations, the ratio

$$1 - \frac{V(x_0)}{V_*(x_0)} = 1 - \frac{\int_{\mathbb{R}^d} (u \circ I) \left(\frac{\mathcal{K}(x_0) e^{-rT}}{F(T, z)} \right) F(T, z) \varphi_T(z) dz}{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (u \circ I) \left(\frac{\mathcal{K}_*(x_0) e^{-rT}}{\exp(\vartheta^* w + T \|\vartheta\|^2/2)} \right) \varphi_T(w) dw \mu(d\vartheta)} \quad (5.8)$$

has the significance of *relative cost for the uncertainty associated with the prior distribution μ* , in the context of a utility function $u(\cdot)$ from terminal wealth.

Example 5.2 In the case of the logarithmic utility function $u(x) = \log(x)$, we have $\mathcal{K}_*(x_0) = 1/x_0$ from (5.4). The $(\mathbf{G}, \tilde{\mathbf{P}}_T)$ -martingale of (5.5) takes the form

$$e^{-rt} \check{X}(t) = x_0 \cdot \tilde{\mathbf{E}}_T [Z(T) | \mathcal{G}(t)] = x_0 Z(t) = x_0 + \int_0^t x_0 Z(s) \Theta^* dY(s), \quad 0 \leq t \leq T \quad (5.9)$$

from Lemma 2.1, and thus admits the representation (5.7) with

$$\tilde{\pi}(t) \triangleq (\sigma^*)^{-1} \Theta \check{X}(t) = x_0 Z(t) e^{rt} (\sigma^*)^{-1} \Theta, \quad 0 \leq t \leq T. \quad (5.10)$$

This pair $(x_0, \tilde{\pi}(\cdot)) \in \mathcal{A}_*(x_0)$ is therefore optimal for the problem (5.1), whose value function is then given by (5.3) as

$$\begin{aligned} V_*(x_0) &= \log x_0 + rT + \mathbf{E}(\Theta^* W(T) + T \|\Theta\|^2/2) \\ &= \log x_0 + rT + \frac{T}{2} \int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta). \end{aligned} \quad (5.11)$$

From the computations of Examples 3.3 and 4.4, the relative-cost ratio of (5.8) takes the form

$$\begin{aligned} 1 - \frac{V(x_0)}{V_*(x_0)} &= \frac{T \int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta) - 2 \int_{\mathbb{R}^d} F(T, y) \log F(T, y) \varphi_T(y) dy}{2 \log x_0 + 2rT + T \int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta)} \\ &= \frac{\int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta) - \frac{2}{T} \rho(T, 0)}{\frac{2}{T} \log x_0 + 2r + \int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta)} \end{aligned} \quad (5.12)$$

in the notation of (4.24), for any distribution μ with $\int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta) < \infty$.

Remark 5.3 In the special case where μ is the multivariate normal distribution $\mathcal{N}(\theta, v^2 I)$, for some $\theta \in \mathbb{R}^d$ and $v^2 > 0$, the function of (3.2) is easily computed as

$$F(t, y) = (1 + tv^2)^{-d/2} \exp \left[\frac{\|\theta + v^2 y\|^2}{2v^2(1 + tv^2)} - \frac{\|\theta\|^2}{2v^2} \right]. \quad (5.13)$$

In particular, we have $F(t, y) \varphi_t(y) = (2\pi t(1 + tv^2))^{-d/2} \exp(-\frac{\|y - t\theta\|^2}{2t(1 + tv^2)})$, and the relative-cost ratio of (5.12) takes the form

$$1 - \frac{V(x_0)}{V_*(x_0)} = \frac{d \log(1 + Tv^2)}{2 \log x_0 + T(2r + \|\theta\|^2 + dv^2)}. \quad (5.14)$$

The expression of (5.14) tends to zero, as $T \rightarrow \infty$; in other words, as the planning horizon gets large, the relative cost of uncertainty becomes negligible.

This property holds in great generality, as our next result shows.

Proposition 5.4 *For a logarithmic utility function, the relative cost of uncertainty in (5.12) tends to zero as $T \rightarrow \infty$, for any prior distribution μ with $\int_{\mathbb{R}^d} \|\vartheta\|^2 \mu(d\vartheta) < \infty$.*

Proof: From (5.12) it suffices to show that $\lim_{T \rightarrow \infty} \frac{2}{T} \rho(T, 0) = \int_{\mathfrak{H}^d} \|\vartheta\|^2 \mu(d\vartheta)$, or equivalently

$$\lim_{T \rightarrow \infty} \frac{1}{T} \mathbf{E} \left[\log \hat{Z}(T) \right] = \frac{1}{2} \int_{\mathfrak{H}^d} \|\vartheta\|^2 \mu(d\vartheta) \quad (5.15)$$

by virtue of (4.24), and (4.27) with $t = 0$. Now, we have

$$\hat{\Lambda}(t) = \frac{1}{\hat{Z}(t)} = \exp \left[- \int_0^t \hat{\Theta}^*(s) dN(s) - \frac{1}{2} \int_0^t \|\hat{\Theta}(s)\|^2 ds \right]$$

from (3.10), and

$$\mathbf{E} \int_0^T \|\hat{\Theta}(t)\|^2 dt \leq \mathbf{E} \int_0^T \|\Theta\|^2 dt = T \int_{\mathfrak{H}^d} \|\vartheta\|^2 \mu(d\vartheta) < \infty,$$

so that

$$\begin{aligned} \mathbf{E}[\log \hat{Z}(T)] &= \mathbf{E} \left[\int_0^T \hat{\Theta}^*(t) dN(t) + \frac{1}{2} \int_0^T \|\hat{\Theta}(t)\|^2 dt \right] \\ &= \frac{1}{2} \int_0^T \mathbf{E} \|\hat{\Theta}(t)\|^2 dt. \end{aligned} \quad (5.16)$$

Clearly from (3.5), $\|\hat{\Theta}(\cdot)\|^2$ is an $(\mathbf{F}, \mathbf{P})$ -submartingale; thus, $\lim_{t \rightarrow \infty} \mathbf{E} \|\hat{\Theta}(t)\|^2$ exists and is dominated by $\mathbf{E} \|\Theta\|^2$. On the other hand, from (3.8) and Fatou's lemma, we have

$$\mathbf{E} \|\Theta\|^2 = \mathbf{E} \left[\lim_{t \rightarrow \infty} \|\hat{\Theta}(t)\|^2 \right] \leq \lim_{t \rightarrow \infty} \mathbf{E} \|\hat{\Theta}(t)\|^2,$$

so that (5.16) yields

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \mathbf{E} [\log \hat{Z}(T)] &= \frac{1}{2} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbf{E} [\|\hat{\Theta}(t)\|^2] dt \\ &= \frac{1}{2} \lim_{t \rightarrow \infty} \mathbf{E} [\|\hat{\Theta}(t)\|^2] = \frac{1}{2} \mathbf{E} \|\Theta\|^2 = \frac{1}{2} \int_{\mathfrak{H}^d} \|\vartheta\|^2 \mu(d\vartheta), \end{aligned}$$

proving (5.15).

Example 5.5 In the case of the utility function $u(x) = x^\alpha / \alpha$ for $0 < \alpha < 1$, and with $\beta = \frac{1}{1-\alpha}$, we have

$$(x_0 e^{rT}) \cdot (\mathcal{K}_*(x_0) e^{-rT})^\beta = \int_{\mathfrak{H}^d} \exp \left(\frac{T}{2} \beta(\beta - 1) \|\vartheta\|^2 \right) \mu(d\vartheta),$$

provided that this last expression is finite, i.e.

$$\int_{\mathfrak{H}^d} \exp \left(\frac{\alpha T \|\vartheta\|^2}{2(1-\alpha)^2} \right) \mu(d\vartheta) < \infty. \quad (5.17)$$

The function of (5.6) takes the form

$$\begin{aligned}\mathcal{X}_*(s, y; \vartheta) &= (\mathcal{K}_*(x_0)e^{-rT})^{-\beta} \exp[\beta y^* \vartheta - \beta s(T - \beta s) \|\vartheta\|^2/2]; \\ 0 \leq s \leq T, \quad y &\in \mathfrak{N}^d\end{aligned}$$

for every $\vartheta \in \mathfrak{N}^d$, and the optimal portfolio $\check{\pi}(\cdot) \in \mathcal{A}_*(x_0)$ and wealth processes $\check{X}(\cdot) \equiv X^{x_0, \check{\pi}}(\cdot)$ are given as

$$\check{X}(t) = \mathcal{X}_*(T - t, Y(t); \Theta), \quad \check{\pi}(t) = \frac{(\sigma^*)^{-1} \Theta}{1 - \alpha} \check{X}(t); \quad 0 \leq t \leq T.$$

Finally, from (5.3) the value function for the problem of (5.1) takes the form

$$\begin{aligned}V_*(x_0) &= \frac{1}{\alpha} (\mathcal{K}_*(x_0)e^{-rT})^{-\alpha\beta} \int_{\mathfrak{N}^d} \int_{\mathfrak{N}^d} e^{\alpha\beta(\vartheta^* w + T \|\vartheta\|^2/2)} (2\pi T)^{-d/2} e^{-\frac{\|w\|^2}{2T}} dw \mu(d\vartheta) \\ &= \frac{(x_0 e^{rT})^\alpha}{\alpha} \left(\int_{\mathfrak{N}^d} \exp\left(\frac{\alpha T \|\vartheta\|^2}{2(1-\alpha)^2}\right) \mu(d\vartheta) \right)^{1-\alpha}.\end{aligned}\quad (5.18)$$

Along with the computations from Examples 3.5 and 4.6, that is

$$V(x_0) = \frac{(x_0 e^{rT})^\alpha}{\alpha} \left(\int_{\mathfrak{N}^d} (F(T, z))^{\frac{1}{1-\alpha}} \varphi_T(z) dz \right)^{1-\alpha},$$

the relative-cost ratio of (5.8) becomes in this case

$$1 - \frac{V(x_0)}{V_*(x_0)} = 1 - \left(\frac{\int_{\mathfrak{N}^d} (F(T, z))^{\frac{1}{1-\alpha}} \varphi_T(z) dz}{\int_{\mathfrak{N}^d} \exp\left(\frac{\alpha T \|b\|^2}{2(1-\alpha)^2}\right) \mu(db)} \right)^{1-\alpha}.\quad (5.19)$$

Remark 5.6 In the case where the prior distribution μ is multivariate normal $\mathcal{N}(\theta, v^2 I)$ for some $\theta \in \mathfrak{N}^d$ and $v^2 > 0$, the condition (5.17) is satisfied if $\alpha T v^2 < (1 - \alpha)^2$. In this case the ratio (5.19) takes the form

$$1 - \frac{V(x_0)}{V_*(x_0)} = 1 - \frac{\left(\frac{1-\alpha\beta^2 v^2 T}{1-\alpha\beta v^2 T}\right)^{d(1-\alpha)/2}}{(1 + v^2 T)^{d\alpha/2}} \exp\left(-\frac{\alpha^3 \beta^3 \|\theta\|^2 v^2 T^2}{2(1 - \alpha\beta^2 v^2 T)(1 - \alpha\beta v^2 T)}\right),$$

which tends to 1 as $T \rightarrow (1 - \alpha)^2 / \alpha v^2 = 1 / \alpha \beta^2 v^2$.

6 The constrained optimization problems

Let us consider now a nonempty, closed and convex set $K \in \mathfrak{N}^d$, and introduce the function

$$\delta(x) \equiv \delta(x|K) \triangleq \sup_{p \in K} (-p^* x) : \mathfrak{N}^d \rightarrow \mathfrak{N} \cup \{+\infty\},\quad (6.1)$$

which is finite on its *effective domain*

$$\tilde{K} \triangleq \{x \in \mathfrak{R}^d; \delta(x|K) < \infty\} = \{x \in \mathfrak{R}^d; \exists \beta \in \mathfrak{R} \text{ s.t. } -p^*x \leq \beta, \forall p \in K\}. \quad (6.2)$$

The function $\delta(\cdot)$ is the *support function* of the convex set $-K$, and $\tilde{K}$ is a convex cone, called the *barrier cone* of $-K$.

Assumption 6.1 *We assume throughout that*

$$\text{the function } \delta(\cdot) \text{ is continuous on } \tilde{K} \quad (6.3)$$

and bounded from below by some real constant:

$$\delta(x|K) \geq \delta_0, \quad \forall x \in \mathfrak{R}^d \text{ for some } \delta_0 \in \mathfrak{R}. \quad (6.4)$$

Remark 6.2 A sufficient condition for (6.3) to hold, is that $\tilde{K}$ be locally simplicial (cf. Rockafellar (1970), Theorem 10.2, p. 84); and (6.4) holds if K contains the origin.

For any $\pi(\cdot) \in \mathcal{A}(x_0)$, we define $\tau_\pi \triangleq \{t \in [0, T] / X^{x_0, \pi}(t) \equiv X(t) = 0\} \wedge T$, following the convention $\inf \emptyset = \infty$. From (2.8), it is clear that $X(\cdot)$ and $\pi(\cdot)$ are identically equal to zero on $[\tau_\pi, T] \triangleq \{(t, \omega) \in [0, T] \times \Omega / \tau_\pi(\omega) \leq t \leq T\}$. We can now introduce the portfolio-weight process $p(\cdot) = (p_1(\cdot), \dots, p_d(\cdot))^*$, where

$$p_i(t) \triangleq \begin{cases} \pi_i(t) / X(t) : & 0 \leq t < \tau_\pi \\ k^* : & \tau_\pi \leq t \leq T \end{cases}, \quad (6.5)$$

for $i = 1, \dots, d$ and an arbitrary but fixed vector $k^* \in K$. It is straightforward to see that $\pi(\cdot) = X(\cdot)p(\cdot)$ on $[[0, T] \triangleq [0, T] \times \Omega$. We have already encountered such portfolio-weight processes in Examples 3.1 and 3.2. It is clear that $p_i(t)$ represents the proportion of the wealth $X(t)$ invested in the i th stock at time t . Thus, from (2.11) and (3.7), the wealth process $X(\cdot)$ satisfies on $[0, T]$ the stochastic differential equation

$$\begin{aligned} dX(t) - rX(t)dt &= X(t)p^*(t)\sigma dY(t) \equiv X(t)p^*(t)\sigma[\hat{\Theta}(t)dt + dN(t)], \\ X(0) &= x_0 > 0. \end{aligned} \quad (6.6)$$

From now on, we shall constrain the portfolio-weight process $p(\cdot)$ to take values in the convex set K . More precisely, we say that a portfolio process $\pi(\cdot)$ is *admissible for the initial wealth $x_0 > 0$ and the constraint set K* , and write $\pi \in \mathcal{A}(x_0; K)$, if $\pi(\cdot) \in \mathcal{A}(x_0)$ and if its corresponding portfolio-weight process $p(\cdot)$ of (6.5) satisfies $p(\cdot) \in K$ almost everywhere on $[[0, T]]$. We can now state the constrained version of Problem 2.4, as follows.

Problem 6.3 For given utility function $u(\cdot)$ and convex set $K \in \mathfrak{R}^d$, maximize the expected utility from $X(\cdot)$ of (6.6) at the terminal time T , over the class $\mathcal{A}(x_0; K)$. The value function of this problem will be denoted by

$$V(x_0; K) \triangleq \sup_{\pi(\cdot) \in \mathcal{A}(x_0; K)} \mathbf{E} u(X^{x_0, \pi}(T)). \quad (6.7)$$

Here are some examples of constraint sets. All of them satisfy the Assumption 6.1.

Example 6.4 *Prohibition of short-selling of stocks:* $p_i(\cdot) \geq 0$, $1 \leq i \leq d$. In other words, $K \triangleq [0, \infty)^d$. Thus, we have $\tilde{K} = [0, \infty)^d$ and $\delta(\cdot) \equiv 0$ on $\tilde{K}$.

Example 6.5 *Incomplete market; only the first n stocks can be traded:* $p_i(\cdot) = 0$, $\forall i = n+1, \dots, d$, for some fixed $n \in \{1, \dots, d-1\}$. In other words, $K \triangleq \{p \in \mathfrak{R}^d / p_{n+1} = \dots = p_d = 0\}$. Thus, we have $\tilde{K} = \{p \in \mathfrak{R}^d / p_1 = \dots = p_n = 0\}$ and $\delta(\cdot) \equiv 0$ on $\tilde{K}$.

Example 6.6 *Constraints on the short-selling of stocks:* $p_i(\cdot) \geq -k$, $1 \leq i \leq d$, for some $k > 0$. In other words, $K = [-k, \infty)^d$. Thus, we have $\delta(x) = k \sum_{i=1}^d x_i$ and $\tilde{K} = [0, \infty)^d$.

Remark 6.7 Under the *full observations* framework, this problem was solved by Cvitanic & Karatzas (1992) using martingale methods, along with duality theory and convex analysis. In the following section, we adapt their methodology to the model $\mathcal{M}$ of Section 2, i.e.

$$dS_0(t) = rS_0(t)dt, \quad S_0(0) = 1 \quad (6.8)$$

$$dS_i(t) = S_i(t) \left[\hat{B}_i(t)dt + \sum_{j=1}^d \sigma_{ij} dN_j(t) \right], \quad S_i(0) > 0 \quad (6.9)$$

where $\hat{B}_i(t) \equiv (\sigma \hat{\Theta}(t))_i + r$, for $i = 1, \dots, d$. We summarize the solution of Problem 6.3 in Theorem 7.3.

7 Auxiliary markets and optimality conditions

Let us consider now the space $\mathcal{H}$ of $\mathbf{F}$ -progressively measurable processes $v : [0, T] \times \Omega \rightarrow \mathfrak{R}^d$, with $\mathbf{E} \int_0^T (\|v(t)\|^2 + \delta(v(t)))dt < \infty$, and define

$$\mathcal{D} \triangleq \left\{ v \in \mathcal{H} / v(t, \omega) \in \tilde{K}, \text{ for } (\ell \otimes \mathbf{P})\text{-a.e. } (t, \omega) \in [0, T] \times \Omega \right\}. \quad (7.1)$$

For any given $v(\cdot) \in \mathcal{D}$, we modify the model $\mathcal{M}$ of (6.8), (6.9) as follows: we introduce an auxiliary financial market $\mathcal{M}_v$ with *money-market*

$$dS_0^{(v)}(t) = S_0^{(v)}(t)[r + \delta(v(t))]dt, \quad (7.2)$$

and d stocks, with price-per-share processes $S_i^{(v)}(\cdot)$ governed by

$$\begin{aligned} dS_i^{(v)}(t) &= S_i^{(v)}(t) \left[(B_i + v_i(t) + \delta(v(t)))dt + \sum_{j=1}^d \sigma_{ij} dW_j(t) \right] \\ &= S_i^{(v)}(t) \left[(\hat{B}_i(t) + v_i(t) + \delta(v(t)))dt + \sum_{j=1}^d \sigma_{ij} dN_j(t) \right] \end{aligned} \quad (7.3)$$

for $i = 1, \dots, d$. In this new market model $\mathcal{M}_v$, the wealth process $X_v(\cdot) \equiv X_v^{x_0, \pi}(\cdot)$, corresponding to initial capital $x_0 > 0$ and portfolio $\pi(\cdot)$, satisfies

$$dX_v^{x_0, \pi}(t) = \left(X_v^{x_0, \pi}(t) - \sum_{i=1}^d \pi_i(t) \right) \frac{dS_0^{(v)}(t)}{S_0^{(v)}(t)} + \sum_{i=1}^d \pi_i(t) \frac{dS_i^{(v)}(t)}{S_i^{(v)}(t)}. \quad (7.4)$$

As in Section 2, we shall denote by $\mathcal{A}_v(x_0)$ the class of the portfolio processes $\pi(\cdot)$ which satisfy (2.7) and

$$X_v^{x_0, \pi}(t) \geq 0, \quad \forall 0 \leq t \leq T, \quad (7.5)$$

P-almost surely. Furthermore, for any $\pi(\cdot) \in \mathcal{A}_v(x_0)$, we can define the portfolio-weight process $p(\cdot)$ through (6.5), so that the wealth-equation (7.4) takes the form

$$\begin{aligned} dX_v^{x_0, \pi}(t) &= X_v^{x_0, \pi}(t) \left[\left(1 - \sum_{i=1}^d p_i(t) \right) \frac{dS_0^{(v)}(t)}{S_0^{(v)}(t)} + \sum_{i=1}^d p_i(t) \frac{dS_i^{(v)}(t)}{S_i^{(v)}(t)} \right] \\ &= X_v^{x_0, \pi}(t) \left[(r + \delta(v(t)) + p^*(t)v(t))dt + p^*(t)\sigma(\hat{\Theta}(t)dt + dN(t)) \right]. \end{aligned} \quad (7.6)$$

The class $\mathcal{A}_v(x_0)$ is the set of our admissible control processes for the *unconstrained optimization problem* in the *auxiliary market* $\mathcal{M}_v$; this is to maximize the expected utility from $X_v^{x_0, \pi}(\cdot)$ of (7.6), for the given utility function $u(\cdot)$ at the terminal time T . The value function of this problem will be denoted by

$$V_v(x_0) \triangleq \sup_{\pi(\cdot) \in \mathcal{A}_v(x_0)} \mathbf{E} u(X_v^{x_0, \pi}(T)). \quad (7.7)$$

Remark 7.1 For any $v(\cdot) \in \mathcal{D}$, $\pi(\cdot) \in \mathcal{A}(x_0; K)$ and its corresponding portfolio-weight process $p(\cdot)$, a comparison of (6.6) with (7.6) gives

$$X_v^{x_0, \pi}(t) \geq X^{x_0, \pi}(t) \geq 0, \quad \forall 0 \leq t \leq T, \quad (7.8)$$

almost surely, because we have $\delta(v(t)) + p^*(t)v(t) \geq 0$ for $p(t) \in K$. Thus, it is straightforward to see that $\mathcal{A}(x_0; K) \subseteq \mathcal{A}_v(x_0)$ and

$$V(x_0; K) \leq V_v(x_0), \quad \forall v \in \mathcal{D}. \quad (7.9)$$

In the new market $\mathcal{M}_v$ of (7.2) and (7.3), we define the analogue

$$d\hat{\Lambda}_v(t) = -\hat{\Lambda}_v(t)[\hat{\Theta}(t) + \sigma^{-1}v(t)]^* dN(t), \quad \hat{\Lambda}_v(0) = 1 \quad (7.10)$$

of the exponential process $\hat{\Lambda}(\cdot)$ of (3.10), and also denote by

$$H_v(t) \triangleq \hat{\Lambda}_v(t)/S_0^{(v)}(t), \quad 0 \leq t \leq T, \quad (7.11)$$

the corresponding *state-price-density* process. For any $\pi(\cdot) \in \mathcal{A}(x_0)$, an application of Itô's rule gives

$$d(H_v(t)X_v^{x_0, \pi}(t)) = H_v(t)X_v^{x_0, \pi}(t)[\sigma^*p(t) - (\hat{\Theta}(t) + \sigma^{-1}v(t))]^* dN(t), \quad (7.12)$$

where $p(\cdot)$ is the portfolio-weight process corresponding to $\pi(\cdot)$. In other words, $H_v(\cdot)X_v^{x_0, \pi}(\cdot)$ is a nonnegative $(\mathbf{F}, \mathbf{P})$ -local martingale, thus also a supermartingale; therefore,

$$\mathbf{E}[H_v(t)X_v^{x_0, \pi}(t)] \leq x_0, \quad \forall \pi(\cdot) \in \mathcal{A}_v(x_0). \quad (7.13)$$

We can now use the methodology of Section 3 to solve the unconstrained optimization problem (7.7) in $\mathcal{M}_v$. Let us start by observing the inequality

$$\mathbf{E}u(X_v^{x_0, \pi}(T)) \leq x_0k + \mathbf{E}\tilde{u}(kH_v(T)), \quad \text{for every } k > 0, \pi(\cdot) \in \mathcal{A}_v(x_0), \quad (7.14)$$

by analogy with (3.14). Equality holds in (7.14) if and only if we have both

$$X_v^{x_0, \pi}(T) = I(kH_v(T)), \quad \text{a.s.}, \quad (7.15)$$

$$\mathbf{E}[H_v(T)X_v^{x_0, \pi}(T)] = x_0; \quad (7.16)$$

these are analogues of (3.15) and (3.16).

Assumption 7.2 Suppose that

$$\mathcal{X}_v(k) \triangleq \mathbf{E}[H_v(T)I(kH_v(T))] < \infty, \quad \forall 0 < k < \infty. \quad (7.17)$$

Under this assumption, the strictly decreasing function $\mathcal{X}_v(\cdot)$ maps $(0, \infty)$ onto itself. We denote by $\mathcal{Y}_v(\cdot)$ the unique inverse function of $\mathcal{X}_v(\cdot)$. Therefore, (7.15) and (7.16) give us the *optimal terminal wealth*

$$\hat{X}_v(T) \equiv C_v \triangleq I(\mathcal{Y}_v(x_0)H_v(T)) \quad (7.18)$$

for the problem of (7.7), whose value function takes the form

$$V_v(x_0) = J_v(\mathcal{Y}_v(x_0)) \quad (7.19)$$

with the notation

$$J_v(k) \triangleq \mathbf{E}[(u \circ I)(kH_v(T))]. \quad (7.20)$$

From the Fujisaki–Kallianpur–Kunita representation theorem (e.g. Kallianpur (1980), Elliott (1982), Rogers & Williams (1987)) there exists an $\mathbf{F}$ -progressively measurable process $\psi_v : [0, T] \times \Omega \rightarrow \Re^d$ with $\int_0^T \|\psi_v(t)\|^2 dt < \infty$, a.s., such that the *optimal wealth process* is given as

$$\begin{aligned} \hat{X}_v(t) &= \frac{1}{H_v(t)} \mathbf{E}[H_v(T)C_v | \mathcal{F}(t)] = \frac{1}{H_v(t)} \left(x_0 + \int_0^t \psi_v^*(s) dN(s) \right), \\ 0 \leq t \leq T. \end{aligned} \quad (7.21)$$

Together with (7.12), this gives us the *optimal portfolio process* $\pi_v(\cdot)$ in the form

$$\hat{\pi}_v(t) = (\sigma^*)^{-1} \left[\frac{\psi_v(t)}{H_v(t)} + (\hat{\Theta}(t) + \sigma^{-1}v(t)) \hat{X}_v(t) \right], \quad 0 \leq t \leq T, \quad (7.22)$$

as well as the *optimal portfolio-weight process*

$$\hat{p}_v(t) = \frac{\hat{\pi}_v(t)}{\hat{X}_v(t)} = (\sigma^*)^{-1} \left[\frac{\psi_v(t)}{H_v(t) \hat{X}_v(t)} + (\hat{\Theta}(t) + \sigma^{-1}v(t)) \right], \quad 0 \leq t \leq T. \quad (7.23)$$

Furthermore, from (7.14), we have

$$\mathbf{E}[\tilde{u}(kH_v(T))] \geq \mathbf{E}u(X_v^{x,\pi}(T)) - xk, \quad \forall x > 0, \pi \in \mathcal{A}_v(x) \quad (7.24)$$

for every $0 < k < \infty$. In particular, this gives

$$\mathbf{E}[\tilde{u}(kH_v(T))] \geq \tilde{V}_v(k), \quad \forall k > 0, \quad (7.25)$$

for the convex dual

$$\tilde{V}_v(k) \triangleq \sup_{x>0} [V_v(x) - xk] \quad (7.26)$$

of the value function (7.7). On the other hand, (7.24) holds as equality when $x = \mathcal{X}_v(k)$ and $\pi(\cdot) \equiv \hat{\pi}_v(\cdot)$ as in (7.22). Thus

$$\begin{aligned} \mathbf{E}[\tilde{u}(kH_v(T))] &= V_v(\mathcal{X}_v(k)) - k\mathcal{X}_v(k) \\ &= J_v(k) - k\mathcal{X}_v(k) \leq \tilde{V}_v(k). \end{aligned} \quad (7.27)$$

Along with (7.25), this leads to

$$\tilde{V}_v(k) = J_v(k) - k\mathcal{X}_v(k) = \mathbf{E}[\tilde{u}(kH_v(T))]. \quad (7.28)$$

We can now solve the constrained Problem 6.3 by the following optimality conditions and Theorem 7.3, which are adapted from Cvitanic & Karatzas (1992).

For a fixed initial capital $x_0 > 0$, let $\hat{\pi}(\cdot) \in \mathcal{A}(x_0; K)$ be a given portfolio process. In the financial market $\mathcal{M}$, its corresponding portfolio-weight process and

wealth process are denoted by $\hat{p}(\cdot)$ and $\hat{X}(\cdot)$, respectively, with $\hat{\pi}(\cdot)$ taking values in the closed, convex set K . Let us consider the statement that $\hat{p}(\cdot)$ is optimal for the constrained Problem 6.3:

(A) *Optimality of $\hat{\pi}$* : We have

$$V(x_0; K) = \mathbf{E}u(\hat{X}(T)) < \infty. \quad (7.29)$$

We shall characterize the optimality condition (A) in terms of the following conditions (B)–(E), which concern a given process $\mu \in \mathcal{D}$.

(B) *Financeability of C_μ* : There exists a portfolio process $\hat{\pi}_\mu(\cdot) \in \mathcal{A}(x_0; K)$, such that its corresponding portfolio-weight process $\hat{p}_\mu(\cdot)$ and wealth process $\hat{X}_\mu(\cdot)$ satisfy the properties

$$\hat{p}_\mu(t) \in K, \quad \delta(\mu(t)) + \hat{p}_\mu^*(t)\mu(t) = 0, \quad X^{x_0, \hat{\pi}_\mu}(t) = \hat{X}_\mu(t)$$

$\ell \otimes \mathbf{P}$ -almost everywhere on $[0, T] \times \Omega$.

(C) *Minimality of μ* : We have

$$V_\mu(x_0) \leq V_\nu(x_0), \quad \forall \nu \in \mathcal{D}. \quad (7.30)$$

(D) *Dual optimality of μ* : We have

$$\tilde{V}_\mu(\mathcal{Y}_\mu(x_0)) \leq \tilde{V}_\nu(\mathcal{Y}_\nu(x_0)), \quad \forall \nu \in \mathcal{D}. \quad (7.31)$$

(E) *Parsimony of μ* : We have

$$\mathbf{E}[H_\nu(T)C_\mu] \leq x_0, \quad \forall \nu \in \mathcal{D}. \quad (7.32)$$

Theorem 7.3 *The conditions (B)–(E) are equivalent, and imply condition (A) with $\hat{\pi}(\cdot) = \hat{\pi}_\mu(\cdot)$. Conversely, condition (A) implies the existence of a process $\mu \in \mathcal{D}$ that satisfies (B)–(E) with $\hat{\pi}_\mu(\cdot) = \hat{\pi}(\cdot)$, provided that the utility function $u(\cdot)$ satisfies the following conditions:*

- (a) $x \mapsto x \cdot u'(x)$ is nondecreasing on $(0, \infty)$; and
- (b) for some $\beta \in (0, 1)$, $\gamma \in (1, \infty)$, we have $\beta \cdot u'(x) \geq u'(\gamma x)$, $\forall x \in (0, \infty)$.

Example 7.4 *Logarithmic utility function* $u(x) = \log(x)$. In this case we have $\mathcal{X}_\nu(k) = 1/k$ and $\hat{X}_\nu(T) = x_0/H_\nu(T)$. This gives $H_\nu(\cdot)\hat{X}_\nu(\cdot) \equiv x_0$, thus $\psi_\nu(\cdot) \equiv 0$ for every $\nu \in \mathcal{D}$, and the optimal portfolio-weight process for the auxiliary, unconstrained problem of (7.7) takes the form

$$\hat{p}_\nu(t) = (\sigma^*)^{-1}[\hat{\Theta}(t) + \sigma^{-1}\nu(t)] = (\sigma^*)^{-1}[G(t, Y(t)) + \sigma^{-1}\nu(t)].$$

Furthermore, the value function for the auxiliary optimization problem (7.7) is given by

$$\begin{aligned} V_v(x_0) &= \mathbf{E} \left[\log \left(\frac{x_0}{H_v(T)} \right) \right] = \log(x_0) - \mathbf{E}(\log(\hat{\Lambda}_v(T)) - \log(S_0^{(v)}(T))) \\ &= \log(x_0) + rT + \mathbf{E} \int_0^T \left[\delta(v(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}v(t)\|^2 \right] dt. \end{aligned} \quad (7.33)$$

Observe that the expression (7.33) is minimized by $\mu(\cdot)$ in $\mathcal{D}$, given by

$$\mu(t) = M(\hat{\Theta}(t)), \quad 0 \leq t \leq T, \quad \text{where} \quad M(\vartheta) \triangleq \arg \min_{v \in \tilde{K}} \left[\delta(v) + \frac{1}{2} \|\vartheta + \sigma^{-1}v\|^2 \right]. \quad (7.34)$$

Now, for the original *constrained* optimization problem, we have $\hat{p}(\cdot) \equiv \hat{p}_\mu(\cdot)$, and

$$V(x_0; K) = V_\mu(x_0) = \log(x_0) + rT + \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt.$$

Example 6.4 (cont'd) *Prohibition of short-selling of stocks*, $\sigma = I_d$. In this case $\delta(\cdot) \equiv 0$, thus $\mu_i(t) = (\hat{\Theta}_i(t))^-$, and $\hat{p}_i(t) = (\hat{p}_\mu)_i(t) = (\hat{\Theta}_i(t))^+$ for $i = 1, \dots, d$, as well as

$$V(x_0; K) = V_\mu(x_0) = \log(x_0) + rT + \frac{1}{2} \mathbf{E} \int_0^T \sum_{i=1}^d \left((\hat{\Theta}_i(t))^+ \right)^2 dt.$$

Example 6.5 (Cont'd) *Incomplete market*, $\sigma = I_d$. In this case $\delta(\cdot) \equiv 0$, thus $\mu_1(\cdot) = \dots = \mu_n(\cdot) \equiv 0$, and $\mu_i(t) = -\hat{\Theta}_i(t)$, $i = n+1, \dots, d$. This gives us $\hat{p}_i(t) = \hat{\Theta}_i(t)$ for $i = 1, \dots, n$ and $\hat{p}_i(\cdot) \equiv 0$ for $i = n+1, \dots, d$, as well as

$$V(x_0; K) = V_\mu(x_0) = \log(x_0) + rT + \frac{1}{2} \mathbf{E} \int_0^T [\hat{\Theta}_1^2(t) + \dots + \hat{\Theta}_n^2(t)] dt.$$

Example 6.6 (Cont'd) *Constraints on the short-selling of stocks*, $\sigma = I_d$. In this case $\delta(v) = k \sum_{i=1}^d v_i$, thus $\mu_i(t) = (\hat{\Theta}_i(t) + k)^-$. This gives us

$$\hat{p}_i(t) = (\hat{p}_\mu)_i(t) = \hat{\Theta}_i(t) + (\hat{\Theta}_i(t) + k)^- = \hat{\Theta}_i(t) \vee (-k),$$

and

$$V(x_0; K) = \log(x_0) + rT + \mathbf{E} \int_0^T \sum_{i=1}^d \left[k(\hat{\Theta}_i(t) + k)^- + \frac{1}{2} (\hat{\Theta}_i(t) \vee (-k))^2 \right] dt.$$

Remark 7.5 Let us consider now the *cost of uncertainty* in the case of Example 7.4. As in our discussion of Section 4, it is easy to see that the optimal portfolio-weight

process for the constrained problem of an investor with “inside information” about the random variable Θ , is

$$\hat{p}_* = (\sigma^*)^{-1}[\Theta + \sigma^{-1}m_*], \quad \text{where } m_* \triangleq M(\Theta),$$

in the notation of (7.34), and that the value function takes the form

$$V_*(x_0; K) = V_{m_*}(x_0) = \log(x_0) + rT + T \cdot \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right].$$

We are assuming here that

$$\begin{aligned} & \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right] \\ &= \int_{\mathfrak{R}^d} \left[\delta(M(\vartheta)) + \frac{1}{2} \|\vartheta + \sigma^{-1}M(\vartheta)\|^2 \right] \mu(d\vartheta) < \infty. \end{aligned}$$

Thus, the relative-cost ratio of (5.8) is now given by the expression

$$\begin{aligned} 1 - \frac{V(x_0; K)}{V_*(x_0; K)} &= 1 - \frac{\log(x_0) + rT + \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt}{\log(x_0) + rT + T \cdot \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right]} \\ &= \frac{\mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right] - \frac{1}{T} \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt}{r + (\log(x_0)/T) + \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right]}. \end{aligned}$$

As in Proposition 5.4, we want to show again that this ratio goes to zero, as T tends to infinity. Clearly, from $V_*(x_0; K) \geq V(x_0; K)$, we have

$$\begin{aligned} & \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right] \\ & \geq \frac{1}{T} \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt, \quad \forall T > 0. \end{aligned}$$

Therefore, it is sufficient to prove that

$$\begin{aligned} & \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right] \\ & \leq \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt. \quad (7.35) \end{aligned}$$

For a given $x \in \mathfrak{R}^d$ and any sequence $\{x_n, n \in \mathbb{N}\}$ which converges to x , we observe that $\{v_n \triangleq M(x_n), n \in \mathbb{N}\}$ is bounded because of Assumption 6.1. Thus, it has a convergent subsequence $\{v_{n_k}, k \in \mathbb{N}\}$, and we denote $\tilde{v} = \lim_{k \rightarrow \infty} v_{n_k}$. From the definition of $M(\cdot)$ in (7.34), we have

$$\delta(v_{n_k}) + \frac{1}{2} \|x_{n_k} + \sigma^{-1}v_{n_k}\|^2 \leq \delta(v) + \frac{1}{2} \|x_{n_k} + \sigma^{-1}v\|^2, \quad \text{for } v \triangleq M(x);$$

letting $k \rightarrow \infty$, we obtain

$$\delta(\tilde{v}) + \frac{1}{2}\|x + \sigma^{-1}\tilde{v}\|^2 \leq \delta(v) + \frac{1}{2}\|x + \sigma^{-1}v\|^2 \quad (7.36)$$

from Assumption 6.1. In conjunction with the strict convexity of $\lambda \mapsto \delta(\lambda) + \frac{1}{2}\|x + \sigma^{-1}\lambda\|^2$, the equality (7.35) leads to $\tilde{v} = v \equiv M(x)$. In other words, we have $\lim_{k \rightarrow \infty} M(x_{n_k}) = M(x)$, which establishes the continuity of the function $M(\cdot)$ of (7.34). Along with (3.8), this gives also $\lim_{t \rightarrow \infty} \mu(t) = \lim_{t \rightarrow \infty} M(\hat{\Theta}(t)) = M(\Theta) = m_*$ almost surely. From Fatou's lemma, we obtain then

$$\begin{aligned} & \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbf{E} \int_0^T \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] dt \\ &= \liminf_{t \rightarrow \infty} \mathbf{E} \left[\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right] \\ &\geq \mathbf{E} \left[\liminf_{t \rightarrow \infty} \left(\delta(\mu(t)) + \frac{1}{2} \|\hat{\Theta}(t) + \sigma^{-1}\mu(t)\|^2 \right) \right] \\ &= \mathbf{E} \left[\delta(m_*) + \frac{1}{2} \|\Theta + \sigma^{-1}m_*\|^2 \right], \end{aligned}$$

proving (7.34).

8 Appendix: proofs of selected results

Proof of Lemma 2.1 From (2.4) we have

$$\begin{aligned} \mathbf{E}[W(t) - W(s) | \mathcal{G}(s)] &= \mathbf{E}[W(t) - W(s) | \sigma(\Theta) \vee \mathcal{F}^W(s)] \\ &= \mathbf{E}[W(t) - W(s)] = 0, \quad \mathbf{P}\text{-a.s.} \end{aligned}$$

for $0 \leq s \leq t < \infty$, as well as

$$\begin{aligned} \mathbf{E}[W^2(t) - W^2(s) | \mathcal{G}(s)] &= \mathbf{E}[(W(t) - W(s))^2 | \mathcal{G}(s)] \\ &= \mathbf{E}[(W(t) - W(s))^2 | \sigma(\Theta) \vee \mathcal{F}^W(s)] \\ &= \mathbf{E}[(W(t) - W(s))^2] = t - s, \quad \mathbf{P}\text{-a.s.} \end{aligned}$$

thanks to our assumptions about the distribution of $(W(\cdot), \Theta)$ under $\mathbf{P}$. In other words, the process $W(\cdot)$ is indeed a $(\mathbf{G}, \mathbf{P})$ -Brownian motion by P. Lévy's theorem (e.g. Karatzas and Shreve (1991)), as it is a continuous $(\mathbf{G}, \mathbf{P})$ -martingale with quadratic variation equal to t . Similarly, because $\mathcal{F}^W(s)$ is independent of both $W(t) - W(s)$ and $\sigma(\Theta)$ under $\mathbf{P}$, we have

$$\begin{aligned} \mathbf{E}[e^{-\Theta(W(t) - W(s))} | \sigma(\Theta) \vee \mathcal{F}^W(s)] &= \mathbf{E}[e^{-\vartheta(W(t) - W(s))}]|_{\vartheta=\Theta} \\ &= e^{\frac{1}{2}\vartheta^2(t-s)}|_{\vartheta=\Theta} = e^{\frac{1}{2}\Theta^2(t-s)}, \quad \mathbf{P}\text{-a.s.} \end{aligned}$$

for $0 \leq s \leq t < \infty$, and this leads to the martingale property of the process $\Lambda(\cdot)$ in (2.4). $\square$

Proof of Lemma 2.2 The process $Y(\cdot)$ is a $(\mathbf{G}, \tilde{\mathbf{P}}_T)$ -Brownian motion, thanks to the Girsanov theorem (e.g. Karatzas & Shreve (1991), Section 3.5) and the fact that $W(\cdot)$ is a $(\mathbf{G}, \mathbf{P})$ -Brownian motion. Now $Y(\cdot)$ is independent of $\mathcal{G}(0) = \sigma(\Theta)$ under $\tilde{\mathbf{P}}_T$, from the definition of Brownian motion (independence of increments). Furthermore, for any $A \in \mathcal{B}(\mathfrak{N}^d)$, we have $\mu_0(A) \stackrel{\Delta}{=} \mathbf{P}[\Theta \in A] = \nu_0(A) = \tilde{\mathbf{P}}_T[\Theta \in A] = \mu(A)$ from (3.3) and (3.4). $\square$

Proof of Theorem 4.7 From definition (4.33), we know that $Q(s, x, y)$ satisfies the boundary condition (4.9). For any $0 < s < T$, $y \in \mathfrak{N}^d$, since $K(0+; s, y) = \infty$, we have $(u \circ I)(K(0+; s, y) \cdot F(T, y + z)) = u(0+)$, thus

$$Q(s, 0+, y) = u(0+) \cdot \int_{\mathfrak{N}^d} F(T, y + z) \varphi_s(z) dz = u(0+) F(T - s, y)$$

from (4.22). In other words, $Q(s, x, y)$ satisfies the boundary condition (4.10). We need to prove that it also satisfies (4.11) and (4.12). From the definition (3.17) we know that the function

$$(s, y) \mapsto L(k; s, y) = \int_{\mathfrak{N}^d} I\left(\frac{k}{F(T, z)}\right) \varphi_s(y - z) dz$$

satisfies the heat-equation

$$L_s = \frac{1}{2} \Delta L \quad \text{on } (0, \infty) \times \mathfrak{N}^d \quad (8.1)$$

for every $k > 0$. We also have

$$L_k(k; s, y) = \int_{\mathfrak{N}^d} \frac{1}{F(T, z)} I'\left(\frac{k}{F(T, z)}\right) \varphi_s(y - z) dz, \quad (8.2)$$

$$\nabla L_k(k; s, y) = \int_{\mathfrak{N}^d} \frac{1}{F(T, z)} I'\left(\frac{k}{F(T, z)}\right) \nabla \varphi_s(y - z) dz. \quad (8.3)$$

From

$$L(K(x; s, y); s, y) = x \quad (8.4)$$

we have

$$L_k(K(x; s, y); s, y) \cdot K_x(x; s, y) = 1 \quad (8.5)$$

and $L_s(K(x; s, y); s, y) + L_k(K(x; s, y); s, y) \cdot K_s(x; s, y) = 0$, so that

$$\frac{1}{K_x(x; s, y)} = L_k(K(x; s, y); s, y) = \int_{\mathbb{R}^d} \frac{1}{F(T, z)} I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \varphi_s(y - z) dz, \quad (8.6)$$

$$L_s(K(x; s, y); s, y) = -\frac{K_s(x; s, y)}{K_x(x; s, y)}. \quad (8.7)$$

From (8.5) we obtain also $\nabla(L_k(K(x; s, y); s, y)) = \nabla(1/K_x(x; s, y))$, which leads to the equation

$$\begin{aligned} & (\nabla L_k)(K(x; s, y); s, y) + L_{kk}(K(x; s, y); s, y) \cdot \nabla K(x; s, y) \\ &= -\left(\frac{\nabla K_x}{K_x^2} \right)(x; s, y). \end{aligned} \quad (8.8)$$

Furthermore, from (8.4) we have $\nabla(L(K(x; s, y); s, y)) = 0$, which yields

$$(\nabla L)(K(x; s, y); s, y) + L_k(K(x; s, y); s, y) \cdot \nabla K(x; s, y) = 0. \quad (8.9)$$

Differentiating (8.9) with respect to y , we get

$$\begin{aligned} & (\Delta L)(K(x; s, y); s, y) + 2(\nabla L_k)(K(x; s, y); s, y) \cdot \nabla K(x; s, y) \\ &+ L_{kk}(K(x; s, y); s, y) \|\nabla K(x; s, y)\|^2 + L_k(K(x; s, y); s, y) \Delta K(x; s, y) = 0. \end{aligned} \quad (8.10)$$

In conjunction with (8.5) and (8.8), this gives

$$\begin{aligned} (\Delta L)(K(x; s, y); s, y) &= 2 \left(\frac{\nabla K_x \cdot \nabla K}{K_x^2} \right)(x; s, y) - \left(\frac{\Delta K}{K_x} \right)(x; s, y) \\ &+ L_{kk}(K(x; s, y); s, y) \|\nabla K(x; s, y)\|^2. \end{aligned} \quad (8.11)$$

Substituting (8.7) and (8.11) back into the heat-equation (8.1), we obtain the equation

$$\frac{K_s}{K_x} + \frac{\nabla K_x \cdot \nabla K}{K_x^2} + \frac{1}{2} L_{kk}(K(x; s, y); s, y) \|\nabla K\|^2 - \frac{\Delta K}{2K_x} = 0. \quad (8.12)$$

On the other hand, starting from the definition (4.33), we get

$$\begin{aligned} \mathcal{Q}_x(s, x, y) &= \int_{\mathbb{R}^d} F(T, z) \frac{K(x; s, y)}{F(T, z)} I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \frac{K_x(x; s, y)}{F(T, z)} \varphi_s(y - z) dz \\ &= K(x; s, y) K_x(x; s, y) \int_{\mathbb{R}^d} \frac{1}{F(T, z)} I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \varphi_s(y - z) dz \\ &= K(x; s, y) \end{aligned} \quad (8.13)$$

in conjunction with (8.6) and the even symmetry of $\varphi_s(\cdot)$, thus also

$$Q_{xx}(s, x, y) = K_x(x, s, y), \quad (8.14)$$

$$\nabla Q_x(s, x, y) = \nabla K(x, s, y). \quad (8.15)$$

Now (8.14), (8.6) and the strict decrease of $I(\cdot)$ imply that the function $Q(x, s, y)$ indeed satisfies the condition (4.11). We can also compute

$$\begin{aligned} Q_s(s, x, y) &= \int_{\mathbb{R}^d} F(T, z) \frac{K(x; s, y)}{F(T, z)} \cdot I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \frac{K_s(x; s, y)}{F(T, z)} \varphi_s(y - z) dz \\ &\quad + \int_{\mathbb{R}^d} F(T, z) \cdot (u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \frac{\partial \varphi_s}{\partial s}(y - z) dz \\ &= \left(\frac{K K_s}{K_x} \right) (x; s, y) \\ &\quad + \int_{\mathbb{R}^d} F(T, x) \cdot (u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \frac{\partial \varphi_s}{\partial s}(y - z) dz \end{aligned} \quad (8.16)$$

and

$$\begin{aligned} \nabla Q(s, x, y) &= \int_{\mathbb{R}^d} F(T, z) \nabla \left[(u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \right] \varphi_s(y - z) dz \\ &\quad + \int_{\mathbb{R}^d} F(T, z) (u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \nabla \varphi_s(y - z) dz \\ &= \left(\frac{K \nabla K}{K_x} \right) (x; s, y) \\ &\quad + \int_{\mathbb{R}^d} F(T, z) (u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \nabla \varphi_s(y - z) dz. \end{aligned} \quad (8.17)$$

Differentiating (8.17) with respect to y , we obtain

$$\begin{aligned} \Delta Q(s, x, y) &= \left(\frac{(\nabla K \cdot \nabla K + K \Delta K) K_x - K \nabla K \cdot \nabla K_x}{K_x^2} \right) (x; s, y) \\ &\quad + \int_{\mathbb{R}^d} F(T, z) \nabla \left[(u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \right] \cdot \nabla \varphi_s(y - z) dz \\ &\quad + \int_{\mathbb{R}^d} F(T, z) \left[(u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \right] \Delta \varphi_s(y - z) dz. \end{aligned} \quad (8.18)$$

Using (8.3), we can rewrite the second term of the right hand side of (8.18) as

$$\int_{\mathbb{R}^d} F(T, z) \nabla \left[(u \circ I) \left(\frac{K(x; s, y)}{F(T, z)} \right) \right] \cdot \nabla \varphi_s(y - z) dz$$

$$\begin{aligned}
&= \int_{\mathbb{R}^d} F(T, z) \frac{K(x; s, y)}{F(T, z)} I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \frac{\nabla K(x; s, y)}{F(T, z)} \cdot \nabla \varphi_s(y - z) dz \\
&= (K \nabla K)(x; s, y) \cdot \int_{\mathbb{R}^d} \frac{1}{F(T, z)} I' \left(\frac{K(x; s, y)}{F(T, z)} \right) \nabla \varphi_s(y - z) dz \\
&= (K \nabla K)(x; s, y) \cdot (\nabla L_k)(K(x; s, y); s, y) \\
&= (K \nabla K)(x; s, y) \cdot \left[- \left(\frac{\nabla K_x}{K_x^2} \right) (x; s, y) - L_{kk}(K(x; s, y); s, y) \nabla K(x; s, y) \right]
\end{aligned} \tag{8.19}$$

from (8.8). Substituting (8.19) back into (8.18), along with (8.14), (8.15), (8.16), and the heat-equation $\frac{\partial \varphi_s}{\partial s} = \frac{1}{2} \Delta \varphi_s$ for the Gaussian kernel $\varphi_s(\cdot)$, we are ready to compute

$$\begin{aligned}
Q_s &- \frac{1}{2} \left[\Delta Q - \frac{\|\nabla Q_x\|^2}{Q_{xx}} \right] \\
&= \frac{K K_s}{K_x} - \frac{1}{2} \left[\frac{\|\nabla K\|^2}{K_x} + \frac{K \Delta K}{K_x} - \frac{K \nabla K \cdot \nabla K_x}{K_x^2} - \frac{K \nabla K \cdot \nabla K_x}{K_x^2} \right. \\
&\quad \left. - K \|\nabla K\|^2 L_{kk}(K(x, s, y), s, y) - \frac{\|\nabla K\|^2}{K_x} \right] \\
&= K \left[\frac{K_s}{K_x} - \frac{1}{2} \frac{\Delta K}{K_x} + \frac{\nabla K \cdot \nabla K_x}{K_x^2} + \frac{1}{2} \|\nabla K\|^2 L_{kk}(K(x, s, y), s, y) \right] = 0,
\end{aligned} \tag{8.20}$$

according to the equation (8.12). In other words, the function $Q(s, x, y)$ satisfies the differential equation (4.12). Along with the identity (4.31), it is straightforward to check that (4.21) holds by the definition (4.18) and (4.33). Thus from (4.20), we have (4.14). On the other hand, differentiating (4.31) with respect to y , we obtain

$$K_x(\mathcal{X}(s, y); s, y) \cdot \nabla \mathcal{X}(s, y) + (\nabla K)(\mathcal{X}(s, y); s, y) = 0.$$

From (8.14) and (8.15), this gives

$$\left(\frac{\nabla Q_x}{Q_{xx}} \right) (s, \mathcal{X}(s, y), y) = \left(\frac{\nabla K}{K_x} \right) (\mathcal{X}(s, y); s, y) = -\nabla \mathcal{X}(s, y),$$

that is, the equality (4.16). Now (4.15) is a straightforward consequence of (3.21) and (3.25). Our proof is complete. $\square$

References

Browne, S. & Whitt, W. (1996) Portfolio choice and the Bayesian Kelly criterion. *Adv. Applied Probability* **28**, 1145–76.

- Cox, J. & Huang, C.F. (1989) Optimal consumption and portfolio policies when asset prices follow a diffusion process. *J. Econom. Theory* **49**, 33–83.
- Cvitanović, J. & Karatzas, I. (1992) Convex duality in constrained portfolio optimization. *Annals of Applied Probability* **2**, 767–818.
- Detemple, J.B. (1986) Asset pricing in a production economy with incomplete information. *J. Finance* **41**, 383–91.
- Dothan, M.U. & Feldman, D. (1986) Equilibrium interest rates and multiperiod bonds in a partially observable economy. *J. Finance* **41**, 369–82.
- Elliott, R.J. (1982) *Stochastic Calculus and Applications*. Springer-Verlag, New York.
- Fleming, W.H. & Rishel, R.W. (1975) *Deterministic and Stochastic Optimal Control*. Springer-Verlag, New York.
- Fleming, W.H. & Soner, H.M. (1993) *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, New York.
- Genotte, G. (1986) Optimal portfolio choice under incomplete information. *J. Finance* **41**, 733–46.
- He, H. & Pearson, N.D. (1991) Consumption and portfolio with incomplete markets and short-sale constraints: the finite-dimensional case. *Math. Finance* **1**, 1–10.
- Kallianpur, G. (1980) *Stochastic Filtering Theory*. Springer-Verlag, New York.
- Karatzas, I. (1997) Adaptive control of a diffusion to a goal and a parabolic Monge–Ampère-type equation. *Asian J. Math.* **1**, 324–41.
- Karatzas, I., Lehoczky, J.P. & Shreve, S.E. (1987) Optimal portfolio and consumption decisions for a “small investor” on a finite horizon. *SIAM J. Control & Optimization* **25**, 1157–586.
- Karatzas, I., Lehoczky, J.P., Shreve, S.E. & Xu, G.L. (1991) Martingale and duality methods for utility maximization in an incomplete market. *SIAM J. Control & Optimization* **29**, 702–30.
- Karatzas, I. & Shreve, S.E. (1991) *Brownian Motion and Stochastic Calculus*. Second Edition, Springer-Verlag, New York.
- Karatzas, I. & Shreve, S.E. (1998) *Methods of Mathematical Finance*. Springer-Verlag, New York.
- Karatzas, I. & Xue, X. (1991) A note on utility maximization under partial observations. *Math. Finance* **1** 57–70.
- Kuwana, Y. (1995) Certainty equivalence and logarithmic utilities in consumption/investment problems. *Math. Finance* **5**, 297–310.
- Lakner, P. (1995) Utility maximization with partial information. *Stochastic Processes & Applications* **56**, 247–73.
- Lakner, P. (1998) Optimal trading strategy for an investor: the case of partial information. *Stochastic Processes & Applications* **76**, 77–97.
- Merton, R.C. (1971) Optimum consumption and portfolio rules in a continuous-time model. *J. Econom. Theory* **3**, 373–413; Erratum, *J. Econom. Theory* **6**, 213–4.
- Pliska, S.R. (1986) A stochastic calculus model of continuous trading: optimal portfolios. *Math. Oper. Research* **11**, 371–82.
- Rishel, R. (1999) Optimal portfolio management with partial observations and power utility function. In *Stochastic Analysis, Control, Optimization and Applications: Volume in Honor of W.H. Fleming* (W. McEneaney, G. Yin & Q. Zhang, Eds.), 605–20. Birkhäuser, Basel and Boston.
- Rockafellar, T. (1970) *Convex Analysis*. Princeton University Press, N.J.
- Rogers, L.C.G. & Williams, D. (1987) *Diffusions, Markov Processes and Martingales*. J. Wiley & Sons, Chichester and New York.
- Spivak, G. (1998) *Maximizing the probability of perfect hedge*. Doctoral Dissertation,

Columbia University.

Zohar, G. (1999) *Dynamic portfolio optimization in the case of partially observed drift process*. Preprint, Columbia University.