HANDBOOKS IN
OPERATIONS RESEARCH AND
MANAGEMENT SCIENCE

Volume 15

FINANCIAL ENGINEERING

John R. Birge Vadim Linetsky
Editors

NORTH-HOLLAND

HANDBOOKS IN OPERATIONS RESEARCH
AND MANAGEMENT SCIENCE
VOLUME 15

Handbooks in Operations Research and Management Science

Volume 15

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo
North-Holland is an imprint of Elsevier

Financial Engineering

Edited by

John R. Birge
University of Chicago, IL, USA

Vadim Linetsky
Northwestern University, IL, USA

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo
North-Holland is an imprint of Elsevier

North-Holland is an imprint of Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

First edition 2008

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-444-51781-4
ISSN: 0927-0507

For information on all North-Holland publications
visit our web site at books.elsevier.com

Printed and bound in The Netherlands

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Contents

PART I

Introduction

This page intentionally left blank

Introduction to the Handbook of Financial Engineering

John R. Birge
Graduate School of Business, University of Chicago, USA

Vadim Linetsky
Department of Industrial Engineering and Management Sciences, Northwestern University, USA

Financial engineering (FE) is an interdisciplinary field focusing on applications of mathematical and statistical modeling and computational technology to problems in the financial services industry. According to the report by the National Academy of Engineering (2003),[1] "Financial services are the foundation of a modern economy. They provide mechanisms for assigning value, exchanging payment, and determining and distributing risk, and they provide the essential underpinnings of global economic activity. The industry provides the wherewithal for the capital investment that drives innovation and productivity growth throughout the economy." Important areas of FE include mathematical modeling of market and credit risk, pricing and hedging of derivative securities used to manage risk, asset allocation and portfolio management.

Market risk is a risk of adverse changes in prices or rates, such as interest rates, foreign exchange rates, stock prices, and commodity and energy prices. Credit risk is a risk of default on a bond, loan, lease, pension or any other type of financial obligation. Modern derivatives markets can be viewed as a global marketplace for financial risks. The function of derivative markets is to facilitate financial risk transfer from risk reducers (hedgers) to risk takers (investors). Organizations wishing to reduce their risk exposure to a particular type of financial risk, such as the risk of increasing commodity and energy prices that will make future production more expensive or the risk of increasing interest rates that will make future financing more expensive, can offset those risks by entering into financial contracts that act as insurance, protecting the company against adverse market events. While the hedger comes to the derivatives market to reduce its risk, the counterparty who takes the other side

[1] National Academy of Engineering, The Impact of Academic Research on Industrial Performance, National Academies Press, Washington, DC, 2003, http://www.nap.edu/books/0309089735/html.

of the contract comes to the market to invest in that risk and expects to be adequately compensated for taking the risk. We can thus talk about buying and selling financial risks.

Global derivatives markets have experienced remarkable growth over the past several decades. According to a recent survey by the Bank for International Settlement in Basel (2006),[2] the aggregate size of the global derivatives markets went from about $50 trillion in notional amounts in 1995 to $343 trillion in notional amounts by the end of 2005 ($283 trillion in notional amounts in over-the-counter derivatives contracts and $58 trillion in futures and options traded on derivatives exchanges worldwide). Major segments of the global derivatives markets include interest rate derivatives, currency derivatives, equity derivatives, commodity and energy derivatives, and credit derivatives.

A *derivative* is a financial contract between two parties that specifies conditions, in particular, dates and the resulting values of underlying variables, under which payments or *payoffs* are to be made between parties (payments can be either in the form of cash or delivery of some specified asset). Call and put options are basic examples of derivatives used to manage market risk. A *call option* is a contract that gives its holder the right to buy some specified quantity of an underlying asset (for example, a fixed number of shares of stock of a particular company or a fixed amount of a commodity) at a predetermined price (called the *strike price*) on or before a specified date in the future (option *expiration*). A *put option* is a contract that gives its holder the right to sell some specified quantity of an underlying asset at a predetermined price on or before expiration. The holder of the option contract locks in the price for future purchase (in the case of call options) or future sale (in the case of put options), thus eliminating any price uncertainty or risk, at the cost of paying the *premium* to purchase the option. The situation is analogous to insurance contracts that pay pre-agreed amounts in the event of fire, flood, car accident, etc. In financial options, the payments are based on financial market moves (and credit events in the case of credit derivatives). Just as in the insurance industry the key problem is to determine the insurance premium to charge for a policy based on actuarial assessments of event probabilities, the option-pricing problem is to determine the premium or option price based on a stochastic model of the underlying financial variables.

Portfolio optimization problems constitute another major class of important problems in financial engineering. Portfolio optimization problems occur throughout the financial services industry as pension funds, mutual funds, insurance companies, university and foundation endowments, and individual investors all face the fundamental problem of allocating their capital across different securities in order to generate investment returns sufficient to achieve a particular goal, such as meeting future pension liabilities. These problems are

[2] Bank for International Settlement Quarterly Review, June 2006, pp. A103–A108, http://www.bis.org/publ/qtrpdf/r_qa0606.pdf.

often very complex owing to their dynamic and stochastic nature, their high dimensionality, and the complexity of real-world constraints.

The remarkable growth of financial markets over the past decades has been accompanied by an equally remarkable explosion in financial engineering research. The goals of financial engineering research are to develop empirically realistic stochastic models describing dynamics of financial risk variables, such as asset prices, foreign exchange rates, and interest rates, and to develop analytical, computational and statistical methods and tools to implement the models and employ them to evaluate financial products used to manage risk and to optimally allocate investment funds to meet financial goals. As financial models are stochastic, probability theory and stochastic processes play a central role in financial engineering. Furthermore, in order to implement financial models, a wide variety of analytical and computational tools are used, including Monte Carlo simulation, numerical PDE methods, stochastic dynamic programming, Fourier methods, spectral methods, etc.

The Handbook is organized in six parts: Introduction, Derivative Securities: Models and Methods, Interest Rate and Credit Risk Models and Derivatives, Incomplete Markets, Risk Management, and Portfolio Optimization. This division is somewhat artificial, as many chapters are equally relevant for several or even all of these areas. Nevertheless, this structure provides an overview of the main areas of the field of financial engineering.

A working knowledge of probability theory and stochastic processes is a prerequisite to reading many of the chapters in the Handbook. Karatzas and Shreve (1991) and Revuz and Yor (1999) are standard references on Brownian motion and continuous martingales. Jacod and Shiryaev (2002) and Protter (2005) are standard references on semimartingale processes with jumps. Shreve (2004) and Klebaner (2005) provide excellent introductions to stochastic calculus for finance at a less demanding technical level. For the financial background at the practical level, excellent overviews of derivatives markets and financial risk management can be found in Hull (2005) and McDonald (2005). Key texts on asset pricing theory include Bjork (2004), Duffie (2001), Jeanblanc et al. (2007), and Karatzas and Shreve (2001). These monographs also contain extensive bibliographies.

In Chapter 1 "A Partial Introduction to Financial Asset Pricing Theory," Robert Jarrow and Philip Protter present a concise introduction to Mathematical Finance theory. The reader is first introduced to derivative securities and the fundamental financial concept of arbitrage in the binomial framework. The core asset pricing theory is then developed in the general semimartingale framework, assuming prices of risky assets follow semimartingale processes. The general fundamental theorems of asset pricing are formulated and illustrated on important examples. In particular, the special case when the risky asset price process is a Markov process is treated in detail, the celebrated Black–Scholes–Merton model is derived, and a variety of results on pricing European- and American-style options and more complex derivative securi-

ties are presented. This chapter summarizes the core of Mathematical Finance theory and is an essential reading.

Part II "Derivative Securities: Models and Methods," contains chapters on a range of topics in derivatives modeling and pricing. The first three chapters survey several important classes of stochastic models used in derivatives modeling. In Chapter 2 "Jump-Diffusion Models," Steven Kou surveys recent developments in option pricing in jump-diffusion models. The chapter discusses empirical evidence of jumps in financial variables and surveys analytical and numerical methods for the pricing of European, American, barrier, and lookback options in jump-diffusion models, with particular attention given to the jump-diffusion model with a double-exponential jump size distribution due to its analytical tractability.

In Chapter 3 "Modeling Financial Security Returns Using Levy Processes," Liuren Wu surveys a class of models based on time-changed Levy processes. Applying stochastic time changes to Levy processes randomizes the clock on which the process runs, thus generating stochastic volatility. If the characteristic exponent of the underlying Levy process and the Laplace transform of the time change process are known in closed form, then the pricing of options can be accomplished by inverting the Fourier transform, which can be done efficiently using the fast Fourier transform (FFT) algorithm. The combination of this analytical and computational tractability and the richness of possible process behaviors (continuous dynamics, as well as jumps of finite activity or infinite activity) make this class of models attractive for a wide range of financial engineering applications. This chapter surveys both the theory and empirical results.

In Chapter 4 "Pricing with Wishart Risk Factors," Christian Gourieroux and Razvan Sufana survey asset pricing based on risk factors that follow a Wishart process. The class of Wishart models can be thought of as multi-factor extensions of affine stochastic volatility models, which model a stochastic variance-covariance matrix as a matrix-valued stochastic process. As for the standard affine processes, the conditional Laplace transforms can be derived in closed form for Wishart processes. This chapter surveys Wishart processes and their applications to building a wide range of multi-variate models of asset prices with stochastic volatilities and correlations, multi-factor interest rate models, and credit risk models, both in discrete and in continuous time.

In Chapter 5 "Volatility," Federico Bandi and Jeff Russell survey the state of the literature on estimating asset price volatility. They provide a unified framework to understand recent advances in volatility estimation by virtue of microstructure noise contaminated asset price data and transaction cost evaluation. The emphasis is on recently proposed identification procedures that rely on asset price data sampled at high frequency. Volatility is the key factor that determines option prices, and, as such, better understanding of volatility is of key interest in options pricing.

In Chapter 6 "Spectral Methods in Derivatives Pricing," Vadim Linetsky surveys a problem of valuing a (possibly defaultable) derivative asset contin-

gent on the underlying economic state modeled as a Markov process. To gain analytical and computational tractability both in order to estimate the model from empirical data and to compute the prices of derivative assets, financial models in applications are often Markovian. In applications, it is important to have a tool kit of analytically tractable Markov processes with known transition semigroups that lead to closed-form expressions for prices of derivative assets. The spectral expansion method is a powerful approach to generate analytical solutions for Markovian problems. This chapter surveys the spectral method in general, as well as those classes of Markov processes for which the spectral representation can be obtained in closed form, thus generating closed form solutions to Markovian derivative pricing problems.

When underlying financial variables follow a Markov jump-diffusion process, the value function of a derivative security satisfies a partial integro-differential equation (PIDE) for European-style exercise or a partial integro-differential variational inequality (PIDVI) for American-style exercise. Unless the Markov process has a special structure (as discussed in Chapter 6), analytical solutions are generally not available, and it is necessary to solve the PIDE or the PIDVI numerically. In Chapter 7 "Variational Methods in Derivatives Pricing," Liming Feng, Pavlo Kovalov, Vadim Linetsky and Michael Marcozzi survey a computational method for the valuation of options in jump-diffusion models based on converting the PIDE or PIDVI to a variational (weak) form, discretizing the weak formulation spatially by the Galerkin finite element method to obtain a system of ODEs, and integrating the resulting system of ODEs in time.

In Chapter 8 "Discrete Path-Dependent Options," Steven Kou surveys recent advances in the development of methods to price discrete path-dependent options, such as discrete barrier and lookback options that sample the minimum or maximum of the asset price process at discrete time intervals, including discrete barrier and lookback options. A wide array of option pricing methods are surveyed, including convolution methods, asymptotic expansion methods, and methods based on Laplace, Hilbert and fast Gauss transforms.

Part III surveys interest rate and credit risk models and derivatives. In Chapter 9 "Topics in Interest Rate Theory" Tomas Bjork surveys modern interest rate theory. The chapter surveys both the classical material on the Heath–Jarrow–Morton forward rate modeling framework and on the LIBOR market models popular in market practice, as well as a range of recent advances in the interest rate modeling literature, including the geometric interest rate theory (issues of consistency and existence of finite-dimensional state space realizations), and potentials and positive interest models.

Chapters 10 and 11 survey the state-of-the-art in modeling portfolio credit risk and multi-name credit derivatives. In Chapter 10 "Computational Aspects of Credit Risk," Paul Glasserman surveys modeling and computational issues associated with portfolio credit risk. A particular focus is on the problem of calculating the loss distribution of a portfolio of credit risky assets, such as corporate bonds or loans. The chapter surveys models of dependence, including

structural credit risk models, copula models, the mixed Poisson model, and associated computational techniques, including recursive convolution, transform inversion, saddlepoint approximation, and importance sampling for Monte Carlo simulation.

In Chapter 11 "Valuation of Basket Credit Derivatives in the Credit Migrations Environment," Tomasz Bielecki, Stephane Crepey, Monique Jeanblanc and Marek Rutkowski present methods to value and hedge basket credit derivatives (such as collateralized debt obligations (CDO) tranches and *n*th to default swaps) and portfolios of credit risky debt. The chapter presents methods for modeling dependent credit migrations of obligors among credit classes and, in particular, dependent defaults. The focus is on specific classes of Markovian models for which computations can be carried out.

Part IV surveys incomplete markets theory and applications. In incomplete markets, dynamic hedging and perfect replication of derivative securities break down and derivatives are no longer redundant assets that can be manufactured via dynamic trading in the underlying primary securities. In Chapter 12 "Incomplete Markets," Jeremy Staum surveys, compares and contrasts many proposed approaches to pricing and hedging derivative securities in incomplete markets, from the perspective of an over-the-counter derivatives market maker operating in an incomplete market. The chapter discusses a wide range of methods, including indifference pricing, good deal bounds, marginal pricing, and minimum-distance pricing measures.

In Chapter 13 "Option Pricing: Real and Risk-Neutral Distributions," George Constantinides, Jens Jackwerth, and Stylianos Perrakis examine the pricing of options in incomplete and imperfect markets in which dynamic trading breaks down either because the market is incomplete or because it is imperfect due to trading costs, or both. Market incompleteness renders the risk-neutral probability measure nonunique and allows one to determine option prices only within some lower and upper bounds. Moreover, in the presence of trading costs, the dynamic replicating strategy does not exist. The authors examine modifications of the theory required to accommodate incompleteness and trading costs, survey testable implications of the theory for option prices, and survey empirical evidence in equity options markets.

In Chapter 14 "Total Risk Minimization Using Monte Carlo Simulation," Thomas Coleman, Yuying Li, and Maria-Cristina Patron study options hedging strategies in incomplete markets. While in an incomplete market it is generally impossible to replicate an option exactly, total risk minimization chooses an optimal self-financing strategy that best approximates the option payoff by its terminal value. Total risk minimization is a computationally challenging dynamic stochastic programming problem. This chapter presents computational approaches to tackle this problem.

In Chapter 15 "Queueing Theoretic Approaches to Financial Price Fluctuations," Erhan Bayraktar, Ulrich Horst, and Ronnie Sircar survey recent research on agent-based market microstructure models. These models of financial prices are based on queueing-type models of order flows and are ca-

pable of explaining many stylized features of empirical data, such as herding behavior, volatility clustering, and fat tailed return distributions. In particular, the chapter examines models of investor inertia, providing a link with behavioral finance.

Part V "Risk Management" contains chapters concerned with risk measurement and its application to capital allocation, liquidity risk, and actuarial risk. In Chapter 16 "Economic Credit Capital Allocation and Risk Contributions," Helmut Mausser and Dan Rosen provide a practical overview of risk measurement and management process, and in particular the measurement of economic capital (EC) contributions and their application to capital allocation. EC acts as a buffer for financial institutions to absorb large unexpected losses, thereby protecting depositors and other claim holders. Once the amount of EC has been determined, it must be allocated among the various components of the portfolio (e.g., business units, obligors, individual transactions). This chapter provides an overview of the process of risk measurement, its statistical and computational challenges, and its application to the process of risk management and capital allocation for financial institutions.

In Chapter 17 "Liquidity Risk and Option Pricing Theory," Robert Jarrow and Phillip Protter survey recent research advances in modeling liquidity risk and including it into asset pricing theory. Classical asset pricing theory assumes that investors' trades have no impact on the prices paid or received. In reality, there is a quantity impact on prices. The authors show how to extend the classical arbitrage pricing theory and, in particular, the fundamental theorems of asset pricing, to include liquidity risk. This is accomplished by studying an economy where the security's price depends on the trade size. An analysis of the theory and applications to market data are presented.

In Chapter 18 "Financial Engineering: Applications in Insurance," Phelim Boyle and Mary Hardy provide an introduction to the insurance area, the oldest branch of risk management, and survey financial engineering applications in insurance. The authors compare the actuarial and financial engineering approaches to risk assessment and focus on the life insurance applications in particular. Life insurance products often include an embedded investment component, and thus require the merging of actuarial and financial risk management tools of analysis.

Part VI is devoted to portfolio optimization. In Chapter 19 "Dynamic Portfolio Choice and Risk Aversion," Costis Skiadas surveys optimal consumption and portfolio choice theory, with the emphasis on the modeling of risk aversion given a stochastic investment opportunity set. Dynamic portfolio choice theory was pioneered in Merton's seminal work, who assumed that the investor maximizes time-additive expected utility and approached the problem using the Hamilton–Jacobi–Bellman equation of optimal control theory. This chapter presents a modern exposition of dynamic portfolio choice theory from a more advanced perspective of recursive utility. The mathematical tools include backward stochastic differential equations (BSDE) and forward–backward stochastic differential equations (FBSDE).

In Chapter 20 "Optimization Methods in Dynamic Portfolio Management," John Birge describes optimization algorithms and approximations that apply to dynamic discrete-time portfolio models including consumption-investment problems, asset-liability management, and dynamic hedging policy design. The chapter develops an overall structure to the many methods that have been proposed by interpreting them in terms of the form of approximation used to obtain tractable models and solutions. The chapter includes the relevant algorithms associated with the approximations and the role that portfolio problem structure plays in enabling efficient implementation.

In Chapter 21 "Simulation Methods for Optimal Portfolios," Jerome Detemple, Rene Garcia and Marcel Rindisbacher survey and compare Monte Carlo simulation methods that have recently been proposed for the computation of optimal portfolio policies. Monte Carlo simulation is the approach of choice for high-dimensional problems with large number of underlying variables. Simulation methods have recently emerged as natural candidates for the numerical implementation of optimal portfolio rules in high-dimensional portfolio choice models. The approaches surveyed include the Monte Carlo Malliavin derivative method, the Monte Carlo covariation method, the Monte Carlo regression method, and the Monte Carlo finite difference method. The mathematical tools include Malliavin's stochastic calculus of variations, a brief survey of which is included in the chapter.

In Chapter 22 "Duality Theory and Approximate Dynamic Programming for Pricing American Options and Portfolio Optimization," Martin Haugh and Leonid Kogan describe how duality and approximate dynamic programming can be applied to construct approximate solutions to American option pricing and portfolio optimization problems when the underlying state space is high-dimensional. While it has long been recognized that simulation is an indispensable tool in financial engineering, it is only recently that simulation has begun to play an important role in control problems in financial engineering. This chapter surveys recent advances in applying simulation to solve optimal stopping and portfolio optimization problems.

In Chapter 23 "Asset Allocation with Multivariate Non-Gaussian Returns," Dilip Madan and Ju-Yi Yen consider a problem of optimal investment in assets with non-Gaussian returns. They present and back test an asset allocation procedure that accounts for higher moments in investment returns. The procedure is made computationally efficient by employing a signal processing technique known as independent component analysis (ICA) to identify long-tailed independent components in the vector of asset returns. The multivariate portfolio allocation problem is then reduced to univariate problems of component investment. They further assume that the ICs follow the variance gamma (VG) Levy processes and build a multivariate VG portfolio and analyze empirical results of the optimal investment strategy in this setting and compare it with the classical mean–variance Gaussian setting.

In Chapter 24 "Large Deviation Techniques and Financial Applications," Phelim Boyle, Shui Feng and Weidong Tian survey recent applications of large

deviation techniques in portfolio management (establishing portfolio selection criteria, construction of performance indexes), risk management (estimation of large credit portfolio losses that occur in the tail of the distribution), Monte Carlo simulation to better simulate rare events for risk management and asset pricing, and incomplete markets models (estimation of the distance of an incomplete model to a benchmark complete model). A brief survey of the mathematics of large deviations is included in the chapter.

A number of important topics that have recently been extensively surveyed elsewhere were not included in the Handbook. Statistical estimation of stochastic models in finance is an important area that has received limited attention in this volume, with the exception of the focused chapter on volatility. Recent advances in this area are surveyed in the forthcoming Handbook of Financial Econometrics edited by Ait-Sahalia and Hansen (2007). In the coverage of credit risk the Handbook is limited to surveying recent advances in multi-name credit portfolios and derivatives in Chapters 10 and 11, leaving out single-name credit models. The latter have recently been extensively surveyed in monographs Bielecki and Rutkowski (2002), Duffie and Singleton (2003), and Lando (2004). The coverage of Monte Carlo simulation methods is limited to applications to multi-name credit portfolios in Chapter 10, to hedging in incomplete markets in Chapter 14, and to portfolio optimization in Chapters 21 and 22. Monte Carlo simulation applications in derivatives valuation have recently been surveyed by Glasserman (2004). Our coverage of risk measurement and risk management is limited to Chapters 16, 17 and 18 on economic capital allocation, liquidity risk, and insurance risk, respectively. We refer the reader to the recently published monograph McNeil et al. (2005) for extensive treatments of Value-at-Risk and related topics. Modeling energy and commodity markets and derivatives is an important area of financial engineering not covered in the Handbook. We refer the reader to the recent monographs by Eydeland and Wolyniec (2002) and Geman (2005) for extensive surveys of energy and commodity markets.

References

Ait-Sahalia, Y., Hansen, L.P. (Eds.) (2007). *Handbook of Financial Econometrics*. Elsevier, Amsterdam, in press.

Bielecki, T., Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*. Springer.

Bjork, T. (2004). *Arbitrage Theory in Continuous Time*, second ed. Oxford University Press, Oxford, UK.

Duffie, D. (2001). *Dynamic Asset Pricing Theory*, third ed. Princeton University Press, Princeton, NJ.

Duffie, D., Singleton, K. (2003). *Credit Risk*. Princeton University Press, Princeton, NJ.

Eydeland, A., Wolyniec, K. (2002). *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*. John Wiley & Sons, New Jersey.

Jacod, J., Shiryaev, A.N. (2002). *Limit Theorems for Stochastic Processes*, second ed. Springer.

Jeanblanc, M., Yor, M., Chesney, M. (2007). *Mathematical Methods for Financial Markets*. Springer, in press.

Geman, H. (2005). *Commodities and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals, and Energy*. John Wiley & Sons, New Jersey.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.
Hull, J. (2005). *Options, Futures, and Other Derivatives*, sixth ed. Prentice Hall.
Karatzas, I., Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*, second ed. Springer.
Karatzas, I., Shreve, S.E. (2001). *Methods of Mathematical Finance*. Springer.
Klebaner, F.C. (2005). *Introduction to Stochastic Calculus with Applications*, second ed. Imperial College Press.
Lando, D. (2004). *Credit Risk Modeling*. Princeton University Press, Princeton, NJ.
McDonald, R.L. (2005). *Derivatives Markets*, second ed. Addison–Wesley.
McNeil, A.J., Frey, R., Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, Princeton, NJ.
Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, second ed. Springer.
Revuz, D., Yor, M. (1999). *Continuous Martingales and Brownian Motion*, third ed. Springer.
Shreve, S.E. (2004). *Stochastic Calculus for Finance II: Continuous-time Models*. Springer.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15001-5

Chapter 1

An Introduction to Financial Asset Pricing

Robert A. Jarrow

Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, USA
E-mail: raj15@cornell.edu

Philip Protter[†]

ORIE – 219 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA
E-mail: pep4@cornell.edu

Abstract

We present an introduction to Mathematical Finance Theory, covering the basic issues as well as some selected special topics.

1 Introduction

Stock markets date back to at least 1531, when one was started in Antwerp, Belgium.[1] Today there are over 150 stock exchanges (see Wall Street Journal, 2000). The mathematical modeling of such markets however, came hundreds of years after Antwerp, and it was embroiled in controversy at its beginnings. The first attempt known to the authors to model the stock market using probability is due to L. Bachelier in Paris about 1900. Bachelier's model was his thesis, and it met with disfavor in the Paris mathematics community, mostly because the topic was not thought worthy of study. Nevertheless we now realize that Bachelier essentially modeled Brownian motion five years before the 1905 paper of Einstein [albeit twenty years after T.N. Thiele of Copenhagen (Hald, 1981)] and of course decades before Kolmogorov gave mathematical legitimacy to the subject of probability theory. Poincaré was hostile to Bachelier's thesis, remarking that his thesis topic was "somewhat remote from those our candidates are in the habit of treating" and Bachelier ended up spending his career in Besançon, far from the French capital. His work was then ignored and forgotten for some time.

[†] Supported in part by NSF grant DMS-0202958 and NSA grant MDA-904-03-1-0092

[1] For a more serious history than this thumbnail sketch, we refer the reader to the recent article (Jarrow and Protter, 2004).

Following work by Cowles, Kendall and Osborne, it was the renowned statistician Savage who re-discovered Bachelier's work in the 1950's, and he alerted Paul Samuelson (see Bernstein, 1992, pp. 22–23). Samuelson further developed Bachelier's model to include stock prices that evolved according to a geometric Brownian motion, and thus (for example) always remained positive. This built on the earlier observations of Cowles and others that it was the increments of the logarithms of the prices that behaved independently.

The development of financial asset pricing theory over the 35 years since Samuelson's 1965 article (Samuelson, 1965) has been intertwined with the development of the theory of stochastic integration. A key breakthrough occurred in the early 1970's when Black, Scholes, and Merton (Black and Scholes, 1973; Merton, 1973) proposed a method to price European options via an explicit formula. In doing this they made use of the Itô stochastic calculus and the Markov property of diffusions in key ways. The work of Black, Merton, and Scholes brought order to a rather chaotic situation, where the previous pricing of options had been done by intuition about ill defined market forces. Shortly after the work of Black, Merton, and Scholes, the theory of stochastic integration for semimartingales (and not just Itô processes) was developed in the 1970's and 1980's, mostly in France, due in large part to P.A. Meyer of Strasbourg and his collaborators. These advances in the theory of stochastic integration were combined with the work of Black, Scholes, and Merton to further advance the theory, by Harrison and Kreps (1979) and Harrison and Pliska (1981) in seminal articles published in 1979 and 1980. In particular they established a connection between complete markets and martingale representation. Much has happened in the intervening two decades, and the subject has attracted the interest and curiosity of a large number of researchers and of course practitioners. The interweaving of finance and stochastic integration continues today. This article has the hope of introducing researchers to the subject at more or less its current state, for the special topics addressed here. We take an abstract approach, attempting to introduce simplifying hypotheses as needed, and we signal when we do so. In this way it is hoped that the reader can see the underlying structure of the theory.

The subject is much larger than the topics of this article, and there are several books that treat the subject in some detail (e.g., Duffie, 2001; Karatzas and Shreve, 1998; Musiela and Rutkowski, 1997; Shiryaev, 1999), including the new lovely book by Shreve (2004). Indeed, the reader is sometimes referred to books such as (Duffie, 2001) to find more details for certain topics. Otherwise references are provided for the relevant papers.

2 Introduction to derivatives and arbitrage

Let $S = (S_t)_{0 \leqslant t \leqslant T}$ represent the (nonnegative) price process of a risky asset (e.g., the price of a stock, a commodity such as "pork bellies," a currency

exchange rate, etc.). The present is often thought of as time $t = 0$. One is interested in the unknown price at some future time T, and thus S_T constitutes a "risk." For example, if an American company contracts at time $t = 0$ to deliver machine parts to Germany at time T, then the unknown price of Euros at time T (in dollars) constitutes a risk for that company. In order to reduce this risk, one may use "derivatives": one can purchase – at time $t = 0$ – the right to buy Euros at time T at a price that is fixed at time 0, and which is called the "strike price." If the price of Euros is higher at time T, then one exercises this right to buy the Euros, and the risk is removed. This is one example of a derivative, called a *call option*.

A *derivative* is any financial security whose value is derived from the price of another asset, financial security, or commodity. For example, the call option just described is a derivative because its value is derived from the value of the underlying Euro. In fact, almost all traded financial securities can be viewed as derivatives.[2] Returning to the *call option* with strike price K, its payoff at time T can be represented mathematically as

$$C = (S_T - K)^+$$

where $x^+ = \max(x, 0)$. Analogously, the payoff to a *put option* with strike price K at time T is

$$P = (K - S_T)^+$$

and this corresponds to the right to *sell* the security at price K at time T. These are two simple examples of derivatives, called a *European call option* and *European put option*, respectively. They are clearly related, and we have

$$S_T - K = (S_T - K)^+ - (K - S_T)^+.$$

This simple equality leads to a relationship between the price of a call option and the price of a put option known as *put–call parity*. We return to this in Section 3.7.

We can also use these two simple options as building blocks for more complicated derivatives. For example, if

$$V = \max(K, S_T)$$

then

$$V = S_T + (K - S_T)^+ = K + (S_T - K)^+.$$

[2] A fun exercise is to try to think of a financial security whose value does not depend on the price of some other asset or commodity. An example is a precious metal itself, like gold, trading as a commodity. But, gold stocks are a derivative as well as gold futures!

More generally, if $f : \mathbb{R}_+ \to \mathbb{R}_+$ is convex then we can use the well-known representation

$$f(x) = f(0) + f'_+(0)x + \int_0^\infty (x - y)^+ \mu(\mathrm{d}y), \tag{1}$$

where $f'_+(x)$ is the right continuous version of the (mathematical) derivative of f, and μ is a positive measure on $\mathbb{R}$ with $\mu = f''$, where the mathematical derivative is in the generalized function sense. In this case if

$$V = f(S_T)$$

is our financial derivative, then V is effectively a portfolio consisting of a continuum of European call options, using (1) (see Brown and Ross, 1991):

$$V = f(0) + f'_+(0)S_T + \int_0^\infty (S_T - K)^+ \mu\,(\mathrm{d}K).$$

For the derivatives discussed so far, the derivative's time T value is a random variable of the form $V = f(S_T)$, that is, a function of the value of S at one fixed and prescribed time T. One can also consider derivatives of the form

$$V = F(S)_T = F(S_t; 0 \leqslant t \leqslant T)$$

which are functionals of the paths of S. For example if S has càdlàg paths (càdlàg is a French acronym for "right continuous with left limits") then $F : D \to \mathbb{R}_+$, where D is the space of functions $f : [0, T] \to \mathbb{R}_+$ which are right continuous with left limits.

If the derivative's value depends on a decision of its holder at only the expiration time T, then they are considered to be of the *European type*, although their analysis for pricing and hedging is more difficult than for simple European call and put options. The decision in the case of a call or put option is whether to exercise the right to buy or sell, respectively.[3] Hence, such decisions are often referred to as *exercise* decisions.

An *American type derivative* is one in which the holder has a decision to make with respect to the security at any time before or at the expiration time. For example, an *American call option* allows the holder to buy the security at a striking price K not only at time T (as is the case for a European call option), but at any time between times $t = 0$ and time T. (It is this type of option that is listed, for example, in the "Listed Options Quotations" in the Wall Street Journal.) Deciding when to exercise such an option is complicated. A strategy for exercising an American call option can be represented mathematically by

[3] This decision is explicitly represented by the maximum operator in the payoff of the call and put options.

a *stopping rule* τ. (That is, if $\mathbb{F} = (\mathcal{F}_t)_{t\geqslant 0}$ is the underlying filtration of S then $\{\tau \leqslant t\} \in \mathcal{F}_t$ for each $t,\ 0 \leqslant t \leqslant T$.) For a given τ, the American call's payoff at time $\tau(\omega)$ is

$$C(\omega) = \big(S_{\tau(\omega)}(\omega) - K\big)^+.$$

We now turn to the *pricing of derivatives*. Let C be a random variable in $\mathcal{F}_T$ representing the time T payoff to a derivative. Let V_t be its *value* (or price) at time t. What then is V_0? From a traditional point of view based on an analysis of fair (gambling) games, classical probability tells us that[4]

$$V_0 = E\{C\}. \tag{2}$$

One should pay the expected payoff of participating in the gamble. But, one should also discount for the time value of money (the interest forgone or earned) and assuming a fixed spot interest rate r, one would have

$$V_0 = E\left\{\frac{C}{(1+r)^T}\right\} \tag{3}$$

instead of (2). Surprisingly, this value is not correct, because it ignores the impact of risk aversion on the part of the purchaser. For simplicity, we will take $r = 0$ and then show why the obvious price given in (2) does not work (!).

Let us consider a simple binary example. At time $t = 0$, 1 Euro = \$1.15. We assume at time $t = T$ that the Euro will be worth either \$0.75 or \$1.45. Let the probability that it goes up to \$1.45 be p and the probability that it goes down be $1 - p$.

Consider a European call option with exercise price $K = \$1.15$. That is, $C = (S_T - \$1.15)^+$, where $S = (S_t)_{0\leqslant t\leqslant T}$ is the price of one Euro in US dollars. The classical rules for calculating probabilities dating back to Huygens and Bernoulli give a fair price of C as

$$E\{C\} = (1.45 - 1.15)\,p = (0.30)\,p.$$

For example if $p = 1/2$ we get $V_0 = 0.15$.

The *Black–Scholes* method[5] for calculating the option's price, however, is quite different. We first replace p with a new probability p^* that (in the absence of interest rates) makes the security price $S = (S_t)_{t=0,T}$ a martingale. Since this is a two-step process, *we need only to choose p^* so that S has a constant expectation under P^*, the probability measure implied by the choice of p^**. Since

[4] This assumes, implicitly, that there are no intermediate cash flows from holding the derivative security.

[5] The "Black–Scholes method" dates back to the fundamental and seminal articles (Black and Scholes, 1973) and (Merton, 1973) of 1973, where partial differential equations were used; the ideas implicit in that (and subsequent) articles are now referred to as the Black–Scholes methods. More correctly, it should be called the Black–Merton–Scholes method. M.S. Scholes and R. Merton received the Nobel prize in economics for (Black and Scholes, 1973; Merton, 1973), and related work. (F. Black died and was not able to share in the prize.)

$S_0 = 1.15$, we need

$$E^*\{S_T\} = 1.45p^* + (1 - p^*)0.75 = 1.15, \tag{4}$$

where E^* denotes mathematical expectation with respect to the probability measure P^* given by $P^*(\text{Euro} = \$1.45 \text{ at time } T) = p^*$, and $P^*(\text{Euro} = \$0.75 \text{ at time } T) = 1 - p^*$. Solving for p^* gives

$$p^* = 4/7.$$

We get now

$$V_0 = E^*\{C\} = (0.30)p^* = \frac{6}{35} \simeq 0.17.$$

The change from p to p^* seems arbitrary. But, there is an *economics* argument to justify it. This is where the economics concept of *no arbitrage opportunities* changes the usual intuition dating back to the 16th and 17th centuries.

Suppose, for example, at time $t = 0$ you sell the call option, giving the buyer of the option the right to purchase 1 Euro at time T for \$1.15. He then gives you the price $v(C)$ of the option. Again we assume $r = 0$, so there is no cost to borrow money. You can then follow a safety strategy to prepare for the option you sold, as follows (calculations are to two decimal places):

Action at time $t = 0$	*Result*
Sell the option at price $v(C)$	$+v(C)$
Borrow $\$\frac{9}{28}$	$+0.32$
Buy $\frac{3}{7}$ euros at \$1.15	-0.49

The balance at time $t = 0$ *is* $v(C) - 0.17$.

At time T there are two possibilities:

What happens to the euro	*Result*
The euro has risen:	
Option is exercised	-0.30
Sell $\frac{3}{7}$ euros at \$1.45	$+0.62$
Pay back loan	-0.32
End balance:	0
The euro has fallen:	
Option is worthless	0
Sell $\frac{3}{7}$ euros at \$0.75	$+0.32$
Pay back loan	-0.32
End balance:	0

Since the balance at time T is zero in both cases, the balance at time 0 should also be 0; therefore we must have $v(C) = 0.17$. Indeed any price other than $v(C) = 0.17$ would allow either the option seller or buyer to make a sure profit

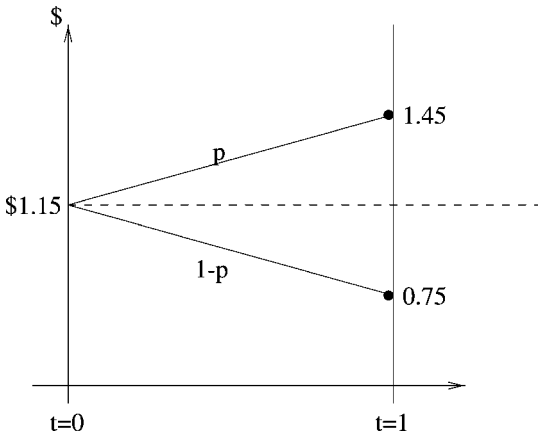


Fig. 1. Binary schematic.

without any risk. Such a sure profit with no risk is called an *arbitrage opportunity* in economics, and it is a standard assumption that such opportunities do not exist. (Of course if they were to exist, market forces would, in theory, quickly eliminate them.)

Thus we see that – at least in the case of this simple example – that the "no arbitrage price" of the derivative C is not $E\{C\}$, but rather it must be $E^*\{C\}$. We emphasize that this is contrary to our standard intuition based on fair games, since P is the probability measure governing the true laws of chance of the security, while P^* is an artificial construct.

Remark 1 *(Heuristic Explanation)*. We offer two comments here. The first is that the change of probability measures from P to P^* is done with the goal of keeping the expectation constant. (See Equation (4).) It is this property of constant expectation of the price process which excludes the possibility of arbitrage opportunities, when the price of the derivative is chosen to be the expectation under P^*. Since one can have many different types of processes with constant expectation, one can ask: what is the connection to martingales? The answer is that a necessary and sufficient condition for a process $M = (M_t)_{t\geqslant 0}$ to be a uniformly martingale is that $E(M_\tau) = E(M_0)$ for every stopping time τ. The key here is that it is required for every stopping time, and not just for fixed times. In words, the price process must have constant expectation at all random times (stopping times) under a measure P^* in order for the expectation of the contingent claim under P^* to be an arbitrage free price of the claim.

The second comment refers to Figure 1 (binary schematic). Intuition tells us that as $p \nearrow 1$, that the price of a call or put option must change, since as it becomes almost certain that the price will go up, the call might be worth less (or more) to the purchaser. And yet our no arbitrage argument tells us that it cannot, and that p^* is fixed for all p, $0 < p < 1$. How can this be? An economics explanation is that if one lets p increase to 1, one is implicitly perverting the

economy. In essence, this perversion of the economy *a fortiori* reflects changes in participants' levels of *risk aversion*. If the price can change to only two prices, and it is near certain to go up, how can we keep the current price fixed at \$1.15? Certainly this change in perceived probabilities should affect the current price too. In order to increase p towards 1 and simultaneously keep the current price fixed at \$1.15, we are forced to assume that people's behavior has changed, and either they are very averse to even a small potential loss (the price going down to \$0.75), or they now value much less the near certain potential price increase to \$1.45.

This simple binary example can do more than illustrate the idea of using the lack of arbitrage to determine a price. We can also use it to approximate some continuous time models for the evolution of an asset's price. We let the time interval become small (Δt), and we let the binomial model already described become a recombinant tree, which moves up or down to a neighboring node at each time "tick" Δt. For an actual time "tick" of interest of length say δ, we can have the price go to 2^n possible values for a given n, by choosing Δt small enough in relation to n and δ. Thus for example if the continuous time process follows geometric Brownian motion:

$$\mathrm{d}S_t = \sigma S_t \,\mathrm{d}B_t + \mu S_t \,\mathrm{d}t$$

(as is often assumed in practice); and if the security price process S has value $S_t = s$, then it will move up or down at the next tick Δt to

$$s\exp(\mu\Delta t + \sigma\sqrt{\Delta t}\,) \quad \text{if up;} \qquad s\exp(\mu\Delta t - \sigma\sqrt{\Delta t}\,) \quad \text{if down;}$$

with p being the probability of going up or down (here take $p = \frac{1}{2}$). Thus for a time t, if $n = \frac{t}{\Delta t}$, we get

$$S_t^n = S_0 \exp\left(\mu t + \sigma\sqrt{t}\left(\frac{2X_n - n}{\sqrt{n}}\right)\right),$$

where X_n counts the number of jumps up. By the central limit theorem S_t^n converges, as n tends to infinity, to a log normal process $S = (S_t)_{t\geqslant 0}$; that is, $\log S_t$ has a normal distribution with mean $\log(S_0 + \mu t)$ and variance $\sigma^2 t$.

Next we use the absence of arbitrage to change p from $\frac{1}{2}$ to p^*. We find p^* by requiring that $E^*\{S_t\} = E^*\{S_0\}$, and we get p^* approximately equal to

$$p^* = \frac{1}{2}\left(1 - \sqrt{\Delta t}\left(\frac{\mu + \frac{1}{2}\sigma^2}{\sigma}\right)\right).$$

Thus under P^*, X_n is still binomial, but now it has mean np^* and variance $np^*(1 - p^*)$. Therefore $\left(\frac{2X_n - n}{\sqrt{n}}\right)$ has mean $-\sqrt{t}(\mu + \frac{1}{2}\sigma^2)/\sigma$ and a variance which converges to 1 asymptotically. The central limit theorem now implies that S_t converges as n tends to infinity to a log normal distribution: $\log S_t$ has

mean $\log S_0 - \frac{1}{2}\sigma^2 t$ and variance $\sigma^2 t$. Thus

$$S_t = S_0 \exp\left(\sigma\sqrt{t}Z - \frac{1}{2}\sigma^2 t\right),$$

where Z is $N(0, 1)$ under P^*. This is known as the "binomial approximation." The binomial approximation can be further used to derive the Black–Scholes equations, by taking limits, leading to simple formulas in the continuous case. (We present these formulas in Section 3.10.) A simple derivation can be found in Cox et al. (1979) or in Duffie (2001, Chapter 12B, pp. 294–299).

3 The core of the theory

3.1 Basic definitions

Throughout this section we will assume that we are given an underlying probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$, where $\mathbb{F} = (\mathcal{F}_t)_{t\geqslant 0}$. We further assume $\mathcal{F}_s \subset \mathcal{F}_t$ if $s < t$; $\mathcal{F}_0$ contains all the P-null sets of $\mathcal{F}$; and also that $\bigcap_{s>t}\mathcal{F}_s \equiv \mathcal{F}_{t+} = \mathcal{F}_t$ by hypothesis. This last property is called the *right continuity of the filtration*. These hypotheses, taken together, are known as the *usual hypotheses*. (When the usual hypotheses hold, one knows that every martingale has a version which is càdlàg, one of the most important consequences of these hypotheses.)

3.2 The price process

We let $S = (S_t)_{t\geqslant 0}$ be a semimartingale[6] which will be the *price process* of a risky security. For simplicity, after the initial purchase or sale, we assume that the security has no cash flows associated with it (for example, if the security is a common stock, we assume that there are no dividends paid). This assumption is easily relaxed, but its relaxation unnecessarily complicates the notation and explanation, so we leave it to outside references.

3.3 Spot interest rates

Let r be a fixed spot rate of interest. If one invests 1 dollar at rate r for one year, at the end of the year one has $1 + r$ dollars. If interest is paid at n evenly spaced times during the year and compounded, then at the end of the year one has $(1 + \frac{r}{n})^n$. This leads to the notion of an *interest rate r compounded*

[6] One *definition of a semimartingale* is a process S that has a decomposition $S = M + A$, with M a local martingale and A an adapted process with càdlàg paths of finite variation on compacts. (See Protter, 2005.)

continuously:

$$\lim_{n\to\infty}\left(1+\frac{r}{n}\right)^n = e^r$$

or, for a fraction t of the year, one has $\$e^{rt}$ after t units of time for a spot interest rate r compounded continuously.

We define

$$R(t) = e^{rt};$$

then R satisfies the ODE (ODE abbreviates ordinary differential equation)

$$\mathrm{d}R(t) = rR(t)\,\mathrm{d}t; \quad R(0) = 1. \tag{5}$$

Using the ODE (5) as a basis for interest rates, one can treat a variable interest rate $r(t)$ as follows: ($r(t)$ can be random: that is $r(t) = r(t, \omega)$)[7]:

$$\mathrm{d}R(t) = r(t)R(t)\,\mathrm{d}t; \quad R(0) = 1 \tag{6}$$

and solving yields $R(t) = \exp(\int_0^t r(s)\,\mathrm{d}s)$. We think of the interest rate process $R(t)$ as the *time t value of a money market account*.

3.4 Trading strategies and portfolios

We will assume as given a risky asset with price process S and a money market account with price process R. Let $(a_t)_{t\geqslant 0}$ and $(b_t)_{t\geqslant 0}$ be our *time t holdings* in the security and the bond, respectively.

We call our holdings of S and R our *portfolio*. Note that for the model to make sense, we must have both the risky asset and the money market account present. When we receive money through the sale of risky assets, we place the cash in the money market account; and when we purchase risky assets, we use the cash from the money market account to pay for the expenditure. The money market account is allowed to have a negative balance.

Definition 1. The *value at time t*[8] *of a portfolio* (a, b) is

$$V_t(a, b) = a_t S_t + b_t R_t.$$

[7] An example is to take $r(t)$ to be a diffusion; one can then make appropriate hypotheses on the diffusion to model the behavior of the spot interest rate.

[8] This concept of value is a commonly used approximation. If one were to liquidate one's risky assets at time t all at once to realize this "value," one would find less money in the savings account, due to liquidity and transaction costs. For simplicity, we are assuming there are no liquidity and transaction costs. Such an assumption is not necessary, however, and we recommend the interested reader to Jarrow and Protter (2007) in this volume.

Now we have our first problem. Later we will want to change probabilities so that $V = (V_t(a,b))_{t\geqslant 0}$ is a martingale. One usually takes the right continuous versions of a martingale, so we want the right side of (4) to be at least càdlàg. Typically this is not a real problem. Even if the process a has no regularity, one can always choose b in such a way that $V_t(a,b)$ is càdlàg.

Let us next define two sigma algebras on the product space $\mathbb{R}_+ \times \Omega$. We recall that we are given an underlying probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$ with $\mathbb{F} = (\mathcal{F}_t)_{t\geqslant 0}$, satisfying the "usual hypotheses."

Definition 2. Let $\mathbb{L}$ denote the space of left continuous processes whose paths have right limits (càglàd), and which are adapted: that is, $H_t \in \mathcal{F}_t$, for $t \geqslant 0$. The *predictable σ-algebra* $\mathcal{P}$ on $\mathbb{R}_+ \times \Omega$ is

$$\mathcal{P} = \sigma\{H\colon\ H \in \mathbb{L}\}.$$

That is $\mathcal{P}$ is the smallest σ-algebra that makes all of $\mathbb{L}$ measurable.

Definition 3. The *optional σ-algebra* $\mathcal{O}$ on $\mathbb{R}_+ \times \Omega$ is

$$\mathcal{O} = \sigma\{H\colon\ H \text{ is càdlàg and adapted}\}.$$

In general we have $\mathcal{P} \subset \mathcal{O}$. In the case where $B = (B_t)_{t\geqslant 0}$ is a standard Wiener process (or "Brownian motion"), and $\mathcal{F}_t^0 = \sigma(B_s; s \leqslant t)$ and $\mathcal{F}_t = \mathcal{F}_t^0 \vee \mathcal{N}$ where $\mathcal{N}$ are the P-null sets of $\mathcal{F}$, then we have $\mathcal{O} = \mathcal{P}$. In general $\mathcal{O}$ and $\mathcal{P}$ are not equal. Indeed if they are equal, then every stopping time is predictable: that is, there are no totally inaccessible stopping times.[9] Since the jump times of (reasonable) Markov processes are totally inaccessible, any model which contains a Markov process with jumps (such as a Poisson Process) will have $\mathcal{P} \subset \mathcal{O}$, where the inclusion is strict.

Remark on filtration issues. The predictable σ-algebra $\mathcal{P}$ is important because it is the natural σ-field for which stochastic integrals are defined. In the special case of Brownian motion one can use the optional σ-algebra (since they are the same). There is a third σ-algebra which is often used, known as

[9] A *totally inaccessible stopping time* is a stopping time that comes with no advance warning: it is a complete surprise. A stopping time T is *totally inaccessible* if whenever there exists a sequence of nondecreasing stopping times $(S_n)_{n\geqslant 1}$ with $\Lambda = \bigcap_{n=1}^{\infty}\{S_n < T\}$, then

$$P\Big(\Big\{w\colon \lim_{n\to\infty} S_n = T\Big\} \cap \Lambda\Big) = 0.$$

A stopping time T is *predictable* if there exists a nondecreasing sequence of stopping times $(S_n)_{n\geqslant 1}$ as above with

$$P\Big(\Big\{w\colon \lim_{n\to\infty} S_n = T\Big\} \cap \Lambda\Big) = 1.$$

Note that the probabilities above need not be only 0 or 1; thus there are in general stopping times which are neither predictable nor totally inaccessible.

the progressively measurable sets, and denoted $\boldsymbol{\pi}$. One has, in general, that $\mathcal{P} \subset \mathcal{O} \subset \boldsymbol{\pi}$; however in practice one gains very little by assuming a process is $\boldsymbol{\pi}$-measurable instead of optional, if – as is the case here – one assumes that the filtration $(\mathcal{F}_t)_{t \geqslant 0}$ is right-continuous (that is $\mathcal{F}_{t+} = \mathcal{F}_t$, all $t \geqslant 0$). The reason is that the primary use of $\boldsymbol{\pi}$ is to show that adapted, right-continuous processes are $\boldsymbol{\pi}$-measurable and in particular that $S_T \in \mathcal{F}_T$ for T a stopping time and S progressive; but such processes are already optional if $(\mathcal{F}_t)_{t \geqslant 0}$ is right continuous. Thus there are essentially no "naturally occurring" examples of progressively measurable processes that are not already optional. An example of such a process, however, is the indicator function $1_G(t)$, where G is described as follows: let $\mathbb{Z} = \{(t, \omega)\colon\ B_t(\omega) = 0\}$. ($B$ is standard Brownian motion.) Then $\mathbb{Z}$ is a perfect (and closed) set on $\mathbb{R}_+$ for almost all ω. For fixed ω, the complement is an open set and hence a countable union of open intervals. $G(\omega)$ denotes the left end-points of these open intervals. One can then show (using the Markov property of B and P.A. Meyer's section theorems) that G is progressively measurable but not optional. In this case note that $1_G(t)$ is zero except for countably many t for each ω, hence $\int 1_G(s)\, \mathrm{d}B_s \equiv 0$. Finally we note that if $a = (a_s)_{s \geqslant 0}$ is progressively measurable, then $\int_0^t a_s\, \mathrm{d}B_s = \int_0^t \dot{a}_s\, \mathrm{d}B_s$, where $\dot{a}$ is the predictable projection of a.[10]

Let us now recall a few details of stochastic integration. First, let S and X be any two càdlàg semimartingales. The integration by parts formula can be used to define the quadratic co-variation of X and S:

$$[X, S]_t = X_t Y_t - \int_0^t X_{s-}\, \mathrm{d}S_s - \int_0^t S_{s-}\, \mathrm{d}X_s.$$

However if a càdlàg, adapted process H is not a semimartingale, one can still give the quadratic co-variation a meaning, by using a limit in probability as the definition. This limit always exists if both H and S are semimartingales:

$$[H, S]_t = \lim_{n \to \infty} \sum_{t_i \in \pi^n[0,t]} (H_{t_{i+1}} - H_{t_i})(S_{t_{i+1}} - S_{t_i}),$$

where $\pi^n[0, t]$ be a sequence of finite partitions of $[0, t]$ with $\lim_{n \to \infty} \mathrm{mesh}(\pi^n) = 0$.

[10] Let H be a bounded, measurable process. (H need not be adapted.) The *predictable projection* of H is the unique predictable process $\dot{H}$ such that

$$\dot{H}_T = E\{H \mid \mathcal{F}_{T-}\} \quad \text{a.s. on } \{T < \infty\}$$

for all predictable stopping times T. Here $\mathcal{F}_{T-} = \sigma\{A \cap \{t < T\};\ A \in \mathcal{F}_t\} \vee \mathcal{F}_0$. For a proof of the existence and uniqueness of $\dot{H}$ see Protter (2005, p. 119).

Henceforth let S be a (càdlàg) semimartingale, and let H be càdlàg and adapted, or alternatively $H \in \mathbb{L}$. Let $H_- = (H_{s-})_{s \geqslant 0}$ denote the left-continuous version of H. (If $H \in \mathbb{L}$, then of course $H = H_-$.) We have:

Theorem 1. *H càdlàg, adapted or $H \in \mathbb{L}$. Then*

$$\lim_{n \to \infty} \sum_{t_i \in \pi^n[0,t]} H_{t_i}(S_{t_{i+1}} - S_{t_i}) = \int_0^t H_{s-} \, \mathrm{d}S_s,$$

with convergence uniform in s on $[0, t]$ in probability.

We remark that it is crucial that we sample H at the left endpoint of the interval $[t_i, t_{i+1}]$. Were we to sample at, say, the right endpoint or the midpoint, then the sums would not converge in general (they converge for example if the quadratic covariation process $[H, S]$ exists); in cases where they do converge, the limit is in general different. Thus while the above theorem gives a pleasing "limit as Riemann sums" interpretation to a stochastic integral, it is not at all a perfect analogy.

The basic idea of the preceding theorem can be extended to bounded predictable processes in a method analogous to the definition of the Lebesgue integral for real-valued functions. Note that

$$\sum_{t_i \in \pi^n[0,t]} H_{t_i}(S_{t_{i+1}} - S_{t_i}) = \int_{0+}^t H_s^n \, \mathrm{d}S_s,$$

where $H_t^n = \sum H_{t_i} 1_{(t_i . t_{i+1}]}$ which is in $\mathbb{L}$; thus these "simple" processes are the building blocks, and since $\sigma(\mathbb{L}) = \mathcal{P}$, it is unreasonable to expect to go beyond $\mathcal{P}$ when defining the stochastic integral.

There is, of course, a maximal space of integrable processes where the stochastic integral is well defined and still gives rise to a semimartingale as the integrated process; without describing it [see any book on stochastic integration such as (Protter, 2005)], we define:

Definition 4. For a semimartingale S we let $L(S)$ denote the space of predictable processes a, where a is integrable with respect to S.

We would like to fix the underlying semimartingale (or vector of semimartingales) S. The process S represents the price process of our risky asset. A way to do that is to introduce the notion of a *model*. We present two versions. The first is the more complete, as it specifies the probability space and the underlying filtration. However it is also cumbersome, and thus we will abbreviate it with the second:

Definition 5. A sextuple $(\Omega, \mathcal{F}, \mathbb{F}, S, L(S), P)$, where $\mathbb{F} = (\mathcal{F}_t)_{t \geqslant 0}$, is called an *asset pricing model*; or more simply, the triple $(S, L(S), P)$ is called a *model*,

where the probability space and σ-algebras are implicit: that is, $(\Omega, \mathcal{F}, \mathbb{F})$ is implicit.

We are now ready for a key definition.

A *trading strategy* in the risky asset is a predictable process $a = (a_t)_{t \geqslant 0}$ with $a \in L(S)$; its economic interpretation is that at time t one holds an amount a_t of the asset. We also remark that it is reasonable that a be predictable: a is the trader's holdings at time t, and this is based on information obtained at times strictly before t, but not t itself. Often one has in concrete situations that a is continuous or at least càdlàg or càglàd (left continuous with right limits). (Indeed, it is difficult to imagine a practical trading strategy with pathological path irregularities.) In the case a is adapted and càglàd, then

$$\int_0^t a_s \, \mathrm{d}S_s = \lim_{n \to \infty} \sum_{t_i \in \pi^n[0,t]} a_{t_i} \Delta_i S,$$

where $\pi^n[0, t]$ is a sequence of partitions of $[0, t]$ with mesh tending to 0 as $n \to \infty$; $\Delta_i S = S_{t_{i+1}} - S_{t_i}$; and with convergence in u.c.p. (uniform in time on compacts and converging in probability). Thus inspired by (1) we let

$$G_t = \int_{0+}^t a_s \, \mathrm{d}S_s$$

and G is called the (*financial*) *gain process generated by* a. A trading strategy in the money market account, $b = (b_t)_{t \geqslant 0}$, is defined in an analogous fashion except that we only require that b is optional and $b \in L(R)$. We will call the pair (a, b), as defined above, a trading strategy.

Definition 6. A trading strategy (a, b) is called *self-financing* if

$$a_t S_t + b_t R_t = a_0 S_0 + b_0 R_0 + \int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s \tag{7}$$

for all $t \geqslant 0$.

Note that the equality (7) above implies that $a_t S_t + b_t R_t$ is càdlàg.

To justify this definition heuristically, let us assume the spot interest rate is constant and zero: that is, $r = 0$ which implies that $R_t = 1$ for all $t \geqslant 0$, a.s. We can do this by the principle of numéraire invariance; see Section 3.6, later in this article. We then have

$$a_t S_t + b_t R_t = a_t S_t + b_t.$$

Assume for the moment that a and b are semimartingales, and as such let us denote them X and Y, respectively.[11] If at time t we change our position in the risky asset, to be self-financing we must change also the amount in our money market account; thus we need to have the equality:

$$(X_{t+\mathrm{d}t} - X_t)S_{t+\mathrm{d}t} = -(Y_{t+\mathrm{d}t} - Y_t),$$

which is algebraically equivalent to

$$(S_{t+\mathrm{d}t} - S_t)(X_{t+\mathrm{d}t} - X_t) + (X_{t+\mathrm{d}t})S_t = -(Y_{t+\mathrm{d}t} - Y_t),$$

which implies in continuous time:

$$S_{t-}\,\mathrm{d}X_t + \mathrm{d}[S, X]_t = -\mathrm{d}Y_t.$$

Using integration by parts, we get

$$X_t S_t - X_{t-}\,\mathrm{d}S_t = -\mathrm{d}Y_t,$$

and integrating yields the desired equality

$$X_t S_t + Y_t = \int_0^t X_{s-}\,\mathrm{d}S_s + X_0 S_0 + Y_0. \tag{8}$$

Finally we drop the assumption that X and Y are semimartingales, and replacing X_- with a and Y with b, respectively, Eq. (8) becomes

$$a_t S_t + b_t R_t = a_0 S_0 + b_0 + \int_0^t a_s\,\mathrm{d}S_s + (b_t - b_0),$$

as we have in Eq. (7).

The next concept is of fundamental importance. An *arbitrage opportunity* is the chance to make a profit *without risk*. The standard way of modeling this mathematically is as follows:

Definition 7. A model is *arbitrage free* if there does not exist a self-financing trading strategy (a, b) such that $V_0(a, b) = 0$, $V_T(a, b) \geqslant 0$, and $P(V_T(a, b) > 0) > 0$.

[11] Since X is assumed to be a semimartingale, it is right continuous, and thus is not in general predictable; hence when it is the integrand of a stochastic integral we need to replace X_s with X_{s-}, which of course denotes the left continuous version of X.

3.5 The fundamental theorem of asset pricing

In Section 2 we saw that with the "no arbitrage" assumption, at least in the case of a very simple example, a reasonable price of a derivative was obtained by taking expectations and changing from the "true" underlying probability measure, P, to an equivalent one, P^*. More formally, under the assumption that $r = 0$, or equivalently that $R_t = 1$ for all t, the price of a derivative C was not $E\{C\}$ as one might expect, but rather $E^*\{C\}$. (If the process R_t is not constant and equal to one, then we consider the expectation of the discounted claim $E^*\{C/R_T\}$.)

The idea underlying the equivalent change of measure was to find a probability P^* that gave the price process S a constant expectation. This simple insight readily generalizes to more complex stochastic processes. In continuous time, a sufficient condition for the price process $S = (S_t)_{t\geqslant 0}$ to have constant expectation is that it be a martingale. That is, if S is a martingale then the function $t \to E\{S_t\}$ is constant. Actually this property is not far from characterizing martingales. A classic theorem from martingale theory is the following (cf., e.g., Protter, 2005):

Theorem 2. *Let $S = (S_t)_{t\geqslant 0}$ be càdlàg and suppose $E\{S_\tau\} = E\{S_0\}$ for any bounded stopping time τ (and of course $E\{|S_\tau|\} < \infty$). Then S is a martingale.*

That is, if we require constant expectation at stopping times (instead of only at fixed times), then S is a martingale.

Based on this simple pricing example and the preceding theorem, one is lead naturally to the following conjecture.

Conjecture. Let S be a price process on a given space $(\Omega, \mathcal{F}, \mathbb{F}, P)$. Then there are no arbitrage opportunities if and only if there exists a probability P^*, equivalent to P, such that S is a martingale under P^*.

The origins of the preceding conjecture can be traced back to Harrison and Kreps (1979) for the case where $\mathcal{F}_T$ is finite, and later to Dalang et al. (1990) for the case where $\mathcal{F}_T$ is infinite, but time is discrete. Before stating a more rigorous theorem [our version is due to Delbaen and Schachermeyer (1994); see also Delbaen and Schachermayer (1998)], let us examine a needed hypothesis.

We need to avoid problems that arise from the classical doubling strategy in gambling. Here a player bets \$1 at a fair bet. If he wins, he stops. If he loses he next bets \$2. Whenever he wins, he stops, and his profit is \$1. If he continues to lose, he continues to play, each time doubling his bet. This strategy leads to a certain gain of \$1 without risk. However, the player needs to be able to tolerate arbitrarily large losses before he gains his certain profit. Of course, no one has such infinite resources to play such a game. Mathematically one can eliminate this type of problem by requiring trading strategies to give martingales that are bounded below by a constant. Thus the player's resources, while they can be

huge, are nevertheless finite and bounded by a nonrandom constant. This leads to the next definition.

Definition 8. Let $\alpha > 0$, and let S be a semimartingale. A predictable trading strategy θ is *α-admissible* if $\theta_0 = 0$, $\int_0^t \theta_s \, \mathrm{d}S_s \geqslant -\alpha$, all $t \geqslant 0$. θ is called *admissible* if there exists $\alpha > 0$ such that θ is α-admissible.

Before we make more definitions, let us recall the basic approach. Suppose θ is an admissible, self-financing trading strategy with $\theta_0 S_0 = 0$ and $\theta_T S_T \geqslant 0$. In the next section we will see that without loss of generality we can neglect the bond or "numéraire" process by a "change of numéraire," so that the self-financing condition reduces to

$$\theta_T S_T = \theta_0 S_0 + \int_0^T \theta_s \, \mathrm{d}S_s.$$

Then if P^* exists such that $\int \theta_s \, \mathrm{d}S_s$ is a martingale, we have

$$E^*\{\theta_T S_T\} = 0 + E^*\left\{ \int_0^T \theta_s \, \mathrm{d}S_s \right\}.$$

In general, if S is continuous then $\int_0^t \theta_s \, \mathrm{d}S_s$ is only a *local* martingale.[12] If S is merely assumed to be a càdlàg semimartingale, then $\int_0^t \theta_s \, \mathrm{d}S_s$ need only be a σ martingale.[13] However if for some reason we do know that it is a true martingale then $E^*\{\int_0^T \theta_s \, \mathrm{d}S_s\} = 0$, whence $E^*\{\theta_T S_T\} = 0$, and since $\theta_T S_T \geqslant 0$ we deduce $\theta_T S_T = 0$, P^* a.s., and since P^* is equivalent to P, we have $\theta_T S_T = 0$ a.s. ($\mathrm{d}P$) as well. This implies no arbitrage exists. The technical part of this argument is to show $\int_0^t \theta_s \, \mathrm{d}S_s$ is a P^* true martingale, and not just a local martingale (see the proof of the Fundamental Theorem that follows). The converse is typically harder: that is, that no arbitrage implies P^* exists. The converse is proved using a version of the Hahn–Banach theorem.

[12] A process M is a *local martingale* if there exists a sequence of stopping times $(T_n)_{n \geqslant 1}$ increasing to ∞ a.s. such that $(M_{t \wedge T_n})_{t \geqslant 0}$ is a martingale for each $n \geqslant 1$.

[13] A process X is a σ martingale if there exists an $\mathbb{R}^d$ valued martingale M and a predictable $\mathbb{R}_+$ valued M-integrable process H such that X is the stochastic integral of H with respect to M. See Protter (2005, pp. 237–239) for more about σ martingales.

Following Delbaen and Schachermayer, we make a sequence of definitions:

$$
\begin{aligned}
K_0 &= \left\{ \int_0^\infty \theta_s \, \mathrm{d}S_s \;\middle|\; \theta \text{ is admissible and } \lim_{t\to\infty} \int_0^t \theta_s \, \mathrm{d}S_s \text{ exists a.s.} \right\} C_0 \\
&= \{\text{all functions dominated by elements of } K_0\} \\
&= K_0 - L^0_+, \text{ where } L^0_+ \text{ are positive, finite random variables,} \\
K &= K_0 \cap L^\infty, \\
C &= C_0 \cap L^\infty, \\
\overline{C} &= \text{the closure of } C \text{ under } L^\infty.
\end{aligned}
$$

Definition 9. A semimartingale price process S satisfies

(i) the *no arbitrage* condition if $\mathbb{C} \cap L^\infty_+ = \{0\}$ (this corresponds to no chance of making a profit without risk);

(ii) the *no free lunch with vanishing risk* condition (NFLVR) if $\overline{\mathbb{C}} \cap L^\infty_+ = \{0\}$, where $\overline{\mathbb{C}}$ is the closure of $\mathbb{C}$ in L^∞.

Definition 10. A probability measure P^* is called an *equivalent martingale measure*, or alternatively a *risk neutral probability*, if P^* is equivalent to P, and if under P^* the price process S is a σ martingale.

Clearly condition (ii) implies condition (i). Condition (i) is slightly too restrictive to imply the existence of an equivalent martingale measure P^*. (One can construct a trading strategy of $H_t(\omega) = 1_{\{[0,1]\setminus\mathbb{Q}\times\Omega\}}(t, \omega)$, which means one sells before each rational time and buys back immediately after it; combining H with a specially constructed càdlàg semimartingale shows that (i) does not imply the existence of P^* – see Delbaen and Schachermayer, 1994, p. 511.)

Let us examine then condition (ii). If NFLVR is not satisfied then there exists an $f_0 \in L^\infty_+$, $f_0 \not\equiv 0$, and also a sequence $f_n \in \mathbb{C}$ such that $\lim_{n\to\infty} f_n = f_0$ a.s., such that for each n, $f_n \geqslant f_0 - \frac{1}{n}$. In particular $f_n \geqslant -\frac{1}{n}$. This is almost the same as an arbitrage opportunity, since any element of $f \in \overline{\mathbb{C}}$ is the limit in the L^∞ norm of a sequence $(f_n)_{n\geqslant 1}$ in $\mathbb{C}$. This means that if $f \geqslant 0$ then the sequence of possible losses $(f_n^-)_{n\geqslant 1}$ tends to zero uniformly as $n \to \infty$, which means that the risk vanishes in the limit.

Theorem 3 *(Fundamental Theorem; Bounded Case). Let S be a bounded semimartingale. There exists an equivalent martingale measure P^* for S if and only if S satisfies NFLVR.*

Proof. Let us assume we have NFLVR. Since S satisfies the no arbitrage property we have $\mathbb{C} \cap L^\infty_+ = \{0\}$. However one can use the property NFLVR to show $\mathbb{C}$ is weak* closed in L^∞ (that is, it is closed in $\sigma(L^1, L^\infty)$), and hence there

will exist a probability P^* equivalent to P with $E^*\{f\} \leqslant 0$, all f in $\mathbb{C}$. (This is the Kreps–Yan separation theorem – essentially the Hahn–Banach theorem; see, e.g., Yan, 1980). For each $s < t$, $B \in \mathcal{F}_s$, $\alpha \in \mathbb{R}$, we deduce $\alpha(S_t - S_s)1_B \in \mathbb{C}$, since S is bounded. Therefore $E^*\{(S_t - S_s)1_B\} = 0$, and S is a martingale under P^*.

For the converse, note that NFLVR remains unchanged with an equivalent probability, so without loss of generality we may assume S is a martingale under P itself. If θ is admissible, then $(\int_0^t \theta_s \, \mathrm{d}S_s)_{t \geqslant 0}$ is a local martingale, hence it is a supermartingale. Since $E\{\theta_0 S_0\} = 0$, we have as well $E\{\int_0^\infty \theta_s \, \mathrm{d}S_s\} \leqslant E\{\theta_s S_0\} = 0$. This implies that for any $f \in \mathbb{C}$, we have $E\{f\} \leqslant 0$. Therefore it is true as well for $f \in \overline{\mathbb{C}}$, the closure of $\mathbb{C}$ in L^∞. Thus we conclude $\overline{\mathbb{C}} \cap L^\infty_+ = \{0\}$. □

Theorem 4 *(Corollary). Let S be a locally bounded semimartingale. There is an equivalent probability measure P^* under which S is a local martingale if and only if S satisfies NFLVR.*

The measure P^* in the corollary is known as a *local martingale measure*. We refer to Delbaen and Schachermayer (1994, p. 479) for the proof of the corollary. Examples show that in general P^* can make S only a local martingale, not a martingale. We also note that any semimartingale with continuous paths is locally bounded. However in the continuous case there is a considerable simplification: the no arbitrage property alone, properly interpreted, implies the existence of an equivalent local martingale measure P^* (see Delbaen, 1995). Indeed using the Girsanov theorem this implies that under the No Arbitrage assumption the semimartingale must have the form

$$S_t = M_t + \int_0^t H_s \, \mathrm{d}[M, M]_s,$$

where M is a local martingale under P, and with restrictions on the predictable process H. Indeed, if one has $\int_0^\epsilon H_s^2 \, \mathrm{d}[M, M]_s = \infty$ for some $\epsilon > 0$, then S admits "immediate arbitrage," a fascinating concept introduced by Delbaen and Schachermayer (1995).

For the general case, we have the impressive theorem of Delbaen and Schachermayer (1995, see for a proof), as follows:

Theorem 5 *(Fundamental Theorem; General Case). Let S be a semimartingale. There exists an equivalent probability measure P^* such that S is a sigma martingale under P^* if and only if S satisfies NFLVR.*[14]

[14] See Protter (2005, Section 9 of Chapter IV, pp. 237ff), for a treatment of sigma martingales; alternatively, see Jacod and Shiryaev (2002, Section 6e of Chapter III, pp. 214ff).

Caveat. *In the remainder of the paper we will abuse language by referring to the equivalent probability measure P^* which makes S into a sigma martingale, as an* equivalent martingale measure*. For clarity let us repeat:* if P^* is an equivalent martingale measure, then we can *a priori* conclude no more than that S is a sigma martingale (*or local martingale, if S has continuous paths*).

3.6 Numéraire invariance

Our portfolio as described in Section 3.4 consists of

$$V_t(a, b) = a_t S_t + b_t R_t,$$

where (a, b) are trading strategies, S is the risky security price, and $R_t = \exp(\int_0^t r_s \, \mathrm{d}s)$ is the price of a money market account. The process R is often called a *numéraire*. One can then deflate future monetary values by multiplying by $\frac{1}{R_t} = \exp(-\int_0^t r_s \, \mathrm{d}s)$. Let us write $Y_t = \frac{1}{R_t}$ and we shall refer to the process Y_t as a *deflator*. By multiplying S and R by $Y = \frac{1}{R}$, we can effectively reduce the situation to the case where the price of the money market account is constant and equal to one. The next theorem allows us to do just that.

Theorem 6 *(Numéraire Invariance). Let (a, b) be a trading strategy for (S, R). Let $Y = \frac{1}{R}$. Then (a, b) is self-financing for (S, R) if and only if (a, b) is self-financing for $(YS, 1)$.*

Proof. Let $Z = \int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s$. Then using integration by parts we have (since Y is continuous and of finite variation)

$$\begin{aligned}
\mathrm{d}(Y_t Z_t) &= Y_t \, \mathrm{d}Z_t + Z_t \, \mathrm{d}Y_t \\
&= Y_t a_t \, \mathrm{d}S_t + Y_t b_t \, \mathrm{d}R_t + \left(\int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s \right) \mathrm{d}Y_t \\
&= a_t (Y_t \, \mathrm{d}S_t + S_t \, \mathrm{d}Y_t) + b_t (Y_t \, \mathrm{d}R_t + R_t \, \mathrm{d}Y_t) \\
&= a_t \, \mathrm{d}(YS)_t + b_t \, \mathrm{d}(YR)_t
\end{aligned}$$

and since $YR = \frac{1}{R} R = 1$, this is

$$= a_t \, \mathrm{d}(YS)_t$$

since $\mathrm{d}YR = 0$ because YR is constant. Therefore

$$a_t S_t + b_t R_t = a_0 S_0 + b_0 + \int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s$$

if and only if

$$a_t \frac{1}{R_t} S_t + b_t = a_0 S_0 + b_0 + \int_0^t a_s \, \mathrm{d}\left(\frac{1}{R} S\right)_s .$$

□

Theorem 6 allows us to assume $R \equiv 1$ without loss of generality. Note that one can easily check that there is no arbitrage for (a, b) with (S, R) if and only if there is no arbitrage for (a, b) with $(\frac{1}{R}S, 1)$. By renormalizing, we no longer write $(\frac{1}{R}S, 1)$, but simply S.

The preceding theorem is the standard version, but in many applications (for example those arising in the modeling of stochastic interest rates), one wants to assume that the numéraire is a strictly positive semimartingale (instead of only a continuous finite variation process as in the previous theorem). We consider here the general case, where the numéraire is a (not necessarily continuous) semimartingale. For examples of how such a change of numéraire theorem can be used (albeit for the case where the deflator is assumed continuous), see for example (Geman et al., 1995). A reference to the literature for a result such as the following theorem is (Huang, 1985, p. 223).

Theorem 7 *(Numéraire Invariance; General Case). Let S, R be semimartingales, and assume R is strictly positive. Then the deflator $Y = \frac{1}{R}$ is a semimartingale and (a, b) is self-financing for (S, R) if and only if (a, b) is self-financing for $(\frac{S}{R}, 1)$.*

Proof. Since $f(x) = \frac{1}{x}$ is $\mathcal{C}^2$ on $(0, \infty)$, we have that Y is a (strictly positive) semimartingale by Itô's formula. By the self-financing hypothesis we have

$$\begin{aligned} V_t(a, b) &= a_t S_t + b_t R_t \\ &= a_0 S_0 + b_0 R_0 + \int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s . \end{aligned}$$

Let us assume $S_0 = 0$, and $R_0 = 1$. The integration by parts formula for semimartingales gives

$$\mathrm{d}(S_t Y_t) = \mathrm{d}\left(\frac{S_t}{R_t}\right) = S_{t-} \, \mathrm{d}\left(\frac{1}{R_t}\right) + \frac{1}{R_{t-}} \, \mathrm{d}S_t + \mathrm{d}\left[S, \frac{1}{R}\right]_t$$

and

$$\mathrm{d}\left(\frac{V_t}{R_t}\right) = V_{t-} \, \mathrm{d}\left(\frac{1}{R_t}\right) + \frac{1}{R_{t-}} \, \mathrm{d}V_t + \mathrm{d}\left[V, \frac{1}{R}\right]_t .$$

We can next use the self-financing assumption to write:

$$\begin{aligned}
\mathrm{d}\left(\frac{V_t}{R_t}\right) &= a_t S_{t-}\,\mathrm{d}\left(\frac{1}{R_t}\right) + b_t R_{t-}\,\mathrm{d}\left(\frac{1}{R_t}\right) + \frac{1}{R_{t-}} a_t\,\mathrm{d}S_t + \frac{1}{R_{t-}} b_t\,\mathrm{d}R_t \\
&\quad + a_t\,\mathrm{d}\left[S, \frac{1}{R}\right]_t + b_t\,\mathrm{d}\left[R, \frac{1}{R}\right]_t \\
&= a_t\left(S_{t-}\,\mathrm{d}\left(\frac{1}{R}\right) + \frac{1}{R_{t-}}\,\mathrm{d}S + \mathrm{d}\left[S, \frac{1}{R}\right]\right) \\
&\quad + b_t\left(R_{t-}\,\mathrm{d}\left(\frac{1}{R}\right) + \frac{1}{R_{t-}}\,\mathrm{d}R + \mathrm{d}\left[R, \frac{1}{R}\right]\right) \\
&= a_t\,\mathrm{d}\left(S\frac{1}{R}\right) + b_t\,\mathrm{d}\left(R\frac{1}{R}\right).
\end{aligned}$$

Of course $R_t \frac{1}{R_t} = 1$, and $\mathrm{d}(1) = 0$; hence

$$\mathrm{d}\left(\frac{V_t}{R_t}\right) = a_t\,\mathrm{d}\left(S_t\frac{1}{R_t}\right).$$

In conclusion we have

$$V_t = a_t S_t + b_t R_t = b_0 + \int_0^t a_s\,\mathrm{d}S_s + \int_0^t b_s\,\mathrm{d}R_s,$$

and

$$a_t\left(\frac{S_t}{R_t}\right) + b_t = \frac{V_t}{R_t} = b_0 + \int_0^t a_s\,\mathrm{d}\left(\frac{S_s}{R_s}\right).$$

□

3.7 Redundant derivatives

Let us assume given a security price process S, and by the results in Section 3.6 we take $R_t \equiv 1$. Let $\mathcal{F}_t^0 = \sigma(S_r; r \leqslant t)$ and let $\mathcal{F}_t^{\sim} = \mathcal{F}_t^0 \vee \mathcal{N}$ where $\mathcal{N}$ are the null sets of $\mathcal{F}$ and $\mathcal{F} = \bigvee_t \mathcal{F}_t^0$, under P, defined on $(\Omega, \mathcal{F}, P)$. Finally we take $\mathcal{F}_t = \bigcap_{u>t} \mathcal{F}_u^{\sim}$. A *derivative* on S is then a random variable $C \in \mathcal{F}_T$, for some fixed time T. Note that we pay a small price here for the simplification of taking $R_t \equiv 1$, since if R_t were to be a nonconstant stochastic process, it might well change the minimal filtration we are taking, because then the processes of interest would be (R_t, S_t), in place of just S_t/R_t.

One goal of Finance Theory is to show there exists a self financing trading strategy (a, b) that one can use either to obtain C at time T, or to come as close as possible – in an appropriate sense – to obtaining C. This is the issue we discuss in this section.

Definition 11. Let S be the price process of a risky security and let R be the price process of a money market account (numéraire), which we setting equal to the constant process 1.[15] A derivative $C \in \mathcal{F}_T$ is said to be *redundant* if there exists an admissible self-financing trading strategy (a, b) such that

$$C = a_0 S_0 + b_0 R_0 + \int_0^T a_s \, \mathrm{d}S_s + \int_0^T b_s \, \mathrm{d}R_s.$$

Let us normalize S by writing $M = \frac{1}{R} S$; then C will still be redundant under M and hence we have (taking $R_t = 1$, all t):

$$C = a_0 M_0 + b_0 + \int_0^T a_s \, \mathrm{d}M_s.$$

Next note that if P^* is any equivalent martingale measure making M a martingale, and if C has finite expectation under P^*, we then have

$$E^*\{C\} = E^*\{a_0 M_0 + b_0\} + E^*\left\{ \int_0^T a_s \, \mathrm{d}M_s \right\}$$

provided all expectations exist,

$$= E^*\{a_0 M_0 + b_0\} + 0.$$

Theorem 8. *Let C be a redundant derivative such that there exists an equivalent martingale measure P^* with $C \in \mathcal{L}^*(M)$. (See the second definition following for a definition of $\mathcal{L}^*(M)$.) Then there exists a unique no arbitrage price of C and it is $E^*\{C\}$.*

Proof. First we note that the quantity $E^*\{C\}$ *is the same* for every equivalent martingale measure. Indeed if Q_1 and Q_2 are both equivalent martingale measures, then

$$E_{Q_i}\{C\} = E_{Q_i}\{a_0 M_0 + b_0\} + E_{Q_i}\left\{ \int_0^T a_s \, \mathrm{d}M_s \right\}.$$

But $E_{Q_i}\{\int_0^T a_s \, \mathrm{d}M_s\} = 0$, and $E_{Q_i}\{a_0 M_0 + b_0\} = a_0 M_0 + b_0$, since we assume a_0, M_0, and b_0 are known at time 0 and thus without loss of generality are taken to be constants.

[15] Although R is taken to be constant and equal to 1, we include it initially in the definition to illustrate the role played by being able to take it a constant process.

Next suppose one offers a price $v > E^*\{C\} = a_0M_0 + b_0$. Then one follows the strategy $a = (a_s)_{s\geqslant 0}$ and (we are ignoring transaction costs) at time T one has C to present to the purchaser of the option. One thus has a sure profit (that is, risk free) of $v - (a_0M_0 + b_0) > 0$. This is an arbitrage opportunity. On the other hand, if one can buy the claim C at a price $v < a_0M_0 + b_0$, analogously at time T one will have achieved a risk-free profit of $(a_0M_0 + b_0) - v$. □

Definition 12. If C is a derivative, and there exists an admissible self-financing trading strategy (a, b) such that

$$C = a_0M_0 + b_0 + \int_0^T a_s \, \mathrm{d}M_s;$$

then the strategy a is said to *replicate* the derivative C.

Theorem 9 *(Corollary). If C is a redundant derivative, then one can replicate C in a self-financing manner with initial capital equal to $E^*\{C\}$, where P^* is any equivalent martingale measure for the normalized price process M.*

At this point we return to the issue of *put–call parity* mentioned in the introduction (Section 2). Recall that we had the trivial relation

$$M_T - K = (M_T - K)^+ - (K - M_T)^+,$$

which, by taking expectations under P^*, shows that the price of a call at time 0 equals the price of a put plus the stock price minus K. More generally at time t, $E^*\{(M_T - K)^+ \mid \mathcal{F}_t\}$ equals the value of a put at time t plus the stock price at time t minus K, by the P^* martingale property of M.

It is tempting to consider markets where all derivatives are redundant. Unfortunately, this is too large a space of random variables; we wish to restrict ourselves to derivatives that have good integrability properties as well.

Let us fix an equivalent martingale measure P^*, so that M is a martingale (or even a local martingale) under P^*. We consider all self-financing trading strategies (a, b) such that the process $(\int_0^t a_s^2 \, \mathrm{d}[M, M]_s)^{1/2}$ is locally integrable: that means that there exists a sequence of stopping times $(T_n)_{n\geqslant 1}$ which can be taken $T_n \leqslant T_{n+1}$, a.s., such that $\lim_{n\to\infty} T_n \geqslant T$ a.s. and $E^*\{(\int_0^{T_n} a_s^2 \, \mathrm{d}[M, M]_s)^{1/2}\} < \infty$, each T_n. Let $\mathcal{L}^*(M)$ denote the class of such strategies, under P^*. We remark that we are cheating a little here: we are letting our definition of a complete market (which follows) depend on the measure P^*, and it would be preferable to define it in terms of the objective probability P. How to go about doing this is a nontrivial issue. In the happy case where the price process is already a local martingale under the objective probability measure, this issue of course disappears.

Definition 13. A market model $(M, \mathcal{L}^*(M), P^*)$ is *complete* if every derivative $C \in L^1(\mathcal{F}_T, \mathrm{d}P^*)$ is redundant for $\mathcal{L}^*(M)$. That is, for any $C \in L^1(\mathcal{F}_T, \mathrm{d}P^*)$,

there exists an admissible self-financing trading strategy (a, b) with $a \in \mathcal{L}^*(M)$ such that

$$C = a_0 M_0 + b_0 + \int_0^T a_s \, \mathrm{d}M_s,$$

and such that $(\int_0^t a_s \, \mathrm{d}M_s)_{t \geqslant 0}$ is uniformly integrable. In essence, then, a complete market is one for which every derivative is redundant.

We point out that the above definition is one of many possible definitions of a complete market. For example, one could limit attention to nonnegative random payoffs and/or payoffs that are in $L^2(\mathcal{F}_T, \mathrm{d}P^*)$.

We note that in probability theory a martingale M is said to have the *predictable representation property* if for any $C \in L^2(\mathcal{F}_T)$ one has

$$C = E\{C\} + \int_0^T a_s \, \mathrm{d}M_s$$

for some predictable $a \in \mathcal{L}(M)$. This is, of course, essentially the property of market completeness. Martingales with predictable representation are well studied and this theory can usefully be applied to Finance. For example, suppose we have a model (S, R) where by a change of numéraire we take $R = 1$. Suppose further there is an equivalent martingale measure P^* such that S is a Brownian motion under P^*. Then the model is complete for all claims C in $L^1(\mathcal{F}_T, P^*)$ such that $C \geqslant -\alpha$, for some $\alpha \geqslant 0$. (α can depend on C.) To see this, we use martingale representation (see, e.g., Protter, 2005) to find a predictable process a such that for $0 \leqslant t \leqslant T$:

$$E^*\{C \mid \mathcal{F}_t\} = E^*\{C\} + \int_0^t a_s \, \mathrm{d}S_s.$$

Let

$$V_t(a, b) = a_0 S_0 + b_0 + \int_0^t a_s \, \mathrm{d}S_s + \int_0^t b_s \, \mathrm{d}R_s;$$

we need to find b such that (a, b) is an admissible, self-financing trading strategy. Since $R_t = 1$, we have $\mathrm{d}R_t = 0$, hence we need

$$a_t S_t + b_t R_t = b_0 + \int_0^t a_s \, \mathrm{d}S_s,$$

and taking $b_0 = E^*\{C\}$, we have

$$b_t = b_0 + \int_0^t a_s \, \mathrm{d}S_s - a_t S_t$$

provides such a strategy. It is admissible since $\int_0^t a_s \, \mathrm{d}S_s \geqslant -\alpha$ for some α which depends on C.

Unfortunately, having the predictable representation property is rather delicate, and few martingales possess this property. Examples include Brownian motion, the compensated Poisson process (but *not* mixtures of the two nor even the difference of two Poisson processes) (although see Jeanblanc and Privault, 2002 for sufficient conditions when one can mix the two and have completeness), and the Azéma martingales. (One can consult Protter, 2005 for background, and Dritschel and Protter, 1999 for more on the Azéma martingales.) One can mimic a complete market in the case (for example) of two independent noises, each of which is complete alone. Several authors have done this with Brownian noise together with compensated Poisson noise, by proposing hedging strategies for each noise separately. A recent example of this is Kusuoka (1999) (where the Poisson intensity can depend on the Brownian motion) in the context of default risk models. A more traditional example is Jeanblanc-Piqué and Pontier (1990).

Most models are therefore *not* complete, and most practitioners believe the a financial world being modeled is at best only approximately complete. We will return again to the notion of an incomplete market later on in this section. First, we need to characterize complete markets. In this regard, we have the following result:

Theorem 10. *Suppose there is an equivalent martingale measure P^* such that M is a local martingale. Then P^* is the unique equivalent martingale measure only if the market is complete.*

This theorem is a trivial consequence of Dellacherie's approach to martingale representation: if there is a unique probability making a process M a local martingale, then M must have the martingale representation property. The theory has been completely resolved in the work of Jacod and Yor. [See for example Protter (2005, Chapter IV, Section 4), for a pedagogic approach to the theory.]

To give an example of what can happen, let $\mathcal{M}^2$ be the set of equivalent probabilities making M an L^2-martingale. Then M has the predictable representation property (and hence market completeness) for every extremal element of the convex set $\mathcal{M}^2$. If $\mathcal{M}^2 = \{P^*\}$, only one element, then of course P^* is extremal. (See Protter, 2005, Theorem 40, p. 186.) Indeed P^* is in fact unique in the proto-typical example of Brownian motion; since many diffusions can be constructed as pathwise functionals of Brownian motion they inherit the

completeness of the Brownian model. But there are examples where one has complete markets without the uniqueness of the equivalent martingale measure (see Artzner and Heath, 1995 in this regard, as well as Jarrow et al., 1999). Nevertheless the situation is simpler when we assume our models have continuous paths.

The next theorem is a version of what is known as *the second fundamental theorem of asset pricing*. We state and prove it for the case of L^2 derivatives only. We note that this theorem has a long and illustrious history, going back to the fundamental paper of Harrison and Kreps (1979, p. 392) for the discrete case, and to Harrison and Pliska (1981, p. 241) for the continuous case, although in Harrison and Pliska (1981) the theorem below is stated only for the "only if" direction.

Theorem 11. *Let M have continuous paths. There is a unique P^* such that M is an L^2 P^*-martingale if and only if the market is complete.*

Proof. The theorem follows easily from Theorems 38, 39, and 340 of Protter (2005, pp. 185–186); we will assume those results and prove the theorem. Theorem 39 shows that if P^* is unique then the market model is complete. If P^* is not unique but the model is nevertheless complete, then by Theorem 37 P^* is nevertheless extremal in the space of probability measures making M an L^2 martingale. Let Q be another such extremal probability, and let $L_\infty = \frac{\mathrm{d}Q}{\mathrm{d}P^*}$ and $L_t = E_P\{L_\infty \mid \mathcal{F}_t\}$, with $L_0 = 1$. Let $T_n = \inf\{t > 0\colon |L_t| \geqslant n\}$. L will be continuous by Theorem 40 of Protter (2005, p. 186), hence $L^n_t = L_{t\wedge T_n}$ is bounded. We then have, for bounded $C \in \mathcal{F}_s$:

$$E_Q\{M_{t\wedge T_n} C\} = E^*\big\{M_{t\wedge T_n} L^n_t C\big\},$$
$$E_Q\{M_{s\wedge T_n} C\} = E^*\big\{M_{s\wedge T_n} L^n_s C\big\}.$$

The two left sides of the above equalities are equal and this implies that ML^n is a martingale, and thus L^n is a bounded P^*-martingale orthogonal to M. It is hence constant by Theorem 39 of Protter (2005, p. 185). We conclude $L_\infty \equiv 1$ and thus $Q = P^*$. □

Note that if C is a redundant derivative, then the no arbitrage price of C is $E^*\{C\}$, for any equivalent martingale measure P^*. (If C is redundant then we have seen the quantity $E^*\{C\}$ is the same under every P^*.) However, if a market model is not complete, then

- there will arise nonredundant claims, and
- there will be more than one equivalent martingale measure P^*.

We now have the conundrum: if C is nonredundant, what is the no arbitrage price of C? We can no longer argue that it is $E^*\{C\}$, because there are many such values! The absence of this conundrum is a large part of the appeal of complete markets. One resolution of this conundrum is to use an

investor's preferences/tastes to select among the set of possible equivalent martingale measures a unique one, that will make them indifferent between holding the derivative in their portfolio or not. This is an interesting area of current research and for more on this topic see Duffie (2001) and references cited therein.

Finally, let us note that when C is redundant there is always a replication strategy a. However, when C is nonredundant it cannot be replicated. In the nonredundant case the best we can do is replicate in some approximate sense (for example expected squared error loss), and we call the strategy we follow a *hedging strategy*. See for example Föllmer and Sondermann (1986) and Jacod et al. (2000) for results about hedging strategies.

3.8 The stochastic price process

For simplicity, we will limit our discussion to one dimension. Let $S = (S_t)_{t \geqslant 0}$ denote our price process for a risky asset. Let $s < t$ and suppose $t - s$ is a small but finite time interval. The randomness in the market price between times s and t comes from the cumulative price changes due to the actions of many traders. Let us enumerate the traders' individual price changes over this interval. Let the random variable θ_i denote the *change in the price* of the asset due to the different sized purchase or sale by the ith trader between the times s and t. No activity corresponds to $\theta_i = 0$. The total effect of the traders' actions on the price is $\Theta = \sum_{i=1}^{n} \theta_i$.

If n is large (even $n = 50$ would suffice in most cases, and typically n is much larger) and if the θ_i are independent with mean μ and finite variance σ^2, then by the Central Limit Theorem we have that $\mathcal{L}(\Theta) = \mathcal{L}(\sum_{i=1}^{n} \theta_i) \approx N(n\mu, n\sigma^2)$[16], where $\mathcal{L}(Y)$ denotes the law, or distribution, of a random variable Y. Under these assumptions, and with $\mu = 0$, it is reasonable to describe the random forces affecting the asset price as Gaussian. We further remark that, as is well known, one can substantially weaken the hypotheses that the random variables $(\theta_i)_{i \geqslant 1}$ are independent, using for example martingale central limit theorems [see, e.g., Jacod and Shiryaev (2002) for a definitive treatment, or Jacod and Protter (2004) for an introductory treatment], and one can also weaken the assumption that all variances are identical. One could then use a stochastic differential equation to give a dynamic model of the risky asset price, where we let B denote a Brownian motion:

$$\mathrm{d}S_t = \sigma(t, S_t)\,\mathrm{d}B_t + \mu(t, S_t)\,\mathrm{d}t, \tag{9}$$

since the increments of the Brownian motion are given by $B_t - B_s \sim N(0, \sigma_0^2(t - s))$.[17] We usually take $\sigma_0^2 = 1$, since otherwise we could sim-

[16] As is customary, $N(\mu, \sigma^2)$ denotes the normal distribution (also known as the Gaussian distribution) with mean μ and variance σ^2.

[17] By choosing the Brownian motion, which has stationary and independent increments, we are implicitly assuming that the distributions of the traders' likelihoods to trade in a time interval (s, t) depends

ply modify the coefficient function $\sigma(t, x)$. The function σ can be thought of as the sensitivity of the price change to "market noise" when $S_t = x$. The term given by $\mu(t, S_t)\,\mathrm{d}t$ is called the "drift," and it corresponds to changes in the risky asset price which are not due to market noise, but rather due to market fundamentals.

There are many problems with the model given by Eq. (9), but the most fundamental one is that price process must always take nonnegative values, and there is no a priori reason that S be positive, even with taking $S_0 > 0$. Let us address this problem. Henceforth we will consider only *autonomous* coefficients.[18] This means that if the noise process has stationary and independent increments, then the price process will be a time homogeneous strong Markov process.[19] With dependence on time in the coefficients, one loses the time homogeneity, although the solution is still Markov.[20] Suppose instead we let the risky asset price process be $Y = (Y_t)_{t\geqslant 0}$ with $Y_0 = 1$ and $Y_t > 0$ for all t, $0 \leqslant t \leqslant T$ for some time horizon T, a.s. Since $Y > 0$ always, we can take its logarithm, and define $X_t = \ln(Y_t)$, and obviously $Y_t = e^{X_t}$.

Let us assume that X is the unique solution of (9), where appropriate hypotheses are made upon the coefficients σ and μ to ensure a unique nonexploding solution. We can use Itô's formula to find a dynamic expression for Y. Indeed,

$$e^{X_t} = e^{X_0} + \int_0^t e^{X_s}\,\mathrm{d}X_s + \frac{1}{2}\int_0^t e^{X_s}\,\mathrm{d}[X, X]_s$$

and substituting Y for e^X we get

$$\begin{aligned} Y_t = Y_0 &+ \int_0^t \sigma\big(s, \ln(Y_s)\big)\,\mathrm{d}B_s + \int_0^t Y_s\mu\big(s, \ln(Y_s)\big)^2\,\mathrm{d}s \\ &+ \frac{1}{2}\int_0^t Y_s\sigma\big(s, \ln(Y_s)\big)^2\,\mathrm{d}s, \end{aligned}$$

only on the length of the interval $t - s$ and does not change with time, and are independent for disjoint time intervals. Both of these assumptions have been questioned repeatedly. See for example Clark (1973) for the case against the stationarity assumption.

[18] That is, coefficients of the form $\sigma(x)$, rather than of the form $\sigma(t, x)$.

[19] See Protter (2005, p. 36 or p. 299) for example, for a definition of a strong Markov process.

[20] By assuming time homogeneity, however, we are depriving ourselves of a useful possibility to allow for excess kurtosis in our models, by allowing time dependence in the diffusion coefficient; see for example Madan and Yor (2002). Kurtosis of a random variable X with mean μ is sometimes defined as $\gamma = E\{(X-\mu)^4\}/(E\{(X-\mu)^2\})^2$, and excess kurtosis is simply $\gamma - 3$, because the kurtosis of a Gaussian random variable is 3.

and letting $\hat{\sigma}(t, y) = \sigma(t, \ln(y))$ and $\hat{\mu}(t, y) = \mu(t, \ln(y))$, we have:

$$\mathrm{d}Y_t = Y_t \hat{\sigma}(t, Y_t)\,\mathrm{d}B_t + Y_t \left\{ \hat{\mu}(t, Y_t) + \frac{1}{2}\hat{\sigma}(t, Y_t)^2 \right\} \mathrm{d}t. \tag{10}$$

Note that we have shown, inter alia, that if there exists a unique nonexploding solution to Equation (9), then there also exists the same for (10), even though the function $y \to y\sigma(t, y)$ need not be globally Lipshitz. We further note that if an equation of the form

$$\mathrm{d}Y_t = f(Y_t)Y_t\,\mathrm{d}B_t + g(Y_t)Y_t\,\mathrm{d}t; \quad Y_0 > 0 \tag{11}$$

has a unique, strong, nonexploding solution, then $P(\omega\colon \exists t$ such that $Y_t(\omega) \leqslant 0) = 0$ (see Protter, 2005, p. 351).

The absolute magnitude that Y changes, that is $Y_{t+\mathrm{d}t} - Y_t$, is not by itself as meaningful as the relative change. For example if $Y_{t+\mathrm{d}t} - Y_t = \$0.12$, this can be a large change if $Y_t = \$1.25$, or it can be an insignificant change if $Y_t = \$105.12$. Therefore we often speak of the *return on the asset*, which is *the change of the price divided by the original value*. Since we now have that $Y > 0$, we can rewrite Eq. (11) as

$$\frac{\mathrm{d}Y_t}{Y_t} = f(Y_t)\,\mathrm{d}B_t + g(Y_t)\,\mathrm{d}t; \quad Y_0 > 0,$$

and indeed this is often the way the price process is written in the literature.[21]

The simplest form of such a price process is when $f = \sigma$ and $g = \mu$ are constants, and then of course

$$Y_t = \exp\left(\sigma B_t + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) \equiv \mathcal{E}(\sigma B_t + \mu t),$$

where $\mathcal{E}(Z)$ denotes the *stochastic exponential* of a semimartingale Z.[22] One reason this simplest form is so popular is that if $f = \sigma$ a constant, it is easy to estimate this parameter. Indeed, a simple procedure is to sample Y at $n + 1$ equal spaced time steps $\{t_1, t_2, \ldots, t_{n+1}\}$ with $t_i - t_{i-1} = \delta$ in chronological order and let

$$\hat{\mu} = \frac{1}{n\delta} \sum_{i=1}^{n} \ln\left(\frac{Y_{t_{i+1}}}{Y_{t_i}}\right)$$

[21] The coefficient μ is called the *drift* and it reflects the *fundamentals* of the asset: its position in the industry and its expected future earnings or losses. The coefficient f is called the *volatility* and it represents a the standard deviation of the returns. It is the volatility that creates risk in the investment, and it is the primary object of study.

[22] For a continuous semimartingale X the stochastic exponential of X, denoted $\mathcal{E}(X)$, is the process Y given by $Y_t = \mathcal{E}(X)_t = \exp(X_t - \frac{1}{2}[X, X]_t)$, and Y satisfies the exponential type differential equation $\mathrm{d}Y_t = Y_t\,\mathrm{d}X_t$; $Y_0 = 1$. See Protter (2005, p. 85) for more about the stochastic exponential, which is also sometimes called the *Doléans–Dade exponential*.

and

$$\hat{\sigma}^2 = \frac{1}{(n-1)\delta} \sum_{i=1}^{n} \left(\ln\left(\frac{Y_{t_{i+1}}}{Y_{t_i}} \right) - \hat{\mu} \right)^2$$

and then $\hat{\sigma}^2$ is an unbiased and consistent estimator for σ^2. (If one takes δ as a fraction of a year, then the parameters are annualized.) Of course there are other reasons the form is so popular; see Section 3.10.

Let us return to the heuristic derivation of the price process. Recall that θ_i denotes the change in the price of the risky asset due to the size of a purchase or sale by the ith trader between the times s and t, with the total effect of the traders' actions being $\Theta = \sum_{i=1}^{n} \theta_i$. In reality, there are many different rubrics of traders. Some examples are (a) a small trader; (b) a trader for a large mutual fund; (c) a trader for a pension fund; (d) corporate traders; and (e) traders for hedge funds. These traders have different goals and different magnitudes of equity supporting their trades. Let us divide the traders into rubrics, and for rubric n we enumerate the traders $(n, 1), (n, 2), \ldots, (n, n)$ and we let the traders' impacts on the price between times s and t be denoted $U_{n,1}, U_{n,2}, \ldots, U_{n,n}$. We assume that the random variables $(U_{n,i})_{1 \leqslant i \leqslant n}$ are i.i.d. for every $n \geqslant 1$ and independent across all the n, and moreover for each fixed n have common law l_n. Set:

$$\Psi_n = \sum_{i=1}^{n} U_{n,i} \quad \text{and} \quad \theta_n = \frac{\Psi_n - E(\Psi_n)}{V_n},$$

where V_n is the standard deviation of Ψ_n. Then

$$\Theta_n = \sum_{i=1}^{n} \theta_i = \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{U_{i,j} - E(U_{i,j})}{V_i}$$

represents the normalized random effect on the market price of the traders' actions between times s and t. We have that Θ_n converges in distribution to a random variable which is *infinitely divisible*. If we denote this random variable as $Z_t - Z_s$, and think of Z as a noise process with stationary and independent increments, then Z must be a *Lévy process* (see for example Protter, 2005, p. 21, for this result, and in general Protter, 2005 or Bertoin, 1996 for more information on Lévy processes in general). Since the only Lévy process with continuous paths is Brownian motion with drift, in order to be different from the classical case, the paths $t \to Z_t(\omega)$ must have jumps.

A discontinuous price process requires a different analysis. Let us understand why. We begin as before and let Z be a Lévy process (with jumps), and then form X by

$$\mathrm{d}X_t = \sigma(X_{t-})\,\mathrm{d}Z_t + \mu(X_{t-})\,\mathrm{d}t \tag{12}$$

and $Y_t = e^{X_t} > 0$. Next using Itô's formula we have

$$e^{X_t} = e^{X_0} + \int_0^t e^{X_{s-}}\,\mathrm{d}X_s + \frac{1}{2}\int_0^t e^{X_{s-}}\,\mathrm{d}[X,X]_s^c + \sum_{s\leqslant t}\bigl(e^{X_s} - e^{X_{s-}} - e^{X_{s-}}\Delta X_s\bigr)$$

and substituting Y_t for e^{X_t}, and using that for a Lévy process Z one a fortiori has that $\mathrm{d}[Z,Z]_t^c = \gamma\,\mathrm{d}t$ for some constant $\gamma \geqslant 0$, we have

$$\begin{aligned} Y_t &= Y_0 + \int_0^t Y_{s-}\sigma\bigl(\ln(Y_{s-})\bigr)\,\mathrm{d}Z_s + \int_0^t Y_{s-}\mu\bigl(\ln(Y_{s-})\bigr)\,\mathrm{d}s \\ &\quad + \frac{1}{2}\int_0^t Y_{s-}\sigma\bigl(\ln(Y_{s-})\bigr)^2\gamma\,\mathrm{d}s \\ &\quad + \sum_{s\leqslant t}\bigl(Y_s - Y_{s-} - Y_{s-}\sigma\bigl(\ln(Y_{s-})\bigr)\Delta Z_s\bigr) \\ &= Y_0 + \hat{\sigma}(Y_{s-})\,\mathrm{d}Z_s + \int_0^t Y_{s-}\left\{\hat{\mu}(Y_{s-}) + \frac{\gamma}{2}\hat{\sigma}(Y_{s-})^2\right\}\mathrm{d}s \\ &\quad + \sum_{s\leqslant t}\bigl(Y_s - Y_{s-} - Y_{s-}\hat{\sigma}(Y_{s-})\Delta Z_s\bigr) \end{aligned}$$

which does not satisfy a stochastic differential equation driven by $\mathrm{d}Z$ and $\mathrm{d}t$. If we were simply to forget about the series term at the end, as many researchers do, and instead were to consider the following equation as our dynamic model:

$$\mathrm{d}Y_t = Y_{t-}f(Y_{t-})\,\mathrm{d}Z_t + Y_{t-}g(Y_{t-})\,\mathrm{d}t, \quad Y_0 > 0, \tag{13}$$

then we could no longer ensure that Y is a positive price process! Indeed, if we consider the simple case where $f = \sigma$ and $g = \mu$ are both constants, with $Y_0 = 1$, we have

$$\mathrm{d}Y_t = \sigma Y_{t-}\,\mathrm{d}Z_t + \mu Y_{t-}\,\mathrm{d}t$$

which has a closed form solution

$$Y_t = \exp\left(\sigma Z_t + \mu t - \frac{1}{2}\sigma^2\gamma t\right)\prod_{s\leqslant t} e^{\{-\sigma\Delta Z_s\}}(1+\sigma\Delta Z_s)$$

and thus as soon as one has a jump $\Delta Z_s \leqslant -\frac{1}{\sigma}$, we have Y becoming zero or negative. In general, for equations of the form (13), a sufficient condition to

have $Y > 0$ always is that $|\Delta Z_s| < \frac{1}{\|\sigma\|_{L^\infty}}$, for all $s \geqslant 0$, a.s. (see Protter, 2005, p. 352).

Should one stop here? Perhaps one should consider more general noise processes than Lévy processes? Indeed, one could consider time changes of Lévy processes, since there is empirical evidence of nonstationarity in the increments of the noise process (see for example Clark, 1973 and more recently Carr and Wu, 2003); or, and not exclusively, one might think there is some historical dependence in the noise process, violating the assumptions of independence of the increments. The advantage that the assumption of independent increments provides is that the solution X of the SDE is a strong Markov process (a time homogeneous strong Markov process if the increments are also stationary and the coefficients are autonomous). Therefore, so is $Y = e^X$, since the function $x \to e^x$ is injective. If however one replaces the Lévy driving process Z with a strong Markov process, call it Q, then the solution X with Z replaced by Q, will no longer be a Markov process, although the vector process (X, Q) will be strong Markov (see Protter, 2005, Theorem 32, p. 300 and Theorem 73, p. 353).

But why do we care if X, and hence Y, is strong Markov? Many researchers claim there is evidence that the price process has short term momentum, which would violate the Markov property. The reason is that it is mathematically convenient to have Y be Markov, since in this case one has a hope of calculating (or at least approximating) a hedging strategy for a financial derivative. If however one is willing to forego having a time homogeneous strong Markov price process, then one can consider a price process of the form

$$\mathrm{d}Y_t = Y_{t-} f(Y_{t-})\,\mathrm{d}Z_t + Y_{t-} g(Y_{t-})\,\mathrm{d}A_t, \quad Y_0 > 0, \tag{14}$$

where Z and A are taken to be semimartingales. There is a danger to this level of generality, since not all such models lead to an absence of arbitrage, as we shall see in Section 3.11.

3.9 Determining the replication strategy

It is rare that we can actually "explicitly" compute a replication strategy for a derivative security. However, there are simple cases where miracles happen; and when there are no miracles, then we can often approximate hedging strategies using numerical techniques.

Let us consider a standard, and relatively simple derivative security of the form

$$C = f(S_T),$$

where S is the price of the risky security. The two most important examples (already discussed in Section 2) are

- *The European call option*: Here $f(x) = (x - K)^+$ for a constant K, so the contingent claim is $C = (S_T - K)^+$. K is referred to as the *strike*

price and T is the expiration time. In words, the European call option gives the holder the right to *buy* one unit of the security at the price K at time T. Thus the (random) value of the option at time T is $(S_T - K)^+$.

- *The European put option*: Here $f(x) = (K - x)^+$. This option gives the holder the right to *sell* one unit of the security at time T at price K. Hence the (random) value of the option at time T is $(K - S_T)^+$.

To illustrate the ideas involved, let us take $R_t \equiv 1$ by a change of numéraire, and let us suppose that $C = f(S_T)$ is a redundant derivative. The *value* of a replicating self-financing trading strategy (a, b) for the claim, at time t, is

$$V_t = E^*\{f(S_T) \mid \mathcal{F}_t\} = a_0 S_0 + b_0 + \int_0^t a_s \, \mathrm{d}S_s.$$

We now make a series of hypotheses in order to obtain a simpler analysis.

Hypothesis 1. *S is a Markov process under some equivalent local martingale measure P^*.*

Under hypothesis 1 we have that

$$V_t = E^*\{f(S_T) \mid \mathcal{F}_t\} = E^*\{f(S_T) \mid S_t\}.$$

But measure theory tells us that there exists a function $\varphi(t, \cdot)$, for each t, such that

$$E^*\{f(S_T) \mid S_t\} = \varphi(t, S_t).$$

Hypothesis 2. *$\varphi(t, x)$ is $\mathcal{C}^1$ in t and $\mathcal{C}^2$ in x.*

This hypothesis enables us to use Itô's formula:

$$\begin{aligned} V_t &= E^*\{f(S_T) \mid \mathcal{F}_t\} = \varphi(t, S_t) \\ &= \varphi(0, S_0) + \int_0^t \varphi'_x(s, S_{s-}) \, \mathrm{d}S_s \\ &\quad + \int_0^t \varphi'_s(s, S_{s-}) \, \mathrm{d}s + \frac{1}{2} \int_0^t \varphi''_{xx}(s, S_{s-}) \, \mathrm{d}[S, S]^c_s \\ &\quad + \sum_{0<s\leqslant t} \{\varphi(s, S_s) - \varphi(s, S_{s-}) - \varphi'_x(s, S_{s-}) \Delta S_s\}. \end{aligned}$$

Hypothesis 3. *S has continuous paths.*

With Hypothesis 3 Itô's formula simplifies:

$$V_t = \varphi(t, S_t) = \varphi(0, S_0) + \int_0^t \varphi'_x(s, S_s)\,\mathrm{d}S_s + \int_0^t \varphi'_s(s, S_s)\,\mathrm{d}s + \frac{1}{2}\int_0^t \varphi''_{xx}(s, S_s)\,\mathrm{d}[S, S]_s. \tag{15}$$

Since V is a P^* martingale, the right side of (15) must also be a P^* martingale. This is true if

$$\int_0^t \varphi'_s(s, S_s)\,\mathrm{d}s + \frac{1}{2}\int_0^t \varphi''_{xx}(s, S_s)\,\mathrm{d}[S, S]_s = 0. \tag{16}$$

For Eq. (16) to hold, it is reasonable to require that $[S, S]$ have paths which are absolutely continuous almost surely. Indeed, we assume more than that. We assume a specific structure for $[S, S]$:

Hypothesis 4. $[S, S]_t = \int_0^t h(s, S_s)^2\,\mathrm{d}s$ *for some jointly measurable function* h *mapping* $\mathbb{R}_+ \times \mathbb{R}$ *to* $\mathbb{R}$.

We then get that (16) certainly holds if φ is the solution of the partial differential equation:

$$\frac{1}{2}h(s, x)^2\frac{\partial^2\varphi}{\partial x^2}(s, x) + \frac{\partial\varphi}{\partial s}(s, x) = 0$$

with boundary condition $\varphi(T, x) = f(x)$. Note that if we combine Hypotheses 1–4 we have a continuous Markov process with quadratic variation $\int_0^t h(s, S_s)^2\,\mathrm{d}s$. An obvious candidate for such a process is the solution of a stochastic differential equation

$$\mathrm{d}S_s = h(s, S_s)\,\mathrm{d}B_s + k(s; S_r; r \leqslant s)\,\mathrm{d}s,$$

where B is a standard Wiener process (Brownian motion) under P. S is a continuous Markov process under P^*, with quadratic variation $[S, S]_t = \int_0^t h(s, S_s)^2\,\mathrm{d}s$ as desired.

The quadratic variation is a path property and is unchanged by changing to an equivalent probability measure P^* (see Protter, 2005 for example). But what about the Markov property? Why is S a Markov process under P^* when b can be path dependent? Here we digress a bit.

Let us analyze P^* in more detail. Since P^* is equivalent to P, we can let $Z = \frac{\mathrm{d}P^*}{\mathrm{d}P}$ and $Z > 0$ a.s. $(\mathrm{d}P)$. Let $Z_t = E\{Z \mid \mathcal{F}_t\}$, which is clearly a martingale.

By Girsanov's theorem (see, e.g., Protter, 2005),

$$\int_0^t h(s, S_s)\,\mathrm{d}B_s - \int_0^t \frac{1}{Z_s}\,\mathrm{d}\left[Z, \int_0^{\cdot} h(r, S_r)\,\mathrm{d}B_r\right]_s \tag{17}$$

is a P^* martingale.

Let us suppose that $Z_t = 1 + \int_0^t H_s Z_s\,\mathrm{d}B_s$, which is reasonable since we have martingale representation for B and Z is a martingale. We then have that (17) becomes

$$\int_0^t h(s, S_s)\,\mathrm{d}B_s - \int_0^t \frac{1}{Z_s} Z_s H_s h(s, S_s)\,\mathrm{d}s$$

$$= \int_0^t h(s, S_s)\,\mathrm{d}B_s - \int_0^t H_s h(s, S_s)\,\mathrm{d}s.$$

If we choose $H_s = \frac{k(s;S_r;r\leqslant s)}{h(s,S_s)}$, then we have

$$S_t = \int_0^t h(s, S_s)\,\mathrm{d}B_s + \int_0^t k(s; S_r; r \leqslant s)\,\mathrm{d}s$$

is a martingale under P^*. Moreover, we have $M_t = B_t + \int_0^t \frac{k(s;S_r;r\leqslant s)}{h(s,S_s)}\,\mathrm{d}s$ is a P^* martingale. Since $[M, M]_t = [B, B]_t = t$, by Lévy's theorem it is a P^*-Brownian motion (see, e.g., Protter, 2005), and we have

$$\mathrm{d}S_t = h(t, S_t)\,\mathrm{d}M_t$$

and thus *S is a Markov process under P^**.

The last step in this digression is to show that it is possible to construct such a P^*! Recall that the *stochastic exponential* of a semimartingale X is the solution of the "exponential equation"

$$\mathrm{d}Y_t = Y_t\,\mathrm{d}X_t; \quad Y_0 = 1.$$

The solution is known in closed form and is given by

$$Y_t = \exp\left(X_t - \frac{1}{2}[X, X]_t^c\right) \prod_{s\leqslant t} (1 + \Delta X_s) e^{-\Delta X_s}.$$

If X is continuous then

$$Y_t = \exp\left(X_t - \frac{1}{2}[X, X]_t\right),$$

and it is denoted $Y_t = \mathcal{E}(X)_t$. Recall that we wanted $\mathrm{d}Z_t = H_t Z_t\,\mathrm{d}B_t$. Let $N_t = \int_0^t H_s\,\mathrm{d}B_s$, and we have $Z_t = \mathcal{E}(N)_t$. Then we set $H_t = \frac{-k(t;S_r,r\leqslant t)}{h(t,S_t)}$ as

planned and let $dP^* = Z_T\,dP$, and we have achieved our goal. Since $Z_T > 0$ a.s. (dP), we have that P and P^* are equivalent.

Let us now summarize the foregoing. We assume we have a price process given by

$$dS_t = h(t, S_t)\,dB_t + k(t; S_r, r \leqslant t)\,dt.$$

We form P^* by $dP^* = Z_T\,dP$, where $Z_T = \mathcal{E}(N)_T$ and $N_t = \int_0^t \frac{-k(s;S_r,r\leqslant s)}{h(s,S_s)}\,dB_s$. We let φ be the (unique) solution of the boundary value problem.

$$\frac{1}{2}h(t,x)^2\frac{\partial^2\varphi}{\partial x^2}(t,x) + \frac{\partial}{\partial s}\varphi(t,x) = 0 \tag{18}$$

and $\varphi(T, x) = f(x)$, where φ is $\mathcal{C}^2$ in x and $\mathcal{C}^1$ in t. Then

$$V_t = \varphi(t, S_t) = \varphi(0, S_0) + \int_0^t \frac{\partial\varphi}{\partial x}(s, S_s)\,dS_s.$$

Thus, under these four rather restrictive hypotheses, we have found our replication strategy! It is $a_s = \frac{\partial\varphi}{\partial x}(s, S_s)$. We have also found our value process $V_t = \varphi(t, S_t)$, provided we can solve the partial differential equation (18). However even if we cannot solve it in closed form, we can always approximate φ numerically.

Remark 2. It is a convenient hypothesis to assume that the price process S of our risky asset follows a stochastic differential equation driven by Brownian motion.

Remark 3. Although our price process is assumed to follow the SDE

$$dS_t = h(t, S_t)\,dB_t + k(t; S_r, r \leqslant t)\,dt,$$

we see that the PDE (4) does not involve the "drift" coefficient k at all! Thus the price and the replication strategy do not involve k either. The economic explanation of this is two-fold: first, the drift term k is already reflected in the market price: it is based on the "fundamentals" of the security; and second, what is important is the risk involved as reflected in the term h.

Remark 4. Hypothesis 2 is not a benign hypothesis. Since φ turns out to be the solution of a partial differential equation (given in (18)), we are asking for regularity of the solution. This is typically true when f is smooth [which of course the canonical example $f(x) = (K - x)^+$ is not!]. The problem occurs at the boundary, not the interior. Thus for reasonable f we can handle the boundary terms. Indeed this analysis works for the cases of European calls and puts as we describe in Section 3.10.

3.10 The Black–Scholes model

In Section 3.9 we saw the convenience of assuming that S solves a stochastic differential equation. Let us now assume S follows a linear SDE (= stochastic differential equation) with constant coefficients:

$$dS_t = \sigma S_t \, dB_t + \mu S_t \, dt; \quad S_0 = 1. \tag{19}$$

Let $X_t = \sigma B_t + \mu t$ and we have

$$dS_t = S_t \, dX_t; \quad S_0 = 1,$$

so that

$$S_t = \mathcal{E}(X)_t = e^{\sigma B_t + (\mu - \frac{1}{2}\sigma^2)t}.$$

The process S of (19) is known as *geometric Brownian motion* and has been used to study stock prices since at least the 1950s and the work of P. Samuelson.

In this simple case the solution of the PDE (18) of Section 3.9 *can be found explicitly*, and it is given by

$$\varphi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f\big(x e^{\sigma u \sqrt{T-t} - \frac{1}{2}\sigma^2(T-t)}\big) e^{-\frac{u^2}{2}} \, du.$$

In the case of a *European call option* we have $f(x) = (x-K)^+$ and

$$\varphi(x,t) = x\Phi\left(\frac{1}{\sigma\sqrt{T-t}}\left(\log\frac{x}{K} + \frac{1}{2}\sigma^2(T-t)\right)\right) - K\Phi\left(\frac{1}{\sigma\sqrt{T-t}}\left(\log\frac{x}{K} - \frac{1}{2}\sigma^2(T-t)\right)\right).$$

Here $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} \, du$. In the case of this call option we can also compute the replication strategy:

$$a_t = \Phi\left(\frac{1}{\sigma\sqrt{T-t}}\left(\log\frac{S_t}{K} + \frac{1}{2}\sigma^2(T-t)\right)\right). \tag{20}$$

And, we can compute the *price* (here we assume $S_0 = s$):

$$V_0 = \varphi(x,0) = x\Phi\left(\frac{1}{\sigma\sqrt{T}}\left(\log\frac{x}{K} + \frac{1}{2}\sigma^2 T\right)\right) - K\Phi\left(\frac{1}{\sigma\sqrt{T}}\left(\log\frac{x}{K} - \frac{1}{2}\sigma^2 T\right)\right). \tag{21}$$

These formulas, (20) and (21) are the celebrated *Black–Scholes option formulas* (or as we prefer to call them, *Black–Scholes–Merton option formulas*), with $R_t \equiv 1$.

This is a good opportunity to show how things change in the presence of interest rates. Let us now assume a constant interest rate r so that $R_t = e^{-rt}$. Then the formula (21) becomes:

$$V_0 = \varphi(x,0) = x\Phi\left(\frac{1}{\sigma\sqrt{T}}\left(\log\frac{x}{K} + \left(r + \frac{1}{2}\sigma^2\right)T\right)\right) - e^{-rT}K\Phi\left(\frac{1}{\sigma\sqrt{T}}\left(\log\frac{x}{K} + \left(r - \frac{1}{2}\sigma^2\right)T\right)\right).$$

These relatively simple, explicit, and easily computable formulas make working with European call and put options easy. It is perhaps because of this beautiful simplicity that security prices are often assumed to follow geometric Brownian motions, even when there is significant evidence to the contrary. Finally note that – as we observed earlier – the drift coefficient μ does not enter into the Black–Scholes formula.

3.11 Reasonable price processes

This section studies reasonable price processes, which we define to be price processes consistent with no arbitrage. The reason, of course, is that if a price process admits arbitrage, it would be unstable. Traders' actions, taking advantage of the arbitrage opportunities, would change the price process into something else (the mechanism is as discussed in Section 3.8).

Here, we consider arbitrary semimartingales as possible price processes, and we study necessary conditions for them to have no arbitrage opportunities. Because of the Delbaen–Schachermayer theory, we know that this is equivalent to finding an equivalent probability measure P^* such that a semimartingale X is a σ martingale under P^*. Note that in Section 3.9 we showed how to construct P^* by constructing the Radon–Nikodym derivative $\frac{\mathrm{d}P^*}{\mathrm{d}P}$, under the assumption that the price process followed a stochastic differential equation of a reasonable form, driven by a Brownian motion. This is of course in the case of a complete market, where P^* is unique. In the incomplete case, there are many equivalent local martingale measures, and for these cases we will indicate in Section 3.12 how to explicitly construct at least one of the equivalent probability measures such that X is a σ martingale.

Definition 14. A *reasonable price process* X is a nonnegative semimartingale on a filtered probability space satisfying 'the usual hypotheses' $(\Omega, \mathcal{F}, \mathbb{F}, P)$, such that there exists at least one equivalent probability measure P^* making X a σ martingale under P^*.

3.11.1 The continuous case

Let $X_t = X_0 + M_t + A_t$, $t \geqslant 0$ be a continuous semimartingale on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$ where $\mathbb{F} = (\mathcal{F}_t)_{t\geqslant 0}$. We seek necessary conditions (and if possible sufficient) such that there exists an equivalent probability

measure P^* where X is a P^* σ martingale. Since X is continuous, and since all continuous sigma martingales are in fact local martingales, we need only concern ourselves with local martingales. We give the appropriate theorem without proof, and instead refer the interested reader to Protter and Shimbo (2006) for a detailed proof.[23]

Theorem 12. *Let $X_t = X_0 + M_t + A_t$, $0 \leqslant t \leqslant T$ be a continuous semimartingale on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$ where $\mathbb{F} = (\mathcal{F}_t)_{0 \leqslant t \leqslant T}$. Let $C_t = [X, X]_t = [M, M]_t$, $0 \leqslant t \leqslant T$. There exists an equivalent probability measure P^* on $\mathcal{F}_T$ such that X is a P^* sigma martingale only if the following two conditions are satisfied*:

1. $\mathrm{d}A \ll \mathrm{d}C$ *a.s.*;
2. *If J is such that $A_t = \int_0^t J_s \, \mathrm{d}C_s$ for $0 \leqslant t \leqslant T$, then $\int_0^T J_s^2 \, \mathrm{d}C_s < \infty$ a.s.*;
   If in addition one has the condition below, then we have sufficient conditions for there to exist an equivalent probability measure P^* on $\mathcal{F}_T$ such that X is a P^* sigma martingale;
3. $E\{\mathcal{E}(-J \cdot M)_T\} = 1$, *where $\mathcal{E}(U)$ denotes the stochastic exponential of a semimartingale U.*

Remark 5. Recall that the decomposition of a continuous semimartingale is unique (assuming one takes the local martingale M to be continuous), so M and A are uniquely defined. If the martingale M is Brownian motion, that is if $M = B$, then since $[B, B]_t = t$ we have as a necessary condition that A must have paths which are absolutely continuous (with respect to Lebesgue measure) almost surely. This means that a semimartingale such as $X_t = 1 + |B_t|$ *cannot be a reasonable price process*, even though it is a nonnegative semimartingale, since by Tanaka's formula we have $X_t = 1 + \beta_t + L_t$ where β is another Brownian motion, and L is the local time of B at level 0.[24] We know that the paths of L are singular with respect to Lebesgue measure, a.s.

Remark 6. The sufficiency is not as useful as it might seem, because of condition (3) of Theorem 12. The first two conditions should be, in principle, possible to verify, but the third condition is in general not. On the other hand, there do exist other sufficient conditions that can be used to verify condition (3) of Theorem 12, such as Kazamaki's condition and the more well-known condition of Novikov (see, e.g., Protter, 2005 for an expository treatment of these conditions). However in practice, both of these conditions are typically quite difficult or impossible to verify, and other more ad hoc methods are used when appropriate. Typically one uses ad hoc methods to show the process in question

[23] In the following, the symbol C is not the payoff to a derivative security as it has been in previous sections.

[24] It is also trivial to construct an arbitrage strategy for this price process: if we buy and hold one share at time 0 for \$1, then at time T we have X_T dollars, and obviously $X_T \geqslant 1$ a.s., and $P(X_T > 1) = 1 > 0$.

is both positive and everywhere finite. Since these process often arise in practice as solutions of stochastic differential equations, this amounts to verifying that there are no explosions. The interested reader can consult (Cheridito et al., in press) for recent results concerning these ad hoc methods.

3.11.2 The general case

A key step in the proofs for the continuous case is the use of Girsanov's theorem. A problem in the general case is that the analog of the predictable version of Girsanov's theorem is not applicable to arbitrary semimartingales (one needs some finiteness, or integrability conditions). Therefore, one needs to use a version of Girsanov's theorem due to Jacod and Mémin, that works for random measures, and this naturally leads us to the framework of semimartingale characteristics. For background on characteristics, we refer the reader to the excellent treatment in Jacod and Shiryaev (2002).

Let X be an arbitrary semimartingale with characteristics (B, C, ν) on our usual filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, P)$, where $\mathbb{F} = (\mathcal{F}_t)_{t\geqslant 0}$. The random measure ν factors as follows: $\nu(\mathrm{d}s, \mathrm{d}x) = \mathrm{d}A_s(\omega)K_s(\omega, \mathrm{d}x)$ in such as way that we can take $C_t = \int_0^t c_s \,\mathrm{d}A_s$ and $B_t = \int_0^t b_s \,\mathrm{d}A_s$. We have the following theorem which gives necessary conditions for X to have no arbitrage in the Delbaen–Schachermayer sense of NFLVR. We give the theorem without proof; a proof can be found in Protter and Shimbo (2006).

Theorem 13. *Let P^* be another probability measure equivalent to P. Then of course X is a semimartingale under P^*, with characteristics (B^*, C, ν^*).*[25] *We then know* [*see Theorem* 3.17 (Jacod and Shiryaev, 2002, p. 170)] *that the random measure ν^* is absolutely continuous with respect to ν, and that there exists a predictable process* (*predictable in the extended sense*) $Y(s, x)_{s\geqslant 0,\, x\in\mathbb{R}}$ *such that*

$$\nu^* = Y \cdot \nu. \tag{22}$$

If X is a $P^*\sigma$ martingale, then we must have the following four conditions satisfied:

1. $b_t + \beta_t c_t + \int \{x(Y(t, x) - 1_{\{|x|\leqslant 1\}})K_t(\mathrm{d}x) = 0$; $P(\mathrm{d}\omega)\,\mathrm{d}A_s(\omega)$ *almost everywhere*;
2. $\int_0^T \beta_s^2 \,\mathrm{d}C_s < \infty$, *a.s.*;
3. $\Delta A_t > 0$ *implies that* $\int xY(s, x)K(s, \mathrm{d}x) = 0$;
4. $\int |x^2| \wedge |x| Y(t, x)K_t(\mathrm{d}x) < \infty$, $P(\mathrm{d}\omega)\,\mathrm{d}A_s(\omega)$ *almost everywhere.*

Remark 7. Distinct from the continuous case, we only have necessary conditions for P^* to exist, and not sufficient conditions. The proof of the sufficiency in the continuous case breaks down here.

[25] We write C instead of C^* because it is the same process for any equivalent probability measure.

Often we impose the assumption of *quasi left continuity*[26] of the underlying filtration. This is a standard assumption in most of Markov process theory, for example, due to Meyer's theorem (cf. Protter, 2005, p. 105). A simple example of a quasi-left continuous filtration is the natural (completed) filtration of a Lévy process.

Theorem 14. *Let X be a semimartingale as in Theorem 13. Suppose in addition that $\mathbb{F}$ is a quasi-left continuous filtration, and that A is continuous. If X is a $P^*\sigma$ martingale, then we must have the following three conditions satisfied*:

1. $b_t + \beta_t c_t + \int \{x(Y(t,x) - 1_{\{|x|\leqslant 1\}})K_t(\mathrm{d}x) = 0$; $P(\mathrm{d}\omega)\,\mathrm{d}A_s(\omega)$ *almost everywhere*;
2. $\int_0^T \beta_s^2\,\mathrm{d}C_s < \infty$, *a.s.*;
3. $\int |x^2| \wedge |x| Y(t,x) K_t(\mathrm{d}x) < \infty$, $P(\mathrm{d}\omega)\,\mathrm{d}A_s(\omega)$ *almost everywhere.*

Remark 8. Since the filtration $\mathbb{F}$ is quasi-left continuous, all martingales jump at totally inaccessible times, so the assumption that A be continuous is not a restriction on the martingale terms, but rather a condition eliminating predictable jump times in the drift. Since A is continuous, obviously we are able to remove the condition on the jumps of A.

Remark 9 *(General Remarks)*. Comparing Theorem 12 and Theorem 13 illustrates how market incompleteness corresponding to the price process X can arise in two different ways. First, Theorem 12 shows that (in the continuous case) the choice of the orthogonal martingale M is essentially arbitrary, and each such choice leads to a different equivalent probability measure rendering X a local martingale. Second, Theorem 13 shows that in the general case (the case where jumps are present) incompleteness can still arise for the same reasons as in the continuous case, but also because of the jumps, through the choice of Y. Indeed, we are free to change Y appropriately at the cost of changing b. Only if K reduces to a point mass is it then possible to have uniqueness of P^* (and hence market completeness), and then of course only if $C = 0$. *What this means is that if there are jumps in the price process, our only hope for a complete market is for there to be only one kind of jump, and no continuous martingale component*.

We also wish to remark that through clever constructions, one can indeed have complete markets with jumps in more interesting settings than point processes; see for example Dritschel and Protter (1999). In addition, one can combine (for example) a Brownian motion and a compensated Poisson process

[26] See Protter (2005, p. 191) for a definition and discussion of quasi-left continuity of a filtration. The primary implication of a filtration being quasi-left continuous is that no martingale can jump at a predictable stopping time.

via a clever trick to get a complete market which has a continuous martingale and a jump martingale as components of the price process. (See Jeanblanc and Privault, 2002.)

3.12 The construction of martingale measures

In Section 3.9 we showed how, in the special continuous case, one can construct (in that section the unique) martingale measure (also known as a risk neutral measure). We can use the same idea more generally in incomplete markets, and we illustrate this technique here. Note that we are not trying for maximum generality, and we will make some rather strong finiteness assumptions in order to keep the presentation simple. Here is the primary result.

Theorem 15. *Let S be a price process, and assume it is a special semimartingale*[27] *with canonical decomposition $S = M + A$. Assume that the conditional quadratic variation process $\langle M, M\rangle$ exists, and that $\mathrm{d}A_t \ll \mathrm{d}\langle M, M\rangle_t$ such that if $\mathrm{d}A_t = K_t\,\mathrm{d}\langle M, M\rangle_t$ for some predictable process K, then $E(e^{\int_0^T K_s^2\,\mathrm{d}\langle M,M\rangle_s}) < \infty$. Assume further that for any stopping time τ, $0 \leqslant \tau \leqslant T$, we have $K_\tau \Delta M_\tau > -1$. Let*

$$Z_t = 1 + \int_0^t Z_{s-}(-K_s)\,\mathrm{d}M_s, \quad 0 \leqslant t \leqslant T,$$

and set $\mathrm{d}P^* = Z_T\,\mathrm{d}P$. Then P^* is a equivalent martingale measure for P.

Proof. Since we know by hypothesis that $K_\tau \Delta M_\tau > -1$ for any stopping time τ with values in $[0, T]$, we have that $Z > 0$ on $[0, T]$ almost surely. Thus Z is a positive supermartingale. The hypothesis $E(e^{\int_0^T K_s^2 d\langle M,M\rangle_s}) < \infty$ allows us to assume that Z is a true martingale, by Shimbo's theorem (see Protter, 2005, p. 142, or Shimbo, 2006).[28] Therefore $E(Z_T) = 1$ and P^* is a true probability measure, equivalent to P. We therefore have, by the Girsanov–Meyer theorem, that the canonical decomposition of S under P^* is

$$S_t = \left\{ S_t - \int_0^t \frac{1}{Z_{s-}}\,\mathrm{d}\langle Z, M\rangle_s \right\} + \left\{ A_t + \int_0^t \frac{1}{Z_{s-}}\,\mathrm{d}\langle Z, M\rangle_s \right\}. \tag{23}$$

[27] A semimartingale is called *special* is it has a decomposition where the finite variation term can be taken predictable. See Protter (2005, pp. 130ff) for more information on special semimartingales.

[28] If S is assumed continuous, we have that the condition $E(e^{\frac{1}{2}\int_0^T K_s^2\,\mathrm{d}\langle M,M\rangle_s}) < \infty$ is sufficient, by Novikov's criterion.

We next note that

$$\int_0^t \frac{1}{Z_{s-}} \, \mathrm{d}\langle Z, M\rangle_s = \int_0^t \frac{1}{Z_{s-}} Z_{s-}(-K_s) \, \mathrm{d}\langle M, M\rangle_s$$

$$= -\int_0^t K_s \, \mathrm{d}\langle M, M\rangle_s \tag{24}$$

and this equals $-A_t$ by our hypothesis on K (and hence A). This implies that P^* renders S into a local martingale, and hence P^* is a choice for an equivalent martingale measure. □

Example 1. Suppose we have a price process which satisfies:

$$\mathrm{d}S_t = \sigma_1(S_t) \, \mathrm{d}B_t + \sigma_2(S_t) \, \mathrm{d}W_t + \sigma_3(S_{t-}) \, \mathrm{d}M_t + \mu(S_t) \, \mathrm{d}t; \quad S_0 > 0, \tag{25}$$

where B and W are independent Brownian motions, $M_t = N_t - \lambda t$, a compensated standard Poisson process with arrival intensity λ. We let $\mathbb{M}$ denote the sum of the three martingales. Moreover we assume that σ_1, σ_2, σ_3 and μ all bounded, Lipshitz functions. (We also assume of course that N is independent from the two Brownian motions.) To find a risk neutral measure P^*, we need only choose it in such a way as to eliminate the drift under P^*. We have four (and as we shall see hence an infinite number of) obvious choices:

1. We can choose Z to be the unique solution of

$$Z_{1,t} = 1 + \int_0^t Z_{1,s}\big(-\mu(S_s)\big) \, \mathrm{d}B_s; \quad Z_{1,0} = 1, \tag{26}$$

and take $\mathrm{d}P_1^* = Z_{1,T} \, \mathrm{d}P$. We then get, using Eqs. (23) and (24), that

$$\int_0^t \frac{1}{Z_{1,s-}} \, \mathrm{d}\langle Z_1, \mathbb{M}\rangle_s = \int_0^t \frac{1}{Z_{1,s-}} Z_{1,s-}\big(-\mu(S_s)\big) \, \mathrm{d}\langle B, \mathbb{M}\rangle_s$$

$$= -\int_0^t \mu(S_s) \, \mathrm{d}\langle B, B\rangle_s = -\int_0^t \mu(S_s) \, \mathrm{d}s \tag{27}$$

where, due to the independence assumption, the second to last equality uses $\langle B, W\rangle = \langle B, M\rangle = 0$, whence $\langle B, \mathbb{M}\rangle = \langle B, B\rangle$. Finally, we have $\mathrm{d}\langle B, B\rangle_s = \mathrm{d}[B, B]_s = \mathrm{d}s$, since B is a Brownian motion.

2. Instead, we can choose Z_2 to satisfy the SDE

$$Z_{2,t} = 1 + \int_0^t Z_{2,s}\big(-\mu(S_s)\big)\,\mathrm{d}W_s; \quad Z_{2,0} = 1.$$

This gives us a new equivalent martingale measure $\mathrm{d}P_2^* = Z_{2,T}\,\mathrm{d}P$ by the same calculations as above. In particular, we get $\langle W, \mathbb{M}\rangle = \langle W, W\rangle$ at the last step.

3. For the third example, we set

$$Z_{3,t} = 1 + \int_0^t Z_{3,s}\big(-\mu(S_s)\big)\frac{1}{\lambda}\,\mathrm{d}M_s; \quad Z_{3,0} = 1,$$

and $\mathrm{d}P_3^* = Z_{3,T}\,\mathrm{d}P$. This time we repeat the calculation of Eq. (27) to get:

$$\begin{aligned}\int_0^t \frac{1}{Z_{1,s-}}\,\mathrm{d}\langle Z_1, \mathbb{M}\rangle_s &= \int_0^t \frac{1}{Z_{1,s-}} Z_{1,s-}\big(-\mu(S_s)\big)\frac{1}{\lambda}\,\mathrm{d}\langle M, \mathbb{M}\rangle_s \\ &= -\int_0^t \mu(S_s)\frac{1}{\lambda}\,\mathrm{d}\langle M, M\rangle_s \\ &= -\int_0^t \mu(S_s)\frac{1}{\lambda}\lambda\,\mathrm{d}s,\end{aligned}$$

since $\mathrm{d}\langle M, M\rangle_s = \lambda\,\mathrm{d}s$, and of course we have used once again the independence of B, W, and M, which implies that $\langle B, M\rangle_t = 0$.

4. In addition to the three equivalent martingale measures constructed above, we can of course combine them, as follows:

$$\begin{aligned}Z_{4,t} = {} & + \int_0^t Z_{4,s-}\Big\{\alpha\big(-\mu(S_s)\big)\,\mathrm{d}B_s + \beta\big(-\mu(S_s)\big)\,\mathrm{d}W_s \\ & \quad + \gamma\big(-\mu(S_s)\big)\frac{1}{\lambda}\,\mathrm{d}M_s\Big\}; \\ & Z_{4,0} = 1,\end{aligned}$$

where α, β, and γ are all nonnegative, and $\alpha + \beta + \gamma = 1$. Then $\mathrm{d}P_4^* = Z_{4,T}\,\mathrm{d}P$.

One can imagine many more constructions, by combinations of the first three examples via random (rather than deterministic and linear) combinations.

Finally, note that these constructions, even with random combinations of the first three fundamental examples, need not exhaust the possibilities for equivalent martingale measures. Depending on the underlying filtration and probability measure, there could be martingales orthogonal[29] to B, W, and M also living on the space, which could generate orthogonal equivalent martingale measures. In this case, there is little hope to explicitly construct these alternative equivalent martingale measures with the given underlying processes. This point is made clear, but in a more abstract setting, in Section 3.11.

3.13 More complex derivatives in the Brownian paradigm: a general view

In Sections 3.9 and 3.10 we studied derivatives of the form $C = f(S_T)$, that depend only on the final value of the price process. There we showed that the computation of the price and also the hedging strategy can be obtained by solving a partial differential equation, provided the price process S is assumed to be Markov under P^*. But, this is a limited perspective. There are many other derivative securities whose payoffs depend on the entire path of the price process, and not only on the final value. In this case, the partial differential equation approach is not applicable and other techniques from the theory of stochastic processes must be applied. This section studies the techniques necessary to handle these more complex derivative securities.

We illustrate these techniques by looking at a *look-back option*, a derivative security whose payoff depends on the maximum value of the asset price S over the entire path from 0 to T. Let us return to geometric Brownian motion:

$$\mathrm{d}S_t = \sigma S_t \,\mathrm{d}B_t + \mu S_t \,\mathrm{d}t.$$

Proceeding as in Section 3.9 we change to an equivalent probability measure P^* such that $B_t^* = B_t + \frac{\mu}{\sigma} t$ is a standard Brownian motion under P^*. Now, S is a martingale satisfying:

$$\mathrm{d}S_t = \sigma S_t \,\mathrm{d}B_t^*.$$

Let F be a functional defined on $\boldsymbol{C}[0, T]$, the continuous functions with domain $[0, T]$. Then $F(u) \in \mathbb{R}$, where $u \in \boldsymbol{C}[0, T]$. Let us suppose that F is *Fréchet* differentiable and let DF denote its Fréchet derivative. Under some technical conditions on F (see, e.g., Clark, 1970), if $C = F(B^*)$, then one can show

$$C = E^*\{C\} + \int_0^T {}^p\big(DF\big(B^*; (t, T]\big)\big)\,\mathrm{d}B_t^*, \tag{28}$$

where ${}^p(X)$ denotes the predictable projection of X. [This is often written "$E^*\{X \mid \mathcal{F}_t\}$" in the literature. The process $X = (X_t)_{0 \leqslant t \leqslant T}$, $E^*\{X_t \mid \mathcal{F}_t\}$ is

[29] See Protter (2005, Section 3 of Chapter IV) for a treatment of orthogonal martingales, and in particular Protter (2005, Corollary 1, p. 183).

defined for each t a.s. The null set N_t depends on t. Thus $E^*\{X_t \mid \mathcal{F}_t\}$ does not uniquely define a process, since if $N = \bigcup_{0\leqslant t\leqslant T} N_t$, then $P(N_t) = 0$ for each t, but $P(N)$ need not be zero. The theory of predictable projections avoids this problem.]

Using (28) we then have a formula for the hedging strategy:

$$a_t = \frac{1}{\sigma S_t}{}^p\big(DF\big(\cdot, (t, T]\big)\big).$$

For the look-back option, we have the payoff: $C(\omega) = \sup_{0\leqslant t\leqslant T} S_t(\omega) = S_T^* = F(B^*)$. Then, we can let $\tau(B^*)$ denote the random time where the trajectory of S attains its maximum on $[0, T]$. Such an operation is Fréchet differentiable and

$$DF(B^*, \cdot) = \sigma F(B^*)\delta_{\tau(B^*)},$$

where δ_α denotes the Dirac measure at α.

Let

$$M_{s,t} = \max_{s\leqslant u\leqslant t}\left(B_u^* - \frac{1}{2}\sigma u\right)$$

with $M_t = M_{0,t}$. Then the Markov property gives

$$\begin{aligned} E^*\big\{DF\big(B^*, (t, T]\big) \mid \mathcal{F}_t\big\}(B^*) &= E^*\big\{\sigma F(B^*)1_{\{M_{t,T}>M_t\}} \mid \mathcal{F}_t\big\}(B^*) \\ &= \sigma S_t E^*\big\{\exp(\sigma M_{T-t});\, M_{T-t} > M_t(B^*)\big\}. \end{aligned}$$

For a given fixed value of B^*, this last expectation depends only on the distribution of the maximum of a Brownian motion with constant drift. But this distribution is explicitly known. Thus, we obtain an explicit hedging strategy for this look-back option (see Goldman et al., 1979):

$$\begin{aligned} a_t(\omega) = &\left(-\log\frac{M_t}{S_t}(\omega) + \frac{\sigma^2(T-t)}{2} + 2\right) \\ &\times \Phi\left(\frac{-\log\frac{M_t}{X_t}(\omega) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}}\right) \\ &+ \sigma\sqrt{T-t}\,\varphi\left(\frac{-\log\frac{M_t}{S_t}(\omega) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}}\right), \end{aligned}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\,du$ and $\varphi(x) = \Phi'(x)$.

The value of this look-back option is then:

$$V_0 = E^*\{C\} = S_0\left(\frac{\sigma^2 T}{2} + 2\right)\Phi\left(\frac{1}{2}\sigma\sqrt{T}\right) + \sigma\sqrt{T}S_0\varphi\left(\frac{1}{2}\sigma\sqrt{T}\right).$$

Requiring that the claim be of the form $C = F(B^*)$ where F is Fréchet differentiable is still restrictive. One can weaken this hypothesis substantially

by requiring that F be only Malliavin differentiable. If we let D denote now the Malliavin derivative of F, then Eq. (28) is still valid. Nevertheless explicit strategies and prices can be computed only in a few very special cases, and usually only when the price process S is geometric Brownian motion.

4 American type derivatives

4.1 The general view

We begin with an abstract definition, when there is *a unique equivalent martingale measure*.

Definition 15. We are given an adapted process U and an expiration time T. An *American type derivative* is a claim to the payoff U_τ at a stopping time $\tau \leqslant T$; the stopping time τ is chosen by the holder of the derivative and is called the *exercise policy*.

We let $V_t =$ the price of the security at time t. One wants to find $(V_t)_{0 \leqslant t \leqslant T}$ and especially V_0. Let $V_t(\tau)$ denote the value of the security at time t if the holder uses exercise policy τ. Let us further assume without loss of generality that $R_t \equiv 1$. Then

$$V_t(\tau) = E^*\{U_\tau \mid \mathcal{F}_t\}$$

where of course E^* denotes expectation with respect to the equivalent martingale measure P^*.

Let $\mathcal{T}(t) = \{\text{all stopping times with values in } [t, T]\}$.

Definition 16. A *rational exercise policy* is a solution to the optimal stopping problem

$$V_0^* = \sup_{\tau \in \mathcal{T}(0)} V_0(\tau). \tag{29}$$

We want to establish a price for an American type derivative. That is, how much should one pay for the right to purchase U in between $[0, T]$ at a stopping rule of one's choice?

Suppose first that the supremum in (29) is achieved. That is, let us assume there exists a rule τ^* such that $V_0^* = V_0(\tau^*)$ where V_0^* is defined in (29).

Theorem 16. *V_0^* is a lower bound for the no arbitrage price of the American type derivative.*

Proof. Suppose it is not. Let $V_0 < V_0^*$ be another price. Then one should buy the security at V_0 and use the stopping rule τ^* to purchase U at time τ^*. One

then spends $-U_{\tau^*}$, which gives an initial payoff of $V_0^* = E^*\{U_{\tau^*} \mid \mathcal{F}_0\}$; one's initial profit is $V_0^* - V_0 > 0$. This is an arbitrage opportunity. □

To prove V_0^* is also an upper bound for the no arbitrage price (and thus finally equal to the price!) is more difficult.

Definition 17. A super-replicating trading strategy θ is a self-financing trading strategy θ such that $\theta_t S_t \geqslant U_t$, all t, $0 \leqslant t \leqslant T$, where S is the price of the underlying risky security on which the American type derivative is based. (We are again assuming $R_t \equiv 1$.)

Theorem 17. *Suppose a super replicating strategy θ exists with $\theta_0 S_0 = V_0^*$. Then, V_0^* is an upper bound for the no arbitrage price of the American type derivative (U, T).*

Proof. If $V_0 > V_0^*$, then one can sell the American type derivative and adopt a super-replicating trading strategy θ with $\theta S_0 = V_0^*$. One then has an initial profit of $V_0 - V_0^* > 0$, while we are also able to cover the payment U_τ asked by the holder of the security at his exercise time τ, since $\theta_\tau S_\tau \geqslant U_\tau$. Thus we have an arbitrage opportunity. □

The existence of super-replicating trading strategies can be established using *Snell Envelopes*. A stochastic process Y is said to be of "class D" if the collection $\mathcal{H} = \{Y_\tau\colon\ \tau \text{ a stopping time}\}$ is uniformly integrable.

Theorem 18. *Let Y be a càdlàg, adapted process, $Y > 0$ a.s., and of "Class D." Then there exists a positive càdlàg supermartingale Z such that*

(i) *$Z \geqslant Y$, and for every other positive supermartingale Z' with $Z' \geqslant Y$, also $Z' \geqslant Z$;*
(ii) *Z is unique and also belongs to Class D;*
(iii) *For any stopping time τ*

$$Z_\tau = \operatorname*{ess\,sup}_{\nu \geqslant \tau} E\{Y_\nu \mid \mathcal{F}_\tau\}$$

(ν is also a stopping time).

For a proof consult Dellacherie and Meyer (1978) or Karatzas and Shreve (1998). Z is called the *Snell Envelope of Y*.

One then needs to make some regularity hypotheses on the American type derivative (U, T). For example, if one assumes U is a continuous semimartingale and $E^*\{[U, U]_T\} < \infty$, it is more than enough. One then uses the existence of Snell envelopes to prove:

Theorem 19. *Under regularity assumptions (for example $E^*\{[U, U]_T\} < \infty$ suffices), there exists a super-replicating trading strategy θ with $\theta_t S_t \geqslant k$ for all t for*

some constant k and such that $\theta_0 S_0 = V_0^*$. *A rational exercise policy is*

$$\tau^* = \inf\{t > 0 \colon Z_t = U_t\},$$

where Z is the Snell Envelope of U under P^*.

4.2 The American call option

Let us here assume that for a price process $(S_t)_{0 \leqslant t \leqslant T}$ and a bond process $R_t \equiv 1$, there exists a unique equivalent martingale measure P^* which means that there is no arbitrage and the market is complete.

Definition 18. An *American call option* with terminal time T and strike price K gives the holder the right to buy the security S at any time τ between 0 and T, at price K.

It is of course reasonable to consider the random time τ where the option is exercised at a stopping time, and the option's payoff is $(S_\tau - K)^+$, corresponding to which rule τ that the holder uses.

First, we note that since the holder of the American call option is free to choose the rule $\tau \equiv T$, he or she is always in a better position than the holder of a European call option, whose worth is $(S_T - K)^+$. Thus, the price of an American call option should be bounded below by the price of the corresponding European call option.

As in Section 4.1 we let

$$V_t(\tau) = E^*\{U_\tau \mid \mathcal{F}_t\} = E^*\big\{(S_\tau - K)^+ \mid \mathcal{F}_t\big\}$$

denote the value of our American call option at time t assuming τ is the exercise rule. The price is then

$$V_0^* = \sup_{\tau; 0 \leqslant \tau \leqslant T} E^*\big\{(S_\tau - K)^+\big\}.$$

We note however that $S = (S_t)_{0 \leqslant t \leqslant T}$ is a martingale under P^*, and since $f(x) = (x - K)^+$ is a convex function, we have $(S_t - K)^+$ is a submartingale under P^*. Hence, from (1) we have that

$$V_0^* = E^*\big\{(S_T - K)^+\big\}$$

since $t \to E^*\{(S_t - K)^+\}$ is an increasing function, and the sup – even for stopping times – of the expectation of a submartingale is achieved at the terminal time (this can be easily seen as a trivial consequence of the Doob–Meyer decomposition theorem). This leads to the following result (however the analogous result is not true for American put options, or even for American call options if the underlying stocks pay dividends):

Theorem 20. *In a complete market* (*with no arbitrage*) *the price of an American call option with terminal time T and strike price K is the same as the price for a European call option with the same terminal time and strike price.*

Theorem 21 *(Corollary). If the price process S_t follows the SDE*

$$\mathrm{d}S_t = \sigma S_t \,\mathrm{d}B_t + \mu S_t \,\mathrm{d}t;$$

then the price of an American call option with strike price K and terminal time T is the same as that of the corresponding European call option and is given by the formula (21) of Black, Scholes, and Merton.

Although the prices of the European and American call options are the same, we have said nothing about the replication strategies. But, the above theorem essentially states that the American call option is never exercised early, and hence, is identical to the European call option. Thus, their replication strategies will be identical as well.

4.3 Backwards stochastic differential equations and the American put option

Let ξ be in L^2 and suppose $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is Lipschitz in space. Then a simple backwards ordinary differential equation (ω by ω) is

$$Y_t(\omega) = \xi(\omega) + \int_t^T f\big(s, Y_s(\omega)\big)\,\mathrm{d}s.$$

However if $\xi \in L^2(\mathcal{F}_T, \mathrm{d}P)$ and one requires that a solution $Y = (Y_t)_{0 \leqslant t \leqslant T}$ be *adapted* (that is, $Y_t \in \mathcal{F}_t$), then the equation is more complex. For example, if $Y_t \in \mathcal{F}_t$ for every t, $0 \leqslant t \leqslant T$, then one has

$$Y_t = E\left\{ \xi + \int_t^T f(s, Y_s)\,\mathrm{d}s \;\middle|\; \mathcal{F}_t \right\}. \tag{30}$$

An equation such as (30) is called a *backwards stochastic differential equation*.

Next, we write

$$\begin{aligned} Y_t &= E\left\{ \xi + \int_0^T f(s, Y_s)\,\mathrm{d}s \;\middle|\; \mathcal{F}_t \right\} - \int_0^t f(s, Y_s)\,\mathrm{d}s \\ &= M_t - \int_0^t f(s, Y_s)\,\mathrm{d}s \end{aligned}$$

where M is the martingale $E\{\xi + \int_0^T f(s, Y_s)\,\mathrm{d}s \mid \mathcal{F}_t\}$. We then have

$$Y_T - Y_t = M_T - M_t - \left(\int_0^T f(s, Y_s)\,\mathrm{d}s - \int_0^t f(s, Y_s)\,\mathrm{d}s \right) \xi - Y_t$$

$$= M_T - M_t - \int_t^T f(s, Y_s)\,\mathrm{d}s$$

or, the equivalent equation:

$$Y_t = \xi + \int_t^T f(s, Y_s)\,\mathrm{d}s - (M_T - M_t). \tag{31}$$

Next, let us suppose that we are solving (30) on the canonical space for Brownian motion. Then, we have that the martingale representation property holds, and hence there exists a predictable $Z \in \mathcal{L}(B)$ such that

$$M_t = M_0 + \int_0^t Z_s\,\mathrm{d}B_s,$$

where B is Brownian motion. We have that (31) becomes

$$Y_t = \xi + \int_t^T f(s, Y_s)\,\mathrm{d}s - \int_t^T Z_s\,\mathrm{d}B_s. \tag{32}$$

Thus, to find an adapted Y that solves (30) is equivalent to find a pair (Y, Z) with Y adapted and Z predictable that solve (32).

Given Z, one can consider a more general version of (32) of the form

$$Y_t = \xi + \int_t^T f(s, Y_s, Z_s)\,\mathrm{d}s - \int_t^T Z_s\,\mathrm{d}B_s. \tag{33}$$

We wish to consider an even more general equation than (33): backward stochastic differential equations where the solution Y is forced to stay above an obstacle. This can be formulated as follows (here we follow El Karoui et al., 1997):

$$Y_t = \xi + \int_t^T f(s, Y_s, Z_s)\,\mathrm{d}s + K_T - K_t - \int_t^T Z_s\,\mathrm{d}B_s$$

where $Y_t \geqslant U_t$ (U is optional),

K is a continuous, increasing, adapted, $K_0 = 0$,

and $\int_0^T (Y_t - U_t)\,\mathrm{d}K_t = 0.$

The obstacle process U is given, as are the random variables ξ and the function f, and the unknowns are (Y, Z, K). Once again it is Z that makes both Y and K adapted.

Theorem 22 *(EKPPQ). Let f be Lipshitz in (y, z) and assume $E\{\sup_{0\leqslant t\leqslant T}(U_t^+)^2\} < \infty$. Then there exists a unique solution (Y, Z, K) to Eq. (5).*

Two proofs are given in El Karoui et al. (1997): one uses the Skorohod problem, a priori estimates and Picard iteration; the other uses a penalization method.

Now let us return to American type derivatives. Let S be the price process of a risky security and let us take $R_t \equiv 1$. For an American put option, by definition, the payoff takes the form $(K - S_\tau)^+$ where K is a strike price and the exercise rule τ is a stopping time with $0 \leqslant \tau \leqslant T$. Thus, we should let $U_t = (K - S_t)^+$, and if X is the Snell envelope of U, we see from Section 4.1 that a rational exercise policy is

$$\tau^* = \inf\{t > 0\colon\ X_t = U_t\}$$

and that the price is $V_0^* = V_0(\tau^*) = E^*\{U_{\tau^*} \mid \mathcal{F}_0\} = E^*\{(K - S_{\tau^*})^+\}$. Therefore, finding the price of an American put option is related to finding the Snell envelope of U. Recall that the Snell envelope is a supermartingale such that

$$X_\tau = \operatorname*{ess\,sup}_{\nu\geqslant\tau} E\{U_\nu \mid \mathcal{F}_\tau\}$$

where ν is also a stopping time.

We consider the situation where $U_t = (K - S_t)^+$ and $\xi = (K - S_T)^+$. We then have

Theorem 23 *(EKPPQ). Let (Y, K, Z) be the solution of (5). Then*

$$Y_t = \operatorname*{ess\,sup}_{\nu \text{ a stopping time; } t\leqslant\nu\leqslant T} E\left\{\int_t^\nu f(s, Y_s, Z_s)\,\mathrm{d}s + U_\nu \;\middle|\; \mathcal{F}_t\right\}.$$

Proof. [Sketch] In this case

$$Y_t = U_T + \int_t^T f(s, Y_s, Z_s)\,\mathrm{d}s + K_T - K_t - \int_t^T Z_s\,\mathrm{d}B_s,$$

hence

$$Y_\nu - Y_t = -\int_t^\nu f(s, Y_s, Z_s)\,\mathrm{d}s + (K_t - K_\nu) + \int_t^\nu Z_s\,\mathrm{d}B_s$$

and since $Y_t \in \mathcal{F}_t$ we have

$$Y_t = E\left\{ \int_t^\nu f(s, Y_s, Z_s)\, ds + Y_\nu + (K_\nu - K_t) \,\middle|\, \mathcal{F}_t \right\}$$
$$\geqslant E\left\{ \int_t^\nu f(s, Y_s, Z_s)\, ds + U_\nu \,\middle|\, \mathcal{F}_t \right\}.$$

Next let $\gamma_t = \inf\{t \leqslant u \leqslant T\colon\ Y_u = U_u\}$, with $\gamma_t = T$ if $Y_u > U_u$, $t \leqslant u \leqslant T$. Then

$$Y_t = E\left\{ \int_t^{\gamma_t} f(s, Y_s, Z_s)\, ds + Y_{\gamma_t} + K_{\gamma_t} - K_t \,\middle|\, \mathcal{F}_t \right\}.$$

However on $[t, \gamma_t)$ we have $Y > U$, and thus $\int_t^{\gamma_t} (Y_s - U_s)\, dK_s = 0$ implies that $K_{\gamma_{t^-}} - K_t = 0$; however K is continuous by assumption, hence $K_{\gamma_t} - K_t = 0$. Thus (using $Y_{\gamma_t} = U_{\gamma_t}$):

$$Y_t = E\left\{ \int_t^{\gamma_t} f(s, Y_s, Z_s)\, ds + U_{\gamma_t} \,\middle|\, \mathcal{F}_t \right\}$$

and we have the other implication. □

The next corollary shows that we can obtain the price of an American put option via reflected backwards stochastic differential equations.

Theorem 24 *(Corollary). The American put option has the price Y_0, where (Y, K, Z) solves the reflected obstacle backwards SDE with obstacle $U_t = (K - S_t)^+$ and where $f = 0$.*

Proof. In this case the previous theorem becomes

$$Y_0 = \operatorname*{ess\,sup}_{\nu \text{ a stopping time } 0 \leqslant \nu \leqslant T} E\{U_\nu \mid \mathcal{F}_t\},$$

and $U_\nu = (K - S_\nu)^+$. □

The relationship between the American put option and backwards SDEs can be exploited to numerically price an American put option, see Ma et al. (2002), as well as work of Bally et al. (2005), and the more recent and very promising work of Gobet et al. (in press) and also see Lemor (2005). More traditional methods are to use numerical methods associated with variational partial differential equations, Monte Carlo simulation, or lattice (binomial) type approximations.

We note that one can generalize these results to American game options (sometimes called Israeli options), using forward–backward reflected stochastic differential equations. See, e.g., Ma and Cvitanić (2001) or the "Game Options" introduced by Kifer (2000).

Acknowledgements

This chapter had a previous version, published in *Stochastic Processes and Their Applications* (Protter, 2001). Appropriate acknowledgments given there are still valid here. In addition, the second author wishes to thank Jean Jacod, Kazu Shimbo, and Denis Talay for fruitful discussions while working on this article. Peter Carr, Darrell Duffie, and Dilip Madan also helped via personal e-mail exchanges.

References

Artzner, P., Heath, D. (1995). Approximate completeness with multiple martingale measures. *Mathematical Finance* 5, 1–11.

Bally, V., Pagès, G., Printems, J. (2005). A quantization tree method for pricing and hedging multidimensional American options. *Mathematical Finance* 15, 119–168.

Bernstein, P.L. (1992). *Capital Ideas*. Free Press, New York.

Bertoin, J. (1996). *Lévy Processes*. Cambridge Univ. Press, Cambridge, UK.

Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–659.

Brown, D.J., Ross, S.A. (1991). Spanning, valuation and options. *Economic Theory* 1.

Carr, P., Wu, L. (2003). What type of process underlies options? A simple robust test. *The Journal of Finance* 53, 2581–2610.

Cheridito, P., Filipovic, D., Yor, M. (in press). Equivalent and absolutely continuous measure changes for jump-diffusion processes. *Annals of Applied Probability*.

Clark, J.M.C. (1970). The representation of functionals of Brownian motion by stochastic integrals. *Annals of Math Statistics* 41, 1282–1295.

Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 135–155.

Cox, J., Ross, S., Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7, 229–263.

Dalang, R., Morton, A., Willinger, W. (1990). Equivalent martingale measures and no arbitrage in stochastics securities market models. *Stochastics and Stochastic Reports* 29, 185–201.

Delbaen, F., Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* 300, 463–520.

Delbaen, F., Schachermayer, W. (1995). The existence of absolutely continuous local martingale measures. *Annals of Applied Probability* 5, 926–945.

Delbaen, F., Schachermayer, W. (1998). The fundamental theorem for unbounded stochastic processes. *Mathematische Annalen* 312, 215–250.

Dellacherie, C., Meyer, P.A. (1978). *Probabilities and Potential*. Elsevier, North-Holland.

Dritschel, M., Protter, P. (1999). Complete markets with discontinuous security price. *Finance and Stochastics* 3, 203–214.

Duffie, D. (2001). *Dynamic Asset Pricing Theory*, third ed. Princeton Univ. Press, Princeton.

El Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S., Quenez, M.C. (1997). Reflected solutions of backward SDEs, and related obstacle problems for PDEs. *Annals of Probability* 25, 702–737.

Föllmer, H., Sondermann, D. (1986). Hedging of nonredundant contingent claims. In: Hildebrand, W., Mas-Colell, A. (Eds.), *Contributions to Mathematical Economics*, pp. 205–223.
Geman, H., El Karoui, N., Rochet, J.-C. (1995). Changes of numéraire, change of probability measure and option pricing. *Journal of Applied Probability* 32, 443–458.
Gobet, E., Lemor, J.-P., Warin, X. (in press). A regression based Monte-Carlo method to solve backward stochastic differential equations. *Annals of Applied Probability*.
Goldman, M.B., Sosin, H., Gatto, M.A. (1979). Path dependent options: 'Buy at the low, sell at the high'. *Journal of Finance* 34, 1111–1127.
Hald, A. (1981). T.N. Thiele's contributions to statistics. *International Statistical Review* 49, 1–20.
Harrison, J.M., Kreps, D.M. (1979). Martingales and arbitrage in multiperiod security markets. *Journal of Economic Theory* 20, 381–408.
Harrison, J.M., Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11, 215–260.
Huang, C.-F. (1985). Information structures and viable price systems. *Journal of Mathematical Economics* 14, 215–240.
Jacod, J., Protter, P. (2004). *Probability Essentials*, second ed. corrected. Springer-Verlag, Heidelberg.
Jacod, J., Shiryaev, A.N. (2002). *Limit Theorems for Stochastic Processes*, second ed. Springer-Verlag, Heidelberg.
Jacod, J., Méléard, S., Protter, P. (2000). Martingale representation: Formulas and robustness. *Annals of Probability* 28, 1747–1780.
Jarrow, R.A., Protter, P. (2004). A short history of stochastic integration and mathematical finance: The early years, 1880–1970. In: *The Herman Rubin Festschrift*. In: *IMS Lecture Notes*, vol. 45, pp. 75–91.
Jarrow, R.A., Protter, P. (2007). Liquidity risk and option pricing theory. In: Birge, J.R., Linetsky, V. (Eds.), *Handbook in Operations Research and Management Science: Financial Engineering*, vol. 15. Elsevier, Amsterdam (this volume).
Jarrow, R.A., Jin, X., Madan, D.B. (1999). The second fundamental theorem of asset pricing. *Mathematical Finance* 9, 255–273.
Jeanblanc, M., Privault, N. (2002). A complete market model with Poisson and Brownian components. In: *Seminar on Stochastic Analysis, Random Fields and Applications, III*. Ascona, 1999. In: *Progr. Probab.*, vol. 52. Birkhäuser, Basel, pp. 189–204.
Jeanblanc-Piqué, M., Pontier, M. (1990). Optimal portfolio for a small investor in a market model with discontinuous prices. *Applied Mathematics and Optimization* 22, 287–310.
Karatzas, I., Shreve, S.E. (1998). *Methods of Mathematical Finance*. Springer-Verlag, New York.
Kifer, Y. (2000). Game options. *Finance and Stochastics* 4, 443–463.
Kusuoka, S. (1999). A remark on default risk models. *Advances in Mathematics and Economy* 1, 69–82.
Lemor, J.-P. (2005). Approximation par projections et simulations de Monte-Carlo des équations différentielles stochastiques rétrogrades. PhD thesis, École Polytechnique.
Ma, J., Cvitanić, J. (2001). Reflected forward–backward SDE's and obstacle problems with boundary conditions. *Journal of Applied Mathematics and Stochastic Analysis* 14, 113–138.
Ma, J., Protter, P., San Martin, J., Torres, S. (2002). A numerical method for backward stochastic differential equations. *Annals of Applied Probability* 12, 302–316.
Madan, D., Yor, M. (2002). Making Markov martingales meet marginals: With explicit constructions. *Bernoulli* 8, 509–536.
Merton, R. (1973). The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–183.
Musiela, M., Rutkowski, M. (1997). *Martingale Methods in Financial Modelling*. Springer-Verlag, Heidelberg.
Protter, P. (2001). A partial introduction to financial asset pricing theory. *Stochastic Processes and Their Applications* 91, 169–203.
Protter, P. (2005). *Stochastic Integration and Differential Equations*. Version 2.1, second ed. Springer-Verlag, Heidelberg.
Protter, P., Shimbo, K. (2006). No arbitrage and general semimartingales. In preparation.
Samuelson, P. (1965). Rational theory of warrant pricing. *Industrial Management Review* 6, 13–31.
Shimbo, K. (2006). PhD thesis, Cornell University. In preparation.

Shiryaev, A.N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*. World Scientific, Singapore.

Shreve, S.E. (2004). *Stochastic Calculus for Finance II: Continuous Time Models*. Springer-Verlag, New York.

Wall Street Journal (May 15, 2000). Page C1.

Yan, J.A. (1980). Caracterisation d'une Classe d'Ensembles Convexes de L^1 ou H^1. In: *Séminaire de Probabilités XIV*. In: *Springer Lecture Notes in Math.*, vol. 784, pp. 220–222.

This page intentionally left blank

PART II

Derivative Securities: Models and Methods

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15002-7

Chapter 2

Jump-Diffusion Models for Asset Pricing in Financial Engineering

S.G. Kou

Department of Industrial Engineering and Operations Research, Columbia University
E-mail: sk75@columbia.edu

Abstract

In this survey we shall focus on the following issues related to jump-diffusion models for asset pricing in financial engineering. (1) The controversy over tailweight of distributions. (2) Identifying a risk-neutral pricing measure by using the rational expectations equilibrium. (3) Using Laplace transforms to pricing options, including European call/put options, path-dependent options, such as barrier and lookback options. (4) Difficulties associated with the partial integro-differential equations related to barrier-crossing problems. (5) Analytical approximations for finite-horizon American options with jump risk. (6) Multivariate jump-diffusion models.

1 Introduction

There is a large literature on jump-diffusion models in finance, including several excellent books, e.g. the books by Cont and Tankov (2004), Kijima (2002). So a natural question is why another survey article is needed. What we attempt to achieve in this survey chapter is to emphasis some points that have not been well addressed in previous surveys. More precisely we shall focus on the following issues.

(1) The controversy over tailweight of distributions. An empirical motivation for using jump-diffusion models comes from the fact that asset return distributions tend to have heavier tails than those of normal distribution. However, it is not clear how heavy the tail distributions are, as some people favor power-type distributions, others exponential-type distributions. We will stress that, quite surprisingly, it is very difficult to distinguish power-type tails from exponential-type tails from empirical data unless one has extremely large sample size perhaps in the order of tens of thousands or even hundreds of thousands. Therefore,

whether one prefers to use power-type distributions or exponential-type distributions is a subjective issue, which cannot be easily justified empirically. Furthermore, this has significant implications in terms of defining proper risk measures, as it indicates that robust risk measures, such as VaR, are *desirable* for external risk management; see Heyde et al. (2006).

(2) Identifying a risk-neutral pricing measure by using the rational expectations equilibrium. Since jump-diffusion models lead to incomplete markets, there are many ways to choose the pricing measure; popular methods include mean–variance hedging, local mean–variance hedging, entropy methods, indifference pricing, etc. Here we will use the rational expectations equilibrium, which leads to a simple transform from the original physical probability to a risk-neutral probability so that we can pricing many assets, including zero-coupon bonds, stocks, and derivatives on stocks, simultaneously all in one framework.

(3) Using Laplace transforms to pricing options, including European call and put options, path-dependent options, such as barrier options and lookback options. We shall point out that even in the case of European call and put options, Laplace transforms lead to simpler expressions and even faster computations, as direct computations may involve some complicated special functions which may take some time to compute while Laplace transforms do not.

(4) Difficulties associated with the partial integro-differential equations related to barrier-crossing problems. For example: (i) Due to nonsmoothness, it is difficult to apply Itô formula and Feymann–Kac formula directly. (ii) It is generally difficult to solve the partial integro-differential equations unless the jump sizes have an exponential-type distribution. (iii) Even renewal-type arguments may not lead to a unique solution. However martingale arguments may be helpful in solving the problems.

(5) Two analytical approximations for finite-horizon American options, which can be computed efficiently and with reasonable accuracy.

(6) Multivariate jump-diffusion models.

In a survey article, inevitably I will skip some important topics which are beyond the expertise of the author. For example, I will omit numerical solutions for jump-diffusion models; see Cont and Tankov (2004), Cont and Voltchkova (2005) and d'Halluin et al. (2003) on numerical methods for solving partial integro-differential equations, and Feng and Linetsky (2005) and Feng et al. (2004) on how to price path-dependent options numerically via variational methods and extrapolation. Two additional topics omitted are hedging (for a survey, see the book by Cont and Tankov, 2004) and statistical inference and econometric analysis for jump-diffusion models (for a survey, see the book by Singleton, 2006). Due to the page limit, I will also skip various applications of the jump-diffusion models; see the references in Glasserman and Kou (2003) for applications of jump-diffusion models in fixed income derivatives and term

structure models, and Chen and Kou (2005) for applications in credit risk and credit derivatives.

2 Empirical stylized facts

2.1 Are returns normally distributed

Consider the daily closing prices of S&P 500 index (SPX) from Jan 2, 1980 to Dec 31, 2005. We can compute the daily returns of SPX, either using the simple returns or continuously compounded returns. The (one-period) simple return is defined to be $R_t = \{S(t) - S(t-1)\}/S(t-1)$ at time t, where $S(t)$ is the asset price. For mathematical convenience, the continuously compounded return (also called log return) at time t, $r_t = \ln \frac{S(t)}{S(t-1)}$, is very often also used, especially in theoretical modeling. The difference between simple and log returns for daily data is quite small, although it could be substantial for monthly and yearly data. The normalized daily simple returns are plotted in Fig. 1, so that the daily simple returns will have mean zero and standard deviation one.

We see big spikes in 1987. In fact the max and min (which all occurred during 1987) are about 7.9967 and −21.1550 standard deviation. The continuously compounded returns show similar features. Note that for a standard normal

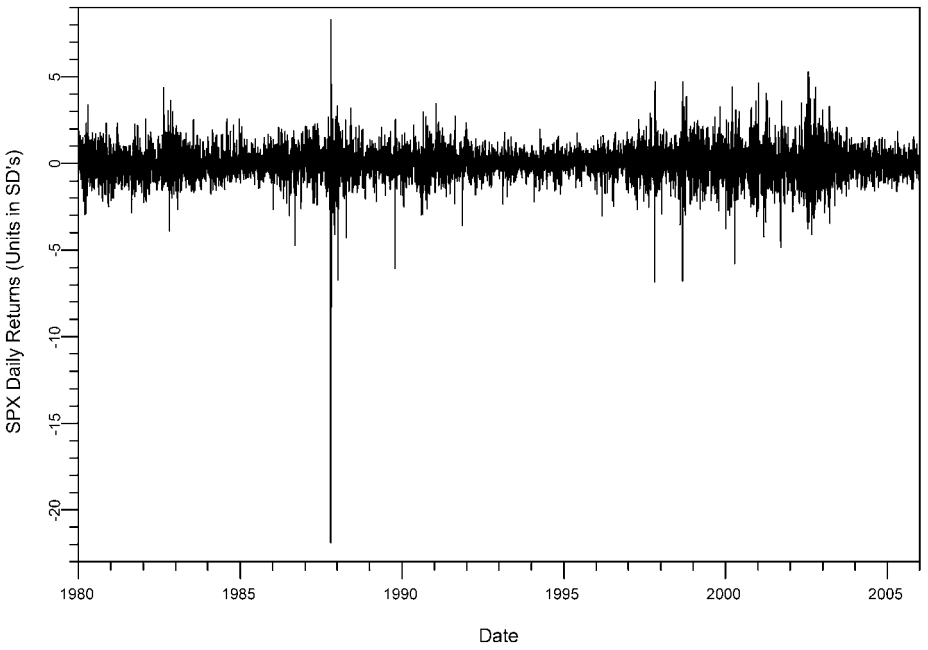


Fig. 1. The normalized daily simple returns of S&P 500 index from Jan 2, 1980 to Dec 31, 2005. The returns have been normalized to have mean zero and standard deviation one.

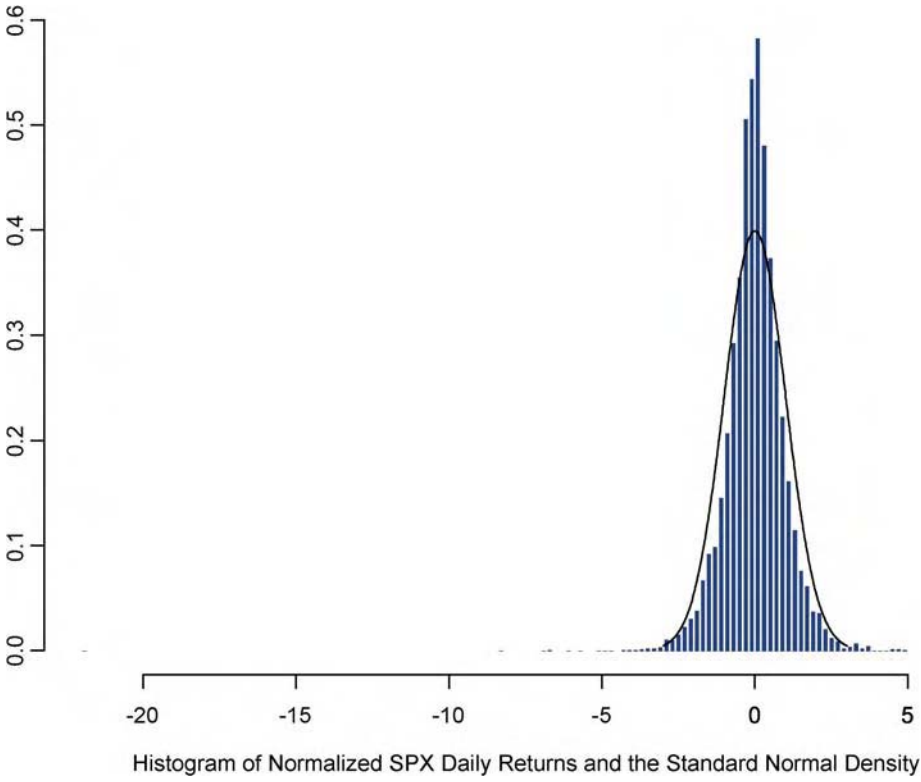


Fig. 2. Comparison of the histogram of the normalized daily returns of S&P 500 index (from Jan 2, 1980 to Dec 31, 2005) and the density of $N(0, 1)$. The feature of a high peak and two heavy tails (i.e. the leptokurtic feature) is quite evident.

random variable Z, $P(Z < -21.1550) \approx 1.4 \times 10^{-107}$; as a comparison, note that the whole universe is believed to have existed for 15 billion years or 5×10^{17} seconds.

Next we plot the histogram of the daily returns of SPX. Figure 2 displays the histogram along with the standard normal density function, which is essentially confined within $(-3, 3)$.

2.1.1 Leptokurtic distributions

Clearly the histogram of SPX displays a high peak and asymmetric heavy tails. This is not only true for SPX, but also for almost all financial asset prices, e.g. US and world wide stock indices, individual stocks, foreign exchange rates, interest rates. In fact it is so evident that a name "leptokurtic distribution" is given, which means the kurtosis of the distribution is large. More precisely, the kurtosis and skewness are defined as $K = E\big(\frac{(X-\mu)^4}{\sigma^4}\big)$, $S = E\big(\frac{(X-\mu)^3}{\sigma^3}\big)$; for the standard normal density $K = 3$. If $K > 3$ then the distribution will be called

leptokurtic and the distribution will have a higher peak and two heavier tails than those of the normal distribution. Examples of leptokurtic distributions include: (1) double exponential distribution with the density given by

$$f(x) = p \cdot \eta_1 e^{-x\eta_1} 1_{\{x>0\}} + (1-p) \cdot \eta_2 e^{x\eta_2} 1_{\{x<0\}}.$$

(2) t-distribution, etc.

To estimate skewness and kurtosis, we shall use

$$\hat{S} = \frac{1}{(n-1)\hat{\sigma}^3} \sum_{i=1}^{n} (X_i - \bar{X})^3, \qquad \hat{K} = \frac{1}{(n-1)\hat{\sigma}^4} \sum_{i=1}^{n} (X_i - \bar{X})^4$$

as sample skewness and sample kurtosis, where $\hat{\sigma}$ is the sample standard deviation. For the daily returns of the SPX data, the sample kurtosis is about 42.23. The skewness is about -1.73; the negative skewness means the return has a heavier left tail than the right tail.

The leptokurtic feature has been observed since 1950's. However classical finance models simply ignore this feature. For example, in the Black–Scholes Brownian motion model, the stock price is modeling as a geometric Brownian motion, $S(t) = S(0)e^{\mu t + \sigma W(t)}$, where the Brownian motion $W(t)$ has a normal distribution with mean 0 and variance t. Here μ is called the drift, which measures the average return, and σ is called the volatility which measures the standard deviation of the return distribution. In this model, the continuous compounded return, $\ln(S(t)/S(0))$, has a normal distribution, which it is not consistent with leptokurtic feature. Many alternative models, e.g. models with jumps and/or stochastic volatility, have been proposed to incorporate the feature, as we will discuss some of them shortly.

2.1.2 Power tails and exponential tails

It is clear that the returns of stocks have two tail distributions heavier than those of normal distribution. However, how heavy the stock tail distributions are is a debatable question. Two main classes proposed in the literature are power-type tails and exponential-type tails. For example, we say that the right tail of a random variable X has a power-type tail if $P(X > x) \approx \frac{c}{x^\alpha}$, $x > 0$, as $x \to \infty$, and the left tail of X has a power-type tail if $P(X < -x) \approx \frac{c}{x^\alpha}$, $x > 0$, as $x \to \infty$. Similarly, we say that X has a right exponential-type tail if $P(X > x) \approx ce^{-\alpha x}$, $x > 0$, and a left exponential-type tail if $P(X < -x) \approx ce^{-\alpha x}$, $x < 0$, as $x \to \infty$.

As pointed out by Kou (2002, p. 1090), one problem with using power-type right tails in modeling return distributions is that the power-type right tails cannot be used in models with continuous compounding. More precisely, suppose that, at time 0, the daily return distribution X has a power-type right tail. Then in models with continuous compounding, the asset price tomorrow $A(\Delta t)$ is given by $A(\Delta t) = A(0)e^X$. Since X has a power-type right tail, it is clear that $\mathsf{E}(e^X) = \infty$. Consequently,

$$\mathsf{E}\big(A(\Delta t)\big) = \mathsf{E}\big(A(0)e^X\big) = A(0)\mathsf{E}\big(e^X\big) = \infty.$$

In other words, the asset price tomorrow has an infinite expectation! The price of call option may also be infinite, if under the risk-neutral probability the return has a power-type right distribution. This is because

$$\mathsf{E}^*\big[\big(S(T)-K\big)^+\big] \geqslant \mathsf{E}^*\big[S(T)\big]-K=\infty.$$

In particular, these paradoxes hold for any t-distribution with any degrees of freedom which has power tails, as long as one considers financial models with continuous compounding. Therefore, the only relevant models with t-distributed returns outside these paradoxes are models with discretely compounded simple returns. However, in models with discrete compounding analytical solutions are in general impossible.

2.1.3 *Difficulties in statistically distinguish power-type tails from exponential-type tails*

Another interesting fact is that, for a sample size of 5000 (corresponding to about 20 years of daily data), it may be very difficult to distinguish empirically the exponential-type tails from power-type tails, although it is quite easy to detect the differences between them and the tails of normal density; see Heyde and Kou (2004). A good intuition may be obtained by simply looking at the quantile tables for both standardized Laplace and standardized t-distributions with mean zero and variance one. Recall that a Laplace distribution has a symmetric density $f(x) = \frac{1}{2}e^{-x}I_{[x>0]} + \frac{1}{2}e^{x}I_{[x<0]}$. The right quantiles for the Laplace and normalized t densities with degrees of freedom from 3 to 7 are given in Table 1.

Table 1 shows that the Laplace distribution may have higher tail probabilities than those of t-distributions with low degrees of freedom, even if asymptotically the Laplace distribution should have lighter tails than those of t-distributions. For example, the 99.9% percentile of the Laplace distribution is actually bigger than that of t-distribution with d.f. 6 and 7! Thus, regardless of the sample size, the Laplace distribution may appear to be heavier tailed than a t-distribution with d.f. 6 or 7, up to the 99.9% percentile. In order to distinguish the distributions it is necessary to use quantiles with very low p values and correspondingly large samples.

If the true quantiles have to be estimated from data, then the problem is even more serious, as confidence intervals need to be considered, resulting

Table 1.
Percentiles of Laplace and t-distributions

Prob.	Laplace	t7	t6	t5	t4	t3
1%	**2.77**	2.53	2.57	2.61	2.65	2.62
0.1%	**4.39**	4.04	4.25	**4.57**	5.07	5.90
0.01%	**6.02**	5.97	**6.55**	7.50	9.22	12.82
0.001%	**7.65**	**8.54**	9.82	12.04	16.50	27.67

in sample sizes typically in the tens of thousands or even hundreds of thousands necessary to distinguish power-type tails from exponential-type tails; see Heyde and Kou (2004).

2.1.4 *Practical implications for risk measures*

The difficulties in distinguishing tail distributions also have implications in risk management. For example, a controversy in axiomatic approaches to risk measures is whether one use should Value-at-Risk (or VaR), which is a measure based on quantiles, or the tail conditional expectation. Unlike the tail conditional expectation, VaR does not in general satisfy an axiom of subadditivity (Artzner et al., 1999). However, VaR is more robust against model assumptions and misspecifications, thus making the VaR more suitable to be used for external risk regulations, because VaR can produce more consistent results which are essential for external law enforcement (see Heyde et al., 2006). Furthermore, VaR at a higher quantile (e.g. 97.5%) can also be represented as tail conditional median (e.g. at 95%), thus taking into consideration of the loss beyond the threshold just as the tail conditional mean does. Indeed, VaR is widely used in practice, e.g. in the recent Basel (II) governmental regulation.

It can be shown that VaR also satisfies a different set of axioms based on commonotonic subadditivity (Heyde et al., 2006), which is consistent with both prospect theory in behavior finance and robustness requirement for external law enforcement. Furthermore, the intuition behind subadditivity (which is the theoretical basis for "coherent risk measures" such as tail conditional expectations) that merger reduces risk is not true in general, in particular in presence of the limited liability law. For details, see Heyde et al. (2006).

In short, although one may use various risk measures for internal risk management, robust risk measures, such as VaR, are needed for external risk regulations. In addition, VaR, though simple, is not irrational because it also satisfies a different set of axioms.

2.2 *Are stock returns predictable: introduction to the dependent structure of stock returns*

To study the question on whether future stock returns can be predicted from the current returns, we can formulate this question mathematically by asking whether returns are correlated in some ways, so that the current returns will provide some information about future returns. For a weakly stationary discrete time series $\{r_t\}$, where the index $t \in (-\infty, \infty)$ can only take integer values (i.e. we have $\ldots, r_{-2}, r_{-1}, r_0, r_1, r_2, \ldots$), we can define the lag-k autocovariance $\gamma_k = \mathrm{Cov}(r_t, r_{t-k}) = \mathrm{Cov}(r_t, r_{t+k})$; the two covariances are equal due to the definition of the weak stationarity. Similarly, we can define lag-k autocorrelation ρ_k:

$$\rho_k = \operatorname{Cor} r(r_t, r_{t-k}) = \operatorname{Cor} r(r_t, r_{t+k}) = \frac{\operatorname{Cov}(r_t, r_{t-k})}{\sqrt{\operatorname{Var}(r_t)\operatorname{Var}(r_{t-k})}} = \frac{\gamma_k}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_k}{\gamma_0}.$$

We can estimate ρ_k by

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(r_t - \bar{r})(r_{t-k} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2}.$$

A plot of $\hat{\rho}_k$, for $k \geqslant 1$, is called autocorrelation function (or ACF) plot. An autocorrelation plot of the simple daily returs of SPX, normalized to have mean 0 and variance 1, in given in Fig. 3.

Note that the two dotted lines in Fig. 3 indicate the 95% significant levels for autocorrelation. More precisely, if $r_t = \mu + a_t$, where a_t is a sequence of i.i.d. random variables with finite mean and variance, then as the total time period $T \to \infty$, it can be shown that $\hat{\rho}_k$ is asymptotic normal with mean 0 and variance $1/T$. This is what is plotted in Fig. 3 as the two dotted lines, which are $\pm 1.96/\sqrt{T}$, as a 95% c.i. for the autocorrelation functions in the above ACF plot.

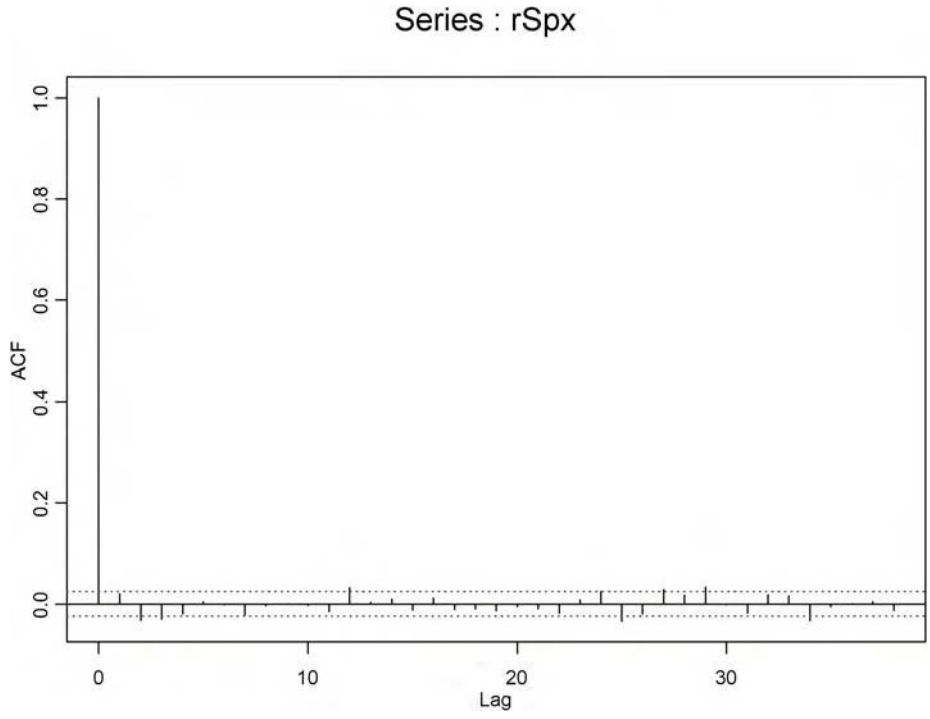


Fig. 3. The ACF plot of the returns of S&P 500 index from Jan 2, 1980 to Dec 31, 2005.

Graphically we can see from the plot that, although the first few autocorrelations significantly exceed the 95% confidence interval, the magnitude of autocorrelations is quite small, only about -0.05 to 0.05 among daily returns; it is even smaller for weekly and monthly returns.

Because of this, many finance models simply ignore the dependent structure, and assume that the stock returns have zero autocorrelations. This is, for example, in the case of Black–Scholes option pricing model, and in the capital asset pricing model, etc. Indeed, most of the classical models assume that the stock prices satisfy "a random walk hypothesis" with independent asset returns. However, starting in 1980's, researches reveal some fascinating dependent structures among asset returns.

In Figs. 4 and 5 we see the autocorrelations for the absolute values and the squared values of the SPX daily returns are quite large. This suggests that returns distributions are dependent in an interesting way that the volatility of returns (which are related to the squared returns) are correlated, but asset returns themselves have almost no autocorrelation. In the literature this is called "volatility clustering effect."

This in particular implies that any model for stock returns with independent increments (such as Lévy processes) cannot incorporate the volatility clustering effect. Since jump-diffusion models are special cases of Lévy processes,

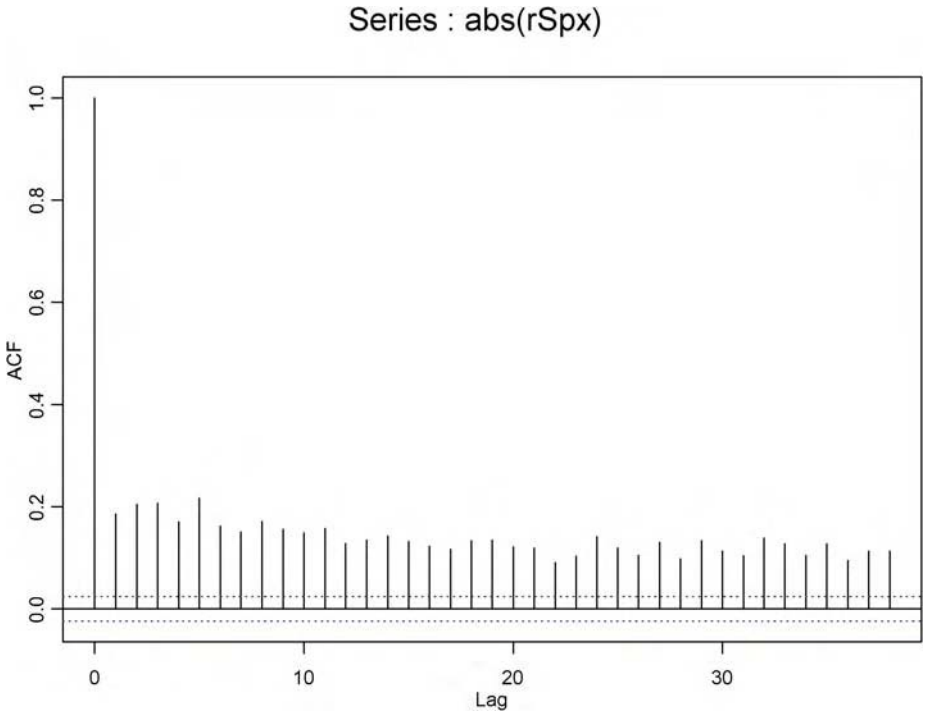


Fig. 4. The ACF plot of the absolute returns of S&P 500 index from Jan 2, 1980 to Dec 31, 2005.

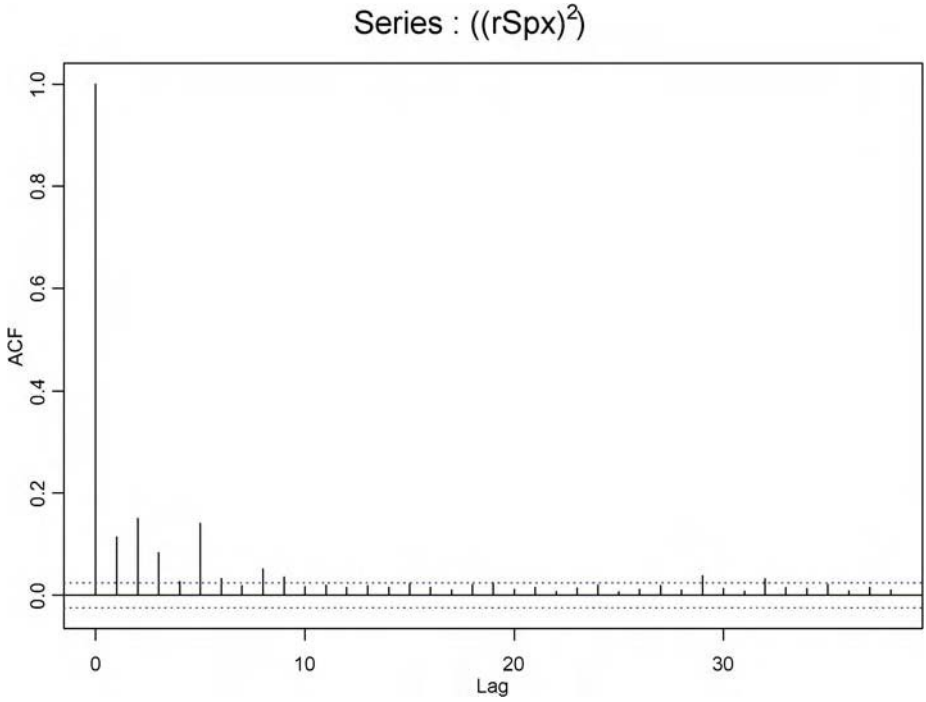


Fig. 5. The ACF plot of the squared returns of S&P 500 index from Jan 2, 1980 to Dec 31, 2005.

they cannot incorporate the volatility clustering effect directly. However, one can combine jump-diffusion processes with other processes (e.g. stochastic volatility) or consider time-changed Lévy processes to incorporate the volatility clustering effect.

2.3 Implied volatility smile

Because in the Black–Scholes formula the call and put option prices are monotone increasing functions of the volatility, we can define an inverse function that maps from a given option price to the volatility parameter, assuming that we know the other parameters in the formula. More precisely, the implied volatility $\sigma(T, K)$ is a parameter associated with a particular strike K and a particular maturity T such that if we use it as the volatility parameter in the Black–Scholes formula for European call and put options, then we should obtain a price that exactly matches the market price of a particular call/put option. In other words $\sigma(T, K)$ is the inverse function of the market option price in terms of volatility.

One immediate question is that whether the above definition is self-consistent. In particular, suppose that one person computes $\sigma(T, K)$ from a call option with maturity T and strike K, and another computes $\sigma(T, K)$ from

a put option with the same maturity T and strike K, will the two people get the same answer? The answer is yes due to the put–call parity, at least in theory. The put–call parity says that no arbitrage implies that the stock price $S(0)$, call price $C(S, K)$, the price $P(S, K)$, and the zero coupon bond price $B(T)$ must satisfy

$$S(0) = C(S, K) - P(S, K) + K \cdot B(T).$$

The relationship is model-free; in other words, no matter what model we use the above put–call parity must hold to prevent arbitrage.

Let $C_{BS}(S, K)$ and $P_{BS}(S, K)$ denote the call and put prices given by the Black–Scholes formula based on the same input variable σ. Then $S(0) = C_{BS}(S, K) - P_{BS}(S, K) + K \cdot B(T)$. Similarly, $S(0) = C_M(S, K) - P_M(S, K) + K \cdot B(T)$, where $C_M(S, K)$ and $P_M(S, K)$ denote the market prices of the call and put. Taking the difference between the two equations, we get

$$C_{BS}(S, K) - C_M(S, K) = P_{BS}(S, K) - P_M(S, K). \tag{1}$$

Now suppose we get the implied volatility $\sigma_c(T, K)$ from the market call option and the implied volatility $\sigma_p(T, K)$ from the market put option. By the definition of implied volatility, if we use $\sigma_c(T, K)$ then we must have $C_{BS}(S, K) - C_M(S, K) = 0$, if $\sigma = \sigma_c(T, K)$. By (1), we must have $P_{BS}(S, K) - P_M(S, K) = 0$, if $\sigma = \sigma_c(T, K)$. Since $\sigma_p(T, K)$ is the unique volatility such that $P_{BS}(S, K) - P_M(S, K) = 0$, we must have $\sigma_p(T, K) = \sigma_c(T, K)$. This shows that the implied volatilities from otherwise identical call and put options must be the same. Of course, in practice, we do have bid–ask spreads for options. So the implied volatility will be different depending whether you use a bid price, an ask price or the average of the bid–ask prices. Therefore, the implied volatilities from otherwise identical call and put options may also be somewhat different.

When one uses the implied volatilities from call and put options to price other options not traded in exchanges, effectively we want to do extrapolation from prices of liquidated options to get prices of less liquidated options. Many practitioners think that implied volatilities are better than the historical volatilities for the purpose of option pricing, as historical volatilities may not reflect the current situation. For example, suppose an extreme event happens to the Wall Street, e.g. a financial crisis, a terrorist attack, etc., then it is hard to find similar events in the historical database, thus making historical volatilities unsuitable.

We can calculate implied volatilities from the market prices of options with different strike prices and maturities. If the geometric Brownian motion assumption is correct, then the implied volatilities should be the same for all the options on the same underlying asset. However, empirically options on the same underlying asset but with different strike prices or maturities tend to have different implied volatilities.

In particular, it is widely recognized that if we plot implied volatilities against strike prices, then the implied volatility curve resembles a "smile,"

meaning the implied volatility is a convex curve of the strike price. In addition, the "smile" curve changes for different maturities. While mispricing exists and statistically significant, the implied volatility smile was not economically significant in early tests before the 1987 market crash (e.g. MacBeth and Merville, 1979; Rubinstein, 1985). However, after the 1987 crash, the implied volatility smile becomes economically significant and the performance of the Black–Scholes model deteriorated.

It is worth mentioning that the leptokurtic features under a risk-neutral measure lead to the "volatility smiles" in option prices; and the volatility clustering effect may lead to implied volatility smile across maturities, especially for long maturity options.

3 Motivation for jump-diffusion models

3.1 Alternative models to the Black–Scholes

Many studies have been conducted to modify the Black–Scholes model to explain the above three empirical stylized facts, namely the leptokurtic feature, volatility clustering effect, and implied volatility smile. Below is a list of some of them.

(a) Chaos theory and fractal Brownian motions. In these models, one typically replaces the Brownian motion by a fractal Brownian motion which has dependent increments (rather than independent increments); see, for example, Mandelbrot (1963). However, as Rogers (1997) pointed out these models may lead to arbitrage opportunities.

(b) Generalized hyperbolic models, including log t model and log hyperbolic model, and stable processes. These models replace the normal distribution assumption by some other distributions; see, for example, Barndorff-Nielsen and Shephard (2001), Samorodnitsky and Taqqu (1994), Blattberg and Gonedes (1974).

(c) Models based on Lévy processes; see, for example, Cont and Tankov (2004) and reference therein.

(d) Stochastic volatility and GARCH models; see, for example, Hull and White (1987), Engle (1995), Fouque et al. (2000), Heston (1993). These models are mainly designed to capture the volatility clustering effect. A typical example of these models is

$$\frac{\mathrm{d}S(t)}{S(t)} = \mu\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_1(t),$$
$$\mathrm{d}\sigma(t) = -\alpha\big(\sigma(t) - \beta\big)\,\mathrm{d}t + \gamma\sqrt{\sigma(t)}\,\mathrm{d}W_2(t),$$

where $W_1(t)$ and $W_2(t)$ are two correlated Brownian motions.

(e) Constant elasticity of variance (CEV) model; see, for example, Cox and Ross (1976) and Davydov and Linetsky (2001). In this model

$$\mathrm{d}S(t) = \mu S(t)\,\mathrm{d}t + \sigma(t)S^{\alpha}(t)\,\mathrm{d}W_1(t), \quad 0 < \alpha \leqslant 1.$$

(f) Jump-diffusion models proposed by Merton (1976) and Kou (2002).

$$S(t) = S(0)e^{(\mu-\frac{1}{2}\sigma^2)t+\sigma W(t)} \prod_{i=1}^{N(t)} e^{Y_i},$$

where $N(t)$ is a Poisson process. In Merton (1976) model, Y has a normal distribution, and in Kou (2002) it has a double exponential distribution. The double exponential distribution enables us to get analytical solutions for many path-dependent options, including barrier and lookback options, and analytical approximations for American options, as we will see later.

(g) A numerical procedure called "implied binomial trees"; see, for example, Derman and Kani (1994) and Dupire (1994).

There are models combining several features, such as stochastic volatility, jumps, and time changes. Below are two examples of them.

(h) Time changed Brownian motions and time changed Lévy processes. In these models, the asset price $S(t)$ is modeled as

$$S(t) = G\big(M(t)\big),$$

as G is a either geometric Brownian motion or a Lévy process, and $M(t)$ is a nondecreasing stochastic process modeling the stochastic activity time in the market. The activity process $M(t)$ may link to trading volumes. See, for example, Clark (1973), Madan and Seneta (1990), Madan et al. (1998), Heyde (2000), Carr et al. (2003).

(i) Affine stochastic-volatility and affine jump-diffusion models; see, for example, Duffie et al. (2000), which combines both stochastic volatilities and jump-diffusions.

3.2 Jump-diffusion models

In jump-diffusion models under the physical probability measure P the asset price, $S(t)$, is modeled as

$$\frac{\mathrm{d}S(t)}{S(t-)} = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) + \mathrm{d}\left(\sum_{i=1}^{N(t)}(V_i - 1)\right), \tag{2}$$

where $W(t)$ is a standard Brownian motion, $N(t)$ is a Poisson process with rate λ, and $\{V_i\}$ is a sequence of independent identically distributed (i.i.d.) nonnegative random variables. In the model, all sources of randomness, $N(t)$, $W(t)$, and Y's, are assumed to be independent. Solving the stochastic differential equation (2) gives the dynamics of the asset price:

$$S(t) = S(0)\exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right\} \prod_{i=1}^{N(t)} V_i. \tag{3}$$

In Merton (1976) model, $Y = \log(V)$ has a normal distribution. In Kou (2002) $Y = \log(V)$ has an asymmetric double exponential distribution with the density

$$f_Y(y) = p \cdot \eta_1 e^{-\eta_1 y} 1_{\{y \geqslant 0\}} + q \cdot \eta_2 e^{\eta_2 y} 1_{\{y<0\}}, \quad \eta_1 > 1,\ \eta_2 > 0,$$

where $p, q \geqslant 0$, $p+q = 1$, represent the probabilities of upward and downward jumps. The requirement $\eta_1 > 1$ is needed to ensure that $\mathsf{E}(V) < \infty$ and $\mathsf{E}(S(t)) < \infty$; it essentially means that the average upward jump cannot exceed 100%, which is quite reasonable. For notational simplicity and in order to get analytical solutions for various option pricing problems, the drift μ and the volatility σ are assumed to be constants, and the Brownian motion and jumps are assumed to be one-dimensional. Ramezani and Zeng (2002) independently propose the double exponential jump-diffusion model from an econometric viewpoint as a way of improving the empirical fit of Merton's normal jump-diffusion model to stock price data.

There are two interesting properties of the double exponential distribution that are crucial for the model. First, it has the leptokurtic feature; see Johnson et al. (1995). The leptokurtic feature of the jump size distribution is inherited by the return distribution. Secondly, a unique feature, also inherited from the exponential distribution, of the double exponential distribution is the memoryless property. This special property explains why the closed-form solutions (or approximations) for various option pricing problems, including barrier, lookback, and perpetual American options, are feasible under the double exponential jump-diffusion model, while it seems difficult for many other models, including the normal jump-diffusion model.

3.3 Why jump-diffusion models

Since essentially all models are "wrong" and rough approximations of reality, instead of arguing the "correctness" of a particular model we shall evaluate jump-diffusion models by four criteria.

(1) A model must be internally self-consistent. In the finance context, it means that a model must be arbitrage-free and can be embedded in an equilibrium setting. Note that some of the alternative models may have arbitrage opportunities, and thus are not self-consistent (e.g. the arbitrage opportunities for fractal Brownian motions as shown by Rogers, 1997). In this regard, both the Merton's normal jump-diffusion model and the double exponential jump-diffusion model can be embedded in a rational expectations equilibrium setting.

(2) A model should be able to capture some important empirical phenomena. However, we should emphasize that empirical tests should not be used as the only criterion to judge a model good or bad. Empirical tests tend to favor models with more parameters. However, models with many parameters tend to make calibration more difficult (the calibration may involve high-dimensional numerical optimization with many

local optima), and tend to have less tractability. This is a part of the reason why practitioners still like the simplicity of the Black–Scholes model. Jump-diffusion models are able to reproduce the leptokurtic feature of the return distribution, and the "volatility smile" observed in option prices (see Kou, 2002). The empirical tests performed in Ramezani and Zeng (2002) suggest that the double exponential jump-diffusion model fits stock data better than the normal jump-diffusion model, and both of them fit the data better than the classical geometric Brownian motion model.

(3) A model must be simple enough to be amenable to computation. Like the Black–Scholes model, the double exponential jump-diffusion model not only yields closed-form solutions for standard call and put options, but also leads to a variety of closed form solutions for path-dependent options, such as barrier options, lookback options, perpetual American options (see Kou and Wang, 2003, 2004; Kou et al., 2005), as well as interest rate derivatives (see Glasserman and Kou, 2003).

(4) A model must have some (economical, physical, psychological, etc.) interpretation. One motivation for the double exponential jump-diffusion model comes from behavioral finance. It has been suggested from extensive empirical studies that markets tend to have both overreaction and underreaction to various good news or bad news (see, for example, Fama, 1998 and Barberis et al., 1998, and references therein). One may interpret the jump part of the model as the market response to outside news. More precisely, in the absence of outside news the asset price simply follows a geometric Brownian motion. Good or bad news arrive according to a Poisson process, and the asset price changes in response according to the jump size distribution. Because the double exponential distribution has both a high peak and heavy tails, it can be used to model both the overreaction (attributed to the heavy tails) and underreaction (attributed to the high peak) to outside news. Therefore, the double exponential jump-diffusion model can be interpreted as an attempt to build a simple model, within the traditional random walk and efficient market framework, to incorporate investors' sentiment. Interestingly enough, the double exponential distribution has been widely used in mathematical psychology literature, particularly in vision cognitive studies; see, for example, papers by David Mumford and his authors at the computer vision group, Brown University.

Incidentally, as a by product, the model also suggests that the fact of markets having both overreaction and underreaction to outside news can lead to the leptokurtic feature of asset return distribution.

There are many alternative models that can satisfy at least some of the four criteria listed above. A main attraction of the double exponential jump-diffusion model is its simplicity, particularly its analytical tractability for path-dependent options and interest rate derivatives. Unlike the original Black–Scholes model, many alternative models can only compute prices for standard

call and put options, and analytical solutions for other equity derivatives (such as path-dependent options) are unlikely. Even numerical methods for interest rate derivatives and path-dependent options are not easy, as the convergence rates of binomial trees and Monte Carlo simulation for path-dependent options are typically much slower than those for call and put options (for a survey, see Boyle et al., 1997). This makes it harder to persuade practitioners to switch from the Black–Scholes model to more realistic alternative models. The double exponential jump-diffusion model attempts to improve the empirical implications of the Black–Scholes model, while still retaining its analytical tractability.

3.4 Shortcoming of jump-diffusion models

The main problem with jump-diffusion models is that they cannot capture the volatility clustering effects, which can be captured by other models such as stochastic volatility models. Jump-diffusion models and the stochastic volatility model complement each other: the stochastic volatility model can incorporate dependent structures better, while the double exponential jump-diffusion model has better analytical tractability, especially for path-dependent options and complex interest rate derivatives. For example, one empirical phenomenon worth mentioning is that the daily return distribution tends to have more kurtosis than the distribution of monthly returns. As Das and Foresi (1996) point out, this is consistent with models with jumps, but inconsistent with stochastic volatility models. More precisely, in stochastic volatility models (or essentially any models in a pure diffusion setting) the kurtosis decreases as the sampling frequency increases; while in jump models the instantaneous jumps are independent of the sampling frequency. This, in particular, suggests that jump-diffusion models may capture short-term behavior better, while stochastic volatility may be more useful to model long term behavior.

More general models combine jump-diffusions with stochastic volatilities resulting in "affine jump-diffusion models," as in Duffie et al. (2000) which can incorporate jumps, stochastic volatility, and jumps in volatility. Both normal and double exponential jump diffusion models can be viewed as special cases of their model. However, because of the special features of the exponential distribution, the double exponential jump-diffusion model leads to analytical solutions for path-dependent options, which are difficult for other affine jump-diffusion models (even numerical methods are not easy). Furthermore, jump-diffusion models are simpler than general affine jump-diffusion models; in particular jump-diffusion model have fewer parameters that makes calibration easier. Therefore, jump-diffusion models attempt to strike a balance between reality and tractability, especially for short maturity options and short term behavior of asset pricing.

In summary, many alternative models may give some analytical formulae for standard European call and put options, but analytical solutions for interest rate derivatives and path-dependent options, such as perpetual American options, barrier and lookback options, are difficult, if not impossible. In the dou-

ble exponential jump-diffusion model analytical solution for path-dependent options are possible. However, the jump-diffusion models cannot capture the volatility clustering effect. Therefore, jump-diffusion models are more suitable for pricing short maturity options in which the impact of the volatility clustering effect is less pronounced. In addition jump-diffusion models can provide a useful benchmark for more complicated models (for which one perhaps has to resort to simulation and other numerical procedures).

4 Equilibrium for general jump-diffusion models

4.1 Basic setting of equilibrium

Consider a typical rational expectations economy (Lucas, 1978) in which a representative investor tries to solve a utility maximization problem $\max_c \mathsf{E}[\int_0^\infty U(c(t), t)\,\mathrm{d}t]$, where $U(c(t), t)$ is the utility function of the consumption process $c(t)$. There is an exogenous endowment process, denoted by $\delta(t)$, available to the investor. Also given to the investor is an opportunity to invest in a security (with a finite liquidation date T_0, although T_0 can be very large) which pays no dividends. If $\delta(t)$ is Markovian, it can be shown (see, for example, Stokey and Lucas, 1989, pp. 484–485) that, under mild conditions, the rational expectations equilibrium price (also called the "shadow" price) of the security, $p(t)$, must satisfy the Euler equation

$$p(t) = \frac{\mathsf{E}(U_c(\delta(T), T)\,p(T) \mid \mathcal{F}_t)}{U_c(\delta(t), t)}, \quad \forall T \in [t, T_0], \tag{4}$$

where U_c is the partial derivative of U with respect to c. At this price $p(t)$, the investor will never change his/her current holdings to invest in (either long or short) the security, even though he/she is given the opportunity to do so. Instead, in equilibrium the investor find it optimal to just consume the exogenous endowment, i.e. $c(t) = \delta(t)$ for all $t \geqslant 0$.

In this section we shall derive explicitly the implications of the Euler equation (4) when the endowment process $\delta(t)$ follows a general jump-diffusion process under the physical measure P:

$$\frac{\mathrm{d}\delta(t)}{\delta(t-)} = \mu_1\,\mathrm{d}t + \sigma_1\,\mathrm{d}W_1(t) + \mathrm{d}\left[\sum_{i=1}^{N(t)} (\tilde{V}_i - 1)\right], \tag{5}$$

where the $\tilde{V}_i \geqslant 0$ are any independent identically distributed, nonnegative random variables. In addition, all three sources of randomness, the Poisson process $N(t)$, the standard Brownian motion $W_1(t)$, and the jump sizes $\tilde{V}$, are assumed to be independent.

Although it is intuitively clear that, generally speaking, the asset price $p(t)$ should follow a similar jump-diffusion process as that of the dividend process $\delta(t)$, a careful study of the connection between the two is needed. This is

because $p(t)$ and $\delta(t)$ may not have similar jump dynamics; see (15). Furthermore, deriving explicitly the change of parameters from $\delta(t)$ to $p(t)$ also provides some valuable information about the risk premiums embedded in jump diffusion models.

Naik and Lee (1990) consider the special case that $\tilde{V}_i$ has a lognormal distribution is investigated. In addition, Naik and Lee (1990) also require that the asset pays continuous dividends, and there is no outside endowment process; while here the asset pays no dividends and there is an outside endowment process. Consequently, the pricing formulae here are different even in the case of lognormal jumps.

For simplicity, we shall only consider the utility function of the special forms $U(c,t) = e^{-\theta t}\frac{c^\alpha}{\alpha}$ if $\alpha < 1$, and $U(c,t) = e^{-\theta t}\log(c)$ if $\alpha = 0$, where $\theta > 0$ (although most of the results below hold for more general utility functions), where θ is the discount rate in utility functions. Under these types of utility functions, the rational expectations equilibrium price of (4) becomes

$$p(t) = \frac{\mathsf{E}(e^{-\theta T}(\delta(T))^{\alpha-1}p(T) \mid \mathcal{F}_t)}{e^{-\theta t}(\delta(t))^{\alpha-1}}. \tag{6}$$

4.2 Choosing a risk-neutral measure

We shall assume that the discount rate θ should be large enough so that

$$\theta > -(1-\alpha)\mu_1 + \frac{1}{2}\sigma_1^2(1-\alpha)(2-\alpha) + \lambda\zeta_1^{(\alpha-1)},$$

where the notation $\zeta_1^{(a)}$ means

$$\zeta_1^{(a)} := \mathsf{E}\big[(\tilde{V})^a - 1\big].$$

This assumption guarantees that in equilibrium the term structure of interest rates is positive.

Suppose $\zeta_1^{(\alpha-1)} < \infty$. The following result in Kou (2002) justifies risk-neutral pricing by choosing a particular risk-neutral measure for option pricing:

(1) Letting $B(t,T)$ be the price of a zero coupon bond with maturity T, the yield $r := -\frac{1}{T-t}\log(B(t,T))$ is a constant independent of T,

$$r = \theta + (1-\alpha)\mu_1 - \frac{1}{2}\sigma_1^2(1-\alpha)(2-\alpha) - \lambda\zeta_1^{(\alpha-1)} > 0. \tag{7}$$

(2) Let $Z(t) := e^{rt}U_c(\delta(t), t) = e^{(r-\theta)t}(\delta(t))^{\alpha-1}$. Then $Z(t)$ is a martingale under P,

$$\frac{\mathrm{d}Z(t)}{Z(t-)} = -\lambda\zeta_1^{(\alpha-1)}\,\mathrm{d}t + \sigma_1(\alpha-1)\,\mathrm{d}W_1(t) + \mathrm{d}\left[\sum_{i=1}^{N(t)}\left(\tilde{V}_i^{\alpha-1}-1\right)\right]. \tag{8}$$

Using $Z(t)$, one can define a new probability measure P^*: $\frac{\mathrm{d}\mathsf{P}^*}{\mathrm{d}\mathsf{P}} := Z(t)/Z(0)$. The Euler equation (6) holds if and only if the asset price satisfies

$$S(t) = e^{-r(T-t)}\mathsf{E}^*\big(S(T) \mid \mathcal{F}_t\big), \quad \forall T \in [t, T_0]. \tag{9}$$

Furthermore, the rational expectations equilibrium price of a (possibly path-dependent) European option, with the payoff $\psi_S(T)$ at the maturity T, is given by

$$\psi_S(t) = e^{-r(T-t)}\mathsf{E}^*\big(\psi_S(T) \mid \mathcal{F}_t\big), \quad \forall t \in [0, T]. \tag{10}$$

4.3 The dynamic under the risk-neutral measure

Given the endowment process $\delta(t)$, it must be decided what stochastic processes are suitable for the asset price $S(t)$ to satisfy the equilibrium requirement (6) or (9). Now consider a special jump-diffusion form for $S(t)$,

$$\frac{\mathrm{d}S(t)}{S(t-)} = \mu\,\mathrm{d}t + \sigma\big\{\rho\,\mathrm{d}W_1(t) + \sqrt{1-\rho^2}\,\mathrm{d}W_2(t)\big\} + \mathrm{d}\left(\sum_{i=1}^{N(t)}(V_i-1)\right),$$
$$V_i = \tilde{V}_i^{\beta}, \tag{11}$$

where $W_2(t)$ is a Brownian motion independent of $W_1(t)$. In other words, the same Poisson process affects both the endowment $\delta(t)$ and the asset price $S(t)$, and the jump sizes are related through a power function, where the power $\beta \in (-\infty, \infty)$ is an arbitrary constant. The diffusion coefficients and the Brownian motion part of $\delta(t)$ and $S(t)$, though, are totally different. It remains to determine what constraints should be imposed on this model, so that the jump-diffusion model can be embedded in the rational expectations equilibrium requirement (6) or (9).

Suppose $\zeta_1^{(\alpha+\beta-1)} < \infty$ and $\zeta_1^{(\alpha-1)} < \infty$. It can be shown (Kou, 2002) that the model (11) satisfies the equilibrium requirement (9) if and only if

$$\begin{aligned}\mu &= r + \sigma_1\sigma\rho(1-\alpha) - \lambda\big(\zeta_1^{(\alpha+\beta-1)} - \zeta_1^{(\alpha-1)}\big)\\ &= \theta + (1-\alpha)\left\{\mu_1 - \frac{1}{2}\sigma_1^2(2-\alpha) + \sigma_1\sigma\rho\right\} - \lambda\zeta_1^{(\alpha+\beta-1)}.\end{aligned} \tag{12}$$

If (12) is satisfied, then under P^*

$$\frac{\mathrm{d}S(t)}{S(t-)} = r\,\mathrm{d}t - \lambda^*\mathsf{E}^*\big(\tilde{V}_i^{\beta} - 1\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W^*(t) + \mathrm{d}\left[\sum_{i=1}^{N(t)}\big(\tilde{V}_i^{\beta} - 1\big)\right]. \tag{13}$$

Here, under P^*, $W^*(t)$ is a new Brownian motion, $N(t)$ is a new Poisson process with jump rate $\lambda^* = \lambda\mathsf{E}(\tilde{V}_i^{\alpha-1}) = \lambda(\zeta_1^{(\alpha-1)}+1)$, and $\{\tilde{V}_i\}$ are independent identically distributed random variables with a new density under P^*:

$$f^*_{\tilde{V}}(x) = \frac{1}{\zeta_1^{(\alpha-1)}+1}x^{\alpha-1}f_{\tilde{V}}(x). \tag{14}$$

A natural question is under what conditions all three dynamics, $\delta(t)$ and $S(t)$ under P and $S(t)$ under P^*, have the same jump-diffusion form, which is very convenient for analytical calculation. Suppose the family $\mathcal{V}$ of distributions of the jump size $\tilde{V}$ for the endowment process $\delta(t)$ satisfies that, for any real number a,

$$\text{if } \tilde{V}^a \in \mathcal{V} \text{ then const} \cdot x^a f_{\tilde{V}}(x) \in \mathcal{V}, \tag{15}$$

where the normalizing constant, const, is $\{\zeta_1^{(a-1)}+1\}^{-1}$ (provided that $\zeta_1^{(a-1)} < \infty$). Then the jump sizes for the asset price $S(t)$ under P and the jump sizes for $S(t)$ under the rational expectations risk-neutral measure P^* all belong to the same family $\mathcal{V}$. The result follows immediately from (5), (11), and (14).

The condition (15) essentially requires that the jump size distribution belongs to the exponential family. It is satisfied if $\log(V)$ has a normal distribution or a double exponential distribution. However, the log power-type distributions, such as log t-distribution, do not satisfy (15).

5 Basic setting for option pricing

In the rest of the chapter, we shall focus on option pricing under jump-diffusion models. To do this we shall fix some notations. For a jump-diffusion process, the log-return $X(t) = \ln(S(t)/S(0))$ will be a process such that

$$X(t) = \tilde{\mu}t + \sigma W(t) + \sum_{i=1}^{N_t} Y_i, \quad X_0 \equiv 0. \tag{16}$$

Here $\{W_t;\, t \geqslant 0\}$ is a standard Brownian motion with $W_0 = 0$, $\{N_t;\, t \geqslant 0\}$ is a Poisson process with rate λ, constants $\tilde{\mu}$ and $\sigma > 0$ are the drift and volatility of the diffusion part, respectively, and the jump sizes $\{Y_1, Y_2, \ldots\}$ are independent identically distributed random variables. We also assume that the random processes $\{W_t;\, t \geqslant 0\}$, $\{N_t;\, t \geqslant 0\}$, and random variables $\{Y_1, Y_2, \ldots\}$ are independent representing $Y_i = \log(V_i)$.

The infinitesimal generator of the jump-diffusion process (16) is given by

$$\mathcal{L}u(x) = \frac{1}{2}\sigma^2 u''(x) + \tilde{\mu}u'(x) + \lambda \int_{-\infty}^{\infty} \big[u(x+y) - u(x)\big] f_Y(y)\,\mathrm{d}y, \tag{17}$$

for all twice continuously differentiable functions $u(x)$. In addition, suppose $\theta \in (-\eta_2, \eta_1)$. The moment generating function of $X(t)$ can be obtained as

$$\mathsf{E}\big[e^{\theta X(t)}\big] = \exp\big\{G(\theta)t\big\}, \tag{18}$$

where

$$G(x) := x\tilde{\mu} + \frac{1}{2}x^2\sigma^2 + \lambda\big(\mathsf{E}\big[e^{xY}\big] - 1\big).$$

In the case of Merton's normal jump-diffusion model, Y has a normal density

$$f_Y(y) \sim \frac{1}{\sigma'\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu')^2}{2\sigma'}\right\},$$

where μ' and σ' are the mean and standard deviation for Y. Thus,

$$G(x) = x\tilde{\mu} + \frac{1}{2}x^2\sigma^2 + \lambda\left\{\mu' x + \frac{(\sigma')^2 x^2}{2} - 1\right\}.$$

In the case of double exponential jump-diffusion model

$$f_Y(y) \sim p \cdot \eta_1 e^{-\eta_1 y}\mathbf{1}_{\{y \geqslant 0\}} + q \cdot \eta_2 e^{\eta_2 y}\mathbf{1}_{\{y<0\}}, \quad \eta_1 > 1,\ \eta_2 > 0,$$

and the function $G(x)$ is

$$G(x) = x\tilde{\mu} + \frac{1}{2}x^2\sigma^2 + \lambda\left(\frac{p\eta_1}{\eta_1 - x} + \frac{q\eta_2}{\eta_2 + x} - 1\right). \tag{19}$$

Kou and Wang (2003) show that for $\alpha > 0$ in the case of double exponential jump-diffusion model the equation $G(x) = \alpha$ has exactly four roots $\beta_{1,\alpha}, \beta_{2,\alpha}, -\beta_{3,\alpha}, -\beta_{4,\alpha}$, where

$$0 < \beta_{1,\alpha} < \eta_1 < \beta_{2,\alpha} < \infty, \qquad 0 < \beta_{3,\alpha} < \eta_2 < \beta_{4,\alpha} < \infty. \tag{20}$$

The analytical formulae for the four roots of the equation $G(x) = \alpha$, which is essentially a quartic equation, are given in Kou et al. (2005). The explicit formulae of β's are useful for the Euler algorithm in Laplace inversion.

Under the risk-neutral probability P^* in (13), we have

$$\tilde{\mu} = r - \frac{1}{2}\sigma^2 - \lambda\zeta,$$

where $\zeta := \mathsf{E}^*[e^Y] - 1$. Similarly, if the underlying asset pays continuous dividend at the rate δ, then under P^*

$$\tilde{\mu} = r - \delta - \frac{1}{2}\sigma^2 - \lambda\zeta.$$

In the Merton's model

$$\zeta = \mathsf{E}^*\big[e^Y\big] - 1 = \mu' + \frac{(\sigma')^2}{2} - 1,$$

and in the double exponential jump-diffusion model

$$\zeta = \mathsf{E}^*\big[e^Y\big] - 1 = p\eta_1/(\eta_1 - 1) + q\eta_2/(\eta_2 + 1) - 1.$$

6 Pricing call and put option via Laplace transforms

Laplace transforms have been widely used in valuing financial derivatives. For example, Carr and Madan (1999) propose Fourier transforms with respect to log-strike prices; Geman and Yor (1993), Fu et al. (1999) use Laplace transforms to price Asian options in the Black–Scholes setting; Laplace transforms for double-barrier and lookback options under the CEV model are given in Davydov and Linetsky (2001); Petrella and Kou (2004) use a recursion and Laplace transforms to price discretely monitored barrier and lookback options. For a survey of Laplace transforms in option pricing, see Craddock et al. (2000).

Kou et al. (2005) adapted the method in Carr and Madan (1999), which is based on a change of the order of integration, to price European call and put option via Laplace transforms. In principle, the Laplace transforms for the prices of European call and European put options can also be obtained by using standard results from Fourier transforms for general Lévy processes (see Cont and Tankov, 2004, pp. 361–362).

To fix the notation, the price of a European call with maturity T and strike K, is given by

$$\begin{aligned} C_T(k) &= e^{-rT}\mathsf{E}^*\big[\big(S(T) - K\big)^+\big] \\ &= e^{-rT}\mathsf{E}^*\big[\big(S(0)e^{X(T)} - e^{-k}\big)^+\big], \end{aligned} \tag{21}$$

where $k = -\log(K)$, and the price of a European put is

$$P_T(k') = e^{-rT}\mathsf{E}^*\big[\big(K - S(T)\big)^+\big] = e^{-rT}\mathsf{E}^*\big[\big(e^{k'} - S(0)e^{X(T)}\big)^+\big],$$

where $k' = \log(K)$. The Laplace transform with respect to k of $C_T(k)$ in (21) is given by

$$\begin{aligned} \hat{f}_C(\xi) &:= \int_{-\infty}^{\infty} e^{-\xi k} C_T(k)\,\mathrm{d}k \\ &= e^{-rT}\frac{S(0)^{\xi+1}}{\xi(\xi+1)}\exp\big(G(\xi+1)T\big), \quad \xi > 0, \end{aligned} \tag{22}$$

and the Laplace transform with respect to k' for the put option $P_T(k')$ is

$$\hat{f}_P(\xi) := \int_{-\infty}^{\infty} e^{-\xi k'} P_T(k')\,\mathrm{d}k'$$
$$= e^{-rT}\frac{S(0)^{-(\xi-1)}}{\xi(\xi-1)}\exp\bigl(G\bigl(-(\xi-1)T\bigr)\bigr), \quad \xi > 1, \tag{23}$$

in the notation of (18).

To show this, note that by (21) the Laplace transform for the call options is

$$\hat{f}_C(\xi) = e^{-rT}\int_{-\infty}^{\infty} e^{-\xi k}\mathsf{E}^*\bigl[\bigl(S(0)e^{X(T)} - e^{-k}\bigr)^+\bigr]\,\mathrm{d}k.$$

Applying the Fubini theorem yields for every $\xi > 0$,

$$\hat{f}_C(\xi) = e^{-rT}\mathsf{E}^*\left[\int_{-\infty}^{\infty} e^{-\xi k}\bigl(S(0)e^{X(T)} - e^{-k}\bigr)^+\,\mathrm{d}k\right]$$
$$= e^{-rT}\mathsf{E}^*\left[\int_{-X(T)-\log S(0)}^{\infty} e^{-\xi k}\bigl(S(0)e^{X(T)} - e^{-k}\bigr)\,\mathrm{d}k\right]$$
$$= e^{-rT}\mathsf{E}^*\left[S(0)e^{X(T)}e^{\xi(X(T)+\log S(0))}\frac{1}{\xi}\right.$$
$$\left. - e^{(\xi+1)(X(T)+\log S(0))}\frac{1}{\xi+1}\right]$$
$$= e^{-rT}\frac{S(0)^{\xi+1}}{\xi(\xi+1)}\mathsf{E}^*\bigl[e^{(\xi+1)X(T)}\bigr],$$

from which (22) follows readily from (18). The proof of (23) is similar.

The Laplace transforms can be inverted numerically in the complex plane, using the two-sided extension of the Euler algorithm as described and implemented in Petrella (2004). To check the accuracy of the inversion, Kou et al. (2005) compare the inversion results with the prices of call and put options under the double exponential jump-diffusion model obtained by using the closed-form formulae using *Hh* function as in Kou (2002). They found that the results from the Laplace inversion method agree to the fifth decimal with the analytical solutions for European call and put options. Because of the difficulty in precise calculation of the normal distribution function and the $Hh(x)$ function for very positive and negative x, it is possible that for very large values of the return variance $\sigma^2 T$ and for very high jump rate λ (though perhaps not within the typical parameter ranges seen in finance applications) the closed-form formulae may not give accurate results. In such cases, the inversion method still performs remarkably well.

It is also possible to compute the sensitivities of the option, such as Delta, Gamma, Theta, Vega, etc., by inverting the derivatives of the option's Laplace transform in (22). For example, the delta is given by

$$\begin{aligned}\Delta\big(C_T(k)\big) &= \frac{\partial}{\partial S(0)} C_T(k) \\ &= \mathcal{L}_{\xi}^{-1}\left(e^{-rT}\frac{S(0)^{\xi}}{\xi}\exp\big(G(\xi+1)T\big)\right)\Big|_{k=-\log K},\end{aligned}$$

where $\mathcal{L}_{\xi}^{-1}$ means the Laplace inversion with respect to ξ.

7 First passage times

To price perpetual American options, barrier options, and lookback options for general jump-diffusion processes, it is crucial to study the first passage time of a jump-diffusion process $X(t)$ to a flat boundary:

$$\tau_b := \inf\big\{t \geqslant 0;\ X(t) \geqslant b\big\}, \quad b > 0,$$

where $X(\tau_b) := \limsup_{t\to\infty} X(t)$, on the set $\{\tau_b = \infty\}$.

7.1 The overshoot problem

Without the jump part, the process $X(t)$ simply becomes a Brownian motion with drift $\tilde{\mu}$. The distributions of the first passage times can be obtained either by a combination of a change of measure (Girsanov theorem) and the reflection principle, or by calculating the Laplace transforms via some appropriate martingales and the optional sampling theorem. Details of both methods can be found in many classical textbooks on stochastic analysis, e.g. Karlin and Taylor (1975), Karatzas and Shreve (1991). With the jump part, however, it is very difficult to study the first passage times for general jump-diffusion processes. When a jump-diffusion process crosses boundary level b, sometimes it hits the boundary exactly and sometimes it incurs an "overshoot," $X(\tau_b) - b$, over the boundary. See Fig. 6 for an illustration.

The overshoot presents several problems for option pricing. First, one needs to get the exact distribution of the overshoot. It is well known from stochastic renewal theory that this is in general difficult unless the jump size Y has an exponential-type distribution, thanks to the special memoryless property of the exponential distribution. Second, one needs to know the dependent structure between the overshoot and the first passage time. The two random variables are conditionally independent, given that the overshoot is bigger than 0, if the jump size Y has an exponential-type distribution, thanks to the memoryless property. This conditionally independent structure seems to be very special to the exponential-type distribution, and does not hold for other distributions,

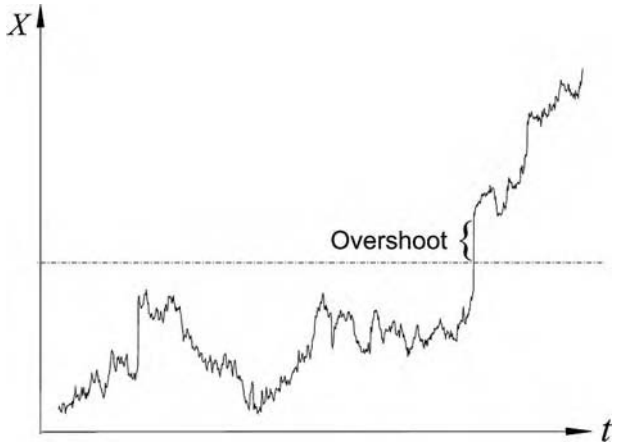


Fig. 6. A simulated sample path with the overshoot problem.

such as the normal distribution. Third, if one wants to use the reflection principle to study the first passage times, the dependent structure between the overshoot and the terminal value X_t is also needed. This is not known to the best of our knowledge, even for the double exponential jump-diffusion process.

Consequently, we can derive closed form solutions for the Laplace transforms of the first passage times for the double exponential jump-diffusion process, yet cannot give more explicit calculations beyond that, as the correlation between $X(t)$ and $X(\tau_b) - b$ is not available. However, for other jump-diffusion processes, even analytical forms of the Laplace transforms seem to be quite difficult. See Asmussen et al. (2004), Boyarchenko and Levendorskiĭ (2002), and Kyprianou and Pistorius (2003) for some representations (though not explicit calculations) based on the Wiener–Hopf factorization related to the overshoot problems for general Lévy processes; and see also Avram et al. (2004) and Rogers (2000) for first passage times with one-sided jumps.

7.2 Conditional independence

The following result shows that the memoryless property of the random walk of exponential random variables leads to the conditional memoryless property of the jump-diffusion process. For any $x > 0$,

$$\mathsf{P}\big(\tau_b \leqslant t, X(\tau_b) - b \geqslant x\big) = e^{-\eta_1 x}\mathsf{P}\big(\tau_b \leqslant t, X(\tau_b) - b > 0\big), \tag{24}$$

$$\mathsf{P}\big(X(\tau_b) - b \geqslant x \mid X(\tau_b) - b > 0\big) = e^{-\eta_1 x}. \tag{25}$$

Furthermore, conditional on $X_{\tau_b} - b > 0$, the stopping time τ_b and the overshoot $X_{\tau_b} - b$ are independent; more precisely, for any $x > 0$,

$$\mathsf{P}\big(\tau_b \leqslant t, X(\tau_b) - b \geqslant x \mid X(\tau_b) - b > 0\big)$$

$$= \mathsf{P}\big(\tau_b \leqslant t \mid X(\tau_b) - b > 0\big)\mathsf{P}\big(X\big(\tau_b\big) - b \geqslant x \mid X(\tau_b) - b > 0\big). \tag{26}$$

It should be pointed out that τ_b and the overshoot $X(\tau_b) - b$ are dependent even in the case of double exponential jump diffusion, although they are conditionally independent.

7.3 Distribution of the first passage times

For any $\alpha \in (0, \infty)$, let $\beta_{1,\alpha}$ and $\beta_{2,\alpha}$ be the only two positive roots for the equation $\alpha = G(\beta)$, where $0 < \beta_{1,\alpha} < \eta_1 < \beta_{2,\alpha} < \infty$. Then Kou and Wang (2003) give the following results regarding the Laplace transform of τ_b

$$\begin{aligned}
\mathsf{E}\big[e^{-\alpha\tau_b}\big] &= \frac{\eta_1 - \beta_{1,\alpha}}{\eta_1} \cdot \frac{\beta_{2,\alpha}}{\beta_{2,\alpha} - \beta_{1,\alpha}} e^{-b\beta_{1,\alpha}} + \frac{\beta_{2,\alpha} - \eta_1}{\eta_1} \\
&\quad \cdot \frac{\beta_{1,\alpha}}{\beta_{2,\alpha} - \beta_{1,\alpha}} e^{-b\beta_{2,\alpha}}, \\
\mathsf{E}^*\big[e^{-\alpha\tau_b}\mathbf{1}_{\{X(\tau_b)>b\}}\big] &= \frac{(\eta_1 - \beta_{1,\alpha})(\beta_{2,\alpha} - \eta_1)}{\eta_1(\beta_{2,\alpha} - \beta_{1,\alpha})}\big[e^{-b\beta_{1,\alpha}} - e^{-b\beta_{2,\alpha}}\big], \\
\mathsf{E}^*\big[e^{-\alpha\tau_b}\mathbf{1}_{\{X(\tau_b)=b\}}\big] &= \frac{\eta_1 - \beta_{1,\alpha}}{\beta_{2,\alpha} - \beta_{1,\alpha}} e^{-b\beta_{1,\alpha}} + \frac{\beta_{2,\alpha} - \eta_1}{\beta_{2,\alpha} - \beta_{1,\alpha}} e^{-b\beta_{2,\alpha}}.
\end{aligned} \tag{27}$$

The results for the down-crossing barrier problem, i.e. $b < 0$, will involve the other two roots, $\beta_{3,\alpha}$ and $\beta_{4,\alpha}$.

For simplicity, we will focus on (27). It is easy to give a heuristic argument for (27). Let $u(x) = \mathsf{E}^x[e^{-\alpha\tau_b}]$, $b > 0$, we expect from a heuristic application of the Feymann–Kac formula that u satisfies the integro-differential equation

$$-\alpha u(x) + \mathcal{L}u(x) = 0, \quad \forall x < b, \tag{28}$$

and $u(x) = 1$ if $x \geqslant b$. This equation can be explicitly solved at least heuristically. Indeed, consider a solution taking form

$$u(x) = \begin{cases} 1, & x \geqslant b, \\ A_1 e^{-\beta_1(b-x)} + B_1 e^{-\beta_2(b-x)}, & x < b, \end{cases} \tag{29}$$

where constants A_1 and B_1 are yet to be determined. Plug in to obtain, after some algebra, that $(-\alpha u + \mathcal{L}u)(x)$ for all $x < b$ is equal to

$$\begin{aligned}
&A_1 e^{-(b-x)\beta_1} f(\beta_1) + B_1 e^{-(b-x)\beta_2} f(\beta_2) \\
&\quad -\lambda p e^{-\eta_1(b-x)}\left(\frac{A_2\eta_1}{\eta_1 - \beta_1} + \frac{B_2\eta_1}{\eta_1 - \beta_2} - e^{-\eta_1 y}\right),
\end{aligned} \tag{30}$$

where $f(\beta) = G(\beta) - \alpha$.

To set $(-\alpha u + \mathcal{L}u)(x) = 0$ for all $x < b$, we can first have $f(\beta_1) = f(\beta_2) = 0$, which means that we shall choose β_1 and β_2 to be two roots of $G(\beta) = \alpha$, although it is not clear which two roots among the four are needed.

Afterward it is enough to set the third term in (30) to be zero by choosing A_1 and B_1 so that

$$A_1 \frac{\eta_1}{\eta_1 - \beta_1} + B_1 \frac{\eta_1}{\eta_1 - \beta_2} = e^{-\eta_1 y}.$$

Furthermore, the continuity of u at $x = b$ implies that

$$A_1 + B_1 = 1.$$

Solve the equations to obtain A_1 and B_1 ($A_1 = 1 - B_1$), which are exactly the coefficients in (27). However, the above heuristic argument has several difficulties.

7.4 Difficulties

7.4.1 Nonsmoothness

Because the function $u(x)$ in (29) is continuous, but not C^1 at $x = b$, we cannot apply the Itô formula and the Feymann–Kac formula directly to the process $e^{-\alpha t}u(X_t)$; $t \geqslant 0$. Furthermore, even if we can use the Feymann–Kac formula, it is not clear whether the solution to the integro-differential equation (28) is well defined and unique. Therefore, Kou and Wang (2003) have to use some approximation of $u(x)$ so that Itô formula can be used, and then they used a martingale method to solve the integro-differential equation (28) directly. In addition, the martingale method also helps to identify which two roots are needed in the formulae. Note that a heuristic argument based on the Feymann–Kac formula for double barrier options (with both upper and lower barriers) is given in Sepp (2004) by extending (28) and (29), and ignoring the nonsmoothness issue.

7.4.2 Explicit calculation

It should be mentioned that the special form of double exponential density function enables us to explicitly solve the integro-differential equation (28) associated with the Laplace transforms using martingale methods, thanks to the explicit calculation in (30). This is made possible as the exponential function has some good properties such as the product of two exponential functions is still an exponential function, and the integral of an exponential function is again an exponential function. For general jump-diffusion processes, however, such explicit solution will be very difficult to obtain.

7.4.3 Nonuniqueness in renewal integral equations

We have used martingale and differential equations to derive closed form solutions of the Laplace transforms for the first-passage-time probabilities. Another possible and popular approach to solving the problems is to set up some integral equations by using renewal arguments. For simplicity, we shall only consider the case of overall drift being nonnegative, i.e. $\bar{u} \geqslant 0$, in which $\tau_b < \infty$ almost surely. For any $x > 0$, define $P(x)$ as the probability that

no overshoot occurs for the first passage time τ_x with $X(0) \equiv 0$, that is $P(x) = \mathsf{P}(X(\tau_x) = x)$. It is easy to see that $P(x)$ satisfies the following renewal-type integral equation:

$$P(x+y) = P(y)P(x) + \big(1 - P(x)\big) \int_0^y P(y-z) \cdot \eta_1 e^{-\eta_1 z}\, \mathrm{d}z.$$

However, the solution to this renewal equation is not unique. Indeed, for every $\xi \geqslant 0$, the function

$$P_\xi(x) = \frac{\eta_1}{\eta_1 + \xi} + \frac{\xi}{\eta_1 + \xi} e^{-(\eta_1+\xi)x}$$

satisfies the integral equation with the boundary condition $P_\xi(0) = 1$.

This shows that, in the presence of two-sided jumps, the renewal-type integral equations may not have unique solutions, mainly because of the difficulty of determining enough boundary conditions based on renewal arguments alone. It is easy to see that $\xi = -P'_\xi(0)$. Indeed, it is possible to use the infinitesimal generator and martingale methods to determine ξ. The point here is, however, that the renewal-type integral equations cannot do the job by themselves.

8 Barrier and lookback options

Barrier and lookback options are among the most popular path-dependent derivatives traded in exchanges and over-the-counter markets worldwide. The payoffs of these options depend on the extrema of the underlying asset. For a complete description of these and other related contracts we refer the reader to Hull (2005). In the standard Black–Scholes setting, closed-form solutions for barrier and lookback options have been derived by Merton (1973) and Gatto et al. (1979).

8.1 Pricing barrier options

We will focus on the pricing of an up-and-in call option (UIC, from now on); other types of barrier options can be priced similarly and using the symmetries described in the Appendix of Petrella and Kou (2004) and Haug (1999). The price of an UIC is given by

$$UIC(k,T) = \mathsf{E}^*\big[e^{-rT}\big(S(T) - e^{-k}\big)^+ \mathbf{1}_{\{\tau_b < T\}}\big], \tag{31}$$

where $H > S(0)$ is the barrier level, $k = -\log(K)$ the transformed strike and $b = \log(H/S(0))$. Using a change of numéraire argument, Kou and Wang (2004) show that under another probability, defined as $\tilde{\mathsf{P}}$, $X(T)$ still has a dou-

ble exponential distribution with drift $r - \delta + \frac{1}{2}\sigma^2 - \lambda\zeta$ and jump parameters

$$\tilde{\lambda} = \lambda(\zeta + 1), \qquad \tilde{p} = \frac{p\eta_1}{(\zeta + 1)(\eta_1 - 1)},$$
$$\tilde{\eta}_1 = \eta_1 - 1, \qquad \tilde{\eta}_2 = \eta_2 + 1.$$

The moment generating function of $X(t)$ under the alternative probability measure $\tilde{\mathsf{P}}$ is given by $\tilde{\mathsf{E}}[e^{\theta X(t)}] = \exp(\tilde{G}(\theta)t)$, with

$$\tilde{G}(x) := x\left(r - \delta + \frac{1}{2}\sigma^2 - \tilde{\lambda}\tilde{\zeta}\right) + \frac{1}{2}x^2\sigma^2 + \tilde{\lambda}\left(\frac{\tilde{p}\tilde{\eta}_1}{\tilde{\eta}_1 - x} + \frac{\tilde{q}\tilde{\eta}_2}{\tilde{\eta}_2 + x} - 1\right).$$

Kou and Wang (2004) further show that

$$UIC(k, T) = S(0)\tilde{\Psi}_{UI}(k, T) - Ke^{-rT}\Psi_{UI}(k, T), \tag{32}$$

where

$$\Psi_{UI}(k, T) = \mathsf{P}^*\big(S(T) \geqslant e^{-k}, M_{0,T} > H\big),$$
$$\tilde{\Psi}_{UI}(k, T) = \tilde{\mathsf{P}}\big(S(T) \geqslant e^{-k}, M_{0,T} > H\big), \tag{33}$$

and show how to price an UIC option by inverting the one-dimensional Laplace transforms for the joint distributions in (32) as in Kou and Wang (2003).

Kou et al. (2005) present an alternative approach that relies on a two-dimensional Laplace transform for both the option price in (31) and the probabilities in (32). The formulae after doing two-dimensional transforms become much simpler than the one-dimensional formulae in Kou and Wang (2003), which involve many special functions.

In particular Kou et al. (2005) show that for ξ and α such that $0 < \xi < \eta_1 - 1$ and $\alpha > \max(G(\xi + 1) - r, 0)$ (such a choice of ξ and α is possible for all small enough ξ as $G(1) - r = -\delta < 0$), the Laplace transform with respect to k and T of $UIC(k, T)$ is given by

$$\hat{f}_{UIC}(\xi, \alpha) = \int_0^\infty \int_{-\infty}^\infty e^{-\xi k - \alpha T} UIC(k, T)\,\mathrm{d}k\,\mathrm{d}T$$
$$= \frac{H^{\xi+1}}{\xi(\xi + 1)} \frac{1}{r + \alpha - G(\xi + 1)} \times \left(A(r + \alpha)\frac{\eta_1}{\eta_1 - (\xi + 1)} + B(r + \alpha)\right), \tag{34}$$

where

$$A(h) := \mathsf{E}^*\big[e^{-h\tau_b}\mathbf{1}_{\{X(\tau_b)>b\}}\big] = \frac{(\eta_1-\beta_{1,h})(\beta_{2,h}-\eta_1)}{\eta_1(\beta_{2,h}-\beta_{1,h})}\big[e^{-b\beta_{1,h}}-e^{-b\beta_{2,h}}\big], \tag{35}$$

$$B(h) := \mathsf{E}^*\big[e^{-h\tau_b}\mathbf{1}_{\{X(\tau_b)=b\}}\big] = \frac{\eta_1-\beta_{1,h}}{\beta_{2,h}-\beta_{1,h}}e^{-b\beta_{1,h}} + \frac{\beta_{2,h}-\eta_1}{\beta_{2,h}-\beta_{1,h}}e^{-b\beta_{2,h}}, \tag{36}$$

with $b = \log(H/S(0))$. If $0 < \xi < \eta_1$ and $\alpha > \max(G(\xi), 0)$ (again this choice of ξ and α is possible for all ξ small enough as $G(0) = 0$), then the Laplace transform with respect to k and T of $\Psi_{UI}(k, T)$ in (33) is

$$\hat{f}_{\Psi_{UI}}(\xi,\alpha) = \int_{-\infty}^{\infty}\left(\int_0^{\infty} e^{-\xi k-\alpha T}\Psi_{UI}(k,T)\,\mathrm{d}T\right)\mathrm{d}k = \frac{H^{\xi}}{\xi}\frac{1}{\alpha-G(\xi)}\left(A(\alpha)\frac{\eta_1}{\eta_1-\xi}+B(\alpha)\right). \tag{37}$$

The Laplace transforms with respect to k and T of $\tilde{\Psi}_{UI}(k, T)$ is given similarly with $\tilde{G}$ replacing G and the functions $\tilde{A}$ and $\tilde{B}$ defined similarly.

Kou et al. (2005) price up-and-in calls using the two-dimensional Laplace transform (using the two-dimensional Euler algorithm developed by Choudhury et al., 1994 and Petrella, 2004) and compare the results with the one-dimensional transform in Kou and Wang (2003) (based on the Gaver–Stehfest algorithm). The two-dimensional Laplace inversion matches to the fourth digit the ones obtained by the one-dimensional Gaver–Stehfest algorithm, and are all within the 95% confidence interval obtained via Monte Carlo simulation.

The two-dimensional Laplace inversion algorithms have three advantages compared to the one-dimensional algorithm: (1) The formulae for the two-dimensional transforms Euler are much easier to compute, simplifying the implementation of the methods. (2) Although we are inverting two-dimensional transforms, the Laplace transform methods are significantly faster, mainly because of the simplicity in the Laplace transform formulae. (3) High-precision calculation (with about 80 digit accuracy) as required by the Gaver–Stehfest inversion is no longer needed in the Euler inversion, which is made possible mainly because of the simplicity of the two-dimensional inversion formulae as no special functions are involved and all the roots of $G(x)$ are given in analytical forms.

8.2 Pricing lookback options via Euler inversion

For simplicity, we shall focus on a standard lookback put option, while the derivation for a standard lookback call is similar. The price of a standard look-

back put is given by

$$LP(T) = \mathsf{E}^*\Big[e^{-rT}\Big\{\max\Big\{M, \max_{0\leqslant t\leqslant T} S(t)\Big\} - S(t)\Big\}\Big]$$
$$= \mathsf{E}^*\Big[e^{-rT}\max\Big\{M, \max_{0\leqslant t\leqslant T} S(t)\Big\}\Big] - S(0),$$

where $M \geqslant S(0)$ is the prefixed maximum at time 0. For any $\xi > 0$, the Laplace transform of the lookback put with respect to the time to maturity T is given by (see Kou and Wang, 2004)

$$\int_0^\infty e^{-\alpha T} LP(T)\,\mathrm{d}T = \frac{S(0)A_\alpha}{C_\alpha}\left(\frac{S(0)}{M}\right)^{\beta_{1,\alpha+r}-1} + \frac{S(0)B_\alpha}{C_\alpha}\left(\frac{S(0)}{M}\right)^{\beta_{2,\alpha+r}-1} + \frac{M}{\alpha+r} - \frac{S(0)}{\alpha}, \quad (38)$$

where

$$A_\alpha = \frac{(\eta_1 - \beta_{1,\alpha+r})\beta_{2,\alpha+r}}{\beta_{1,\alpha+r} - 1}, \qquad B_\alpha = \frac{(\beta_{2,\alpha+r} - \eta_1)\beta_{1,\alpha+r}}{\beta_{2,\alpha+r} - 1},$$
$$C_\alpha = (\alpha + r)\eta_1(\beta_{2,\alpha+r} - \beta_{1,\alpha+r}),$$

and $\beta_{1,\alpha+r}$, $\beta_{2,\alpha+r}$ are the two positive roots of the equation $G(x) = \alpha + r$, as in (20).

The transform in (38) can be inverted in the complex domain by using the one-dimensional Euler inversion (EUL) algorithm developed by Abate and Whitt (1992), rather than in the real domain by the Gaver–Stehfest (GS) algorithm as in Kou and Wang (2004). The main reason for this is that the EUL inversion (which is carried out in the complex-domain) does not require the high numerical precision of the GS algorithm: a precision of 12 digits will suffice for the EUL, compared with the 80 digits accuracy required by the GS. The EUL algorithm is made possible partly due to an explicit formula for the roots of $G(x)$ given. Kou et al. (2005) show that the difference between the EUL and GS results are small. Ultimately, the EUL implementation is preferable, since it is simple to implement, and it converges fast without requiring high numerical precision as in the GS.

9 Analytical approximations for American options

Most of call and put options traded in the exchanges in both US and Europe are American-type options. Therefore, it is of great interest to calculate the prices of American options accurately and quickly. The price of a finite-horizon American option is the solution of a finite horizon free boundary problem. Even within the classical geometric Brownian motion model, except

in the case of the American call option with no dividend, there is no analytical solution available. To price American options under general jump-diffusion models, one may consider numerically solving the free boundary problems via lattice or differential equation methods; see, e.g., Amin (1993), d'Halluin et al. (2003), Feng and Linetsky (2005), Feng et al. (2004), and the book by Cont and Tankov (2004).

9.1 Quadratic approximation

Extending the Barone-Adesi and Whaley (1987) approximation for the classical geometric Brownian motion model, Kou and Wang (2004) considered an alternative approach that takes into consideration of the special structure of the double exponential jump-diffusions. One motivation for such an extension is its simplicity, as it yields an analytic approximation that only involves the price of a European option. The numerical results in Kou and Wang (2004) suggest that the approximation error is typically less than 2%, which is less than the typical bid–ask spread (about 5% to 10%) for American options in exchanges. Therefore, the approximation can serve as an easy way to get a quick estimate that is perhaps accurate enough for many practical situations. The extension of Barone-Adesi and Whaley's quadratic approximation method works nicely for double exponential jump-diffusion models mainly because explicit solutions are available to a class of relevant integro-differential free boundary problems.

To simplify notation, we shall focus only on the finite horizon American put option without dividends, as the methodology is also valid for the finite horizon American call option with dividends. Also, we shall only consider finite-time horizon American put options. Related American calls can be priced by exploiting the symmetric relationship in Schröder (1999).

The analytic approximation involves two quantities, $\mathrm{EuP}(v,t)$ which denotes the price of the European put option with initial stock price v and maturity t, and $\mathsf{P}^v[S(t)\leqslant K]$ which is the probability that the stock price at t is below K with initial stock price v. Both $\mathrm{EuP}(v,t)$ and $\mathsf{P}^v[S(t)\leqslant K]$ can be computed fast by using either the closed form solutions in Kou (2002) or the Laplace transforms in Kou et al. (2005).

We need some notations. Let $z=1-e^{-rt}$, $\beta_3\equiv\beta_{3,\frac{r}{z}}$, $\beta_4\equiv\beta_{4,\frac{r}{z}}$, $C_\beta=\beta_3\beta_4(1+\eta_2)$, $D_\beta=\eta_2(1+\beta_3)(1+\beta_4)$, in the notation of Eq. (20). Define $v_0\equiv v_0(t)\in(0,K)$ as the unique solution to the equation

$$\begin{aligned}&C_\beta K-D_\beta\big[v_0+\mathrm{EuP}(v_0,t)\big]\\&\quad=(C_\beta-D_\beta)Ke^{-rt}\cdot\mathsf{P}^{v_0}\big[S(t)\leqslant K\big].\end{aligned}\tag{39}$$

Note that the left-hand side of (39) is a strictly decreasing function of v_0 (because $v_0+\mathrm{EuP}(v_0,t)=e^{-rt}\mathsf{E}^*[\max(S(t),K)\mid S(0)=v_0]$), and the right hand side of (39) is a strictly increasing function of v_0 [because $C_\beta-D_\beta=\beta_3\beta_4-\eta_2(1+\beta_3+\beta_4)<0$]. Therefore, v_0 can be obtained easily by using, for example, the bisection method.

Approximation. The price of a finite horizon American put option with maturity t and strike K can be approximated by $\psi(S(0), t)$, where the value function ψ is given by

$$\psi(v,t) = \begin{cases} \mathrm{EuP}(v,t) + Av^{-\beta_3} + Bv^{-\beta_4}, & \text{if } v \geqslant v_0, \\ K - v, & \text{if } v \leqslant v_0, \end{cases} \tag{40}$$

with v_0 being the unique root of Eq. (39) and the two constants A and B given by

$$A = \frac{v_0^{\beta_3}}{\beta_4 - \beta_3}\big\{\beta_4 K - (1+\beta_4)\big[v_0 + \mathrm{EuP}(v_0,t)\big] + Ke^{-rt}\mathsf{P}^{v_0}\big[S(t) \leqslant K\big]\big\} > 0, \tag{41}$$

$$B = \frac{v_0^{\beta_4}}{\beta_3 - \beta_4}\big\{\beta_3 K - (1+\beta_3)\big[v_0 + \mathrm{EuP}(v_0,t)\big] + Ke^{-rt}\mathsf{P}^{v_0}\big[S(t) \leqslant K\big]\big\} > 0. \tag{42}$$

In the numerical examples showed by Kou and Wang (2004) the maximum relative error is only about 2.6%, while in most cases the relative errors are below 1%. The approximation runs very fast, taking only about 0.04 s to compute one price on a Pentium 1500 PC, irrespective to the parameter ranges; while the lattice method in Amin (1993) works much slower, taking about over one hour to compute one price.

9.2 Piecewise exponential approximation

A more accurate approximation can be obtained by extending the piecewise exponential approximation in Ju (1998). Extending previous work by Carr et al. (1992), Gukhal (2001) and Pham (1997) show that under jump-diffusion models the value at time t of an American put option with maturity $T > t$ on an asset with value S_t at time t ($P_A(S_t, t, T)$ from now on) is given by

$$\begin{aligned} P_A(S_t,t,T) = P_E(S_t,t,T) &+ \int_t^T e^{-r(s-t)} rK\mathsf{E}^*[\mathbf{1}_{\{S_s \leqslant S_s^*\}} \mid S_t]\,\mathrm{d}s \\ &- \delta\int_t^T e^{-r(s-t)}\mathsf{E}^*[S_s\mathbf{1}_{\{S_s \leqslant S_s^*\}} \mid S_t]\,\mathrm{d}s \\ &- \lambda\int_t^T e^{-r(s-t)}\mathsf{E}^*\big[\big\{P_A(VS_{s^-},s,T) - (K - VS_{s^-})\big\} \\ &\times \mathbf{1}_{\{S_{s^-} \leqslant S_{s^-}^*\}}\mathbf{1}_{\{VS_{s^-} > S_{s^-}^*\}} \mid S_t\big]\,\mathrm{d}s, \end{aligned} \tag{43}$$

where $P_E(S_t, t, T)$ is the value of the corresponding European put option, $\log(V) = Y$ with an independent double exponential distribution, and S_s^* is the early exercise boundary at time s, such that if the stocks price S_s goes below S_s^* at time s, then it is optimal to exercise immediately. Gukhal (2001) provides an interpretation of the four terms in (43): the value of an American put is given by the corresponding European put option value $P_E(X, t, T)$ to which we add the present value of interest accrued on the strike price in the exercise region (*IA*, from now), subtract the present value of dividends lost in the exercise region (*DL*, from now on), and subtract the last term in (43), to be denoted by $RCJ(t, T)$, which represents the rebalancing costs due to jumps from the early exercise region to the continuation region and is absent in the case of pure-diffusion.

The term $RCJ(t, T)$ takes into account of the possibility of an upward jump that will move the asset price from the early exercise to the continuation region. Consequently, this term diminishes when the upward jump rate λp is small. Furthermore, intuitively this term should also be very small whenever a jump from the early exercise to the continuation region only causes minimal changes in the American option value, which in particular requires that the overshoot over the exercise boundary is not too large. This happens if the overshoot jump size has small mean, which in the double exponential case is $1/\eta_1$; in most practical cases $\eta_1 > 10$. In other words, the term $RCJ(t, T)$ should be negligible for either small λp or large η_1. Indeed, Kou et al. (2005) show that for $T > t$, under the double exponential jump-diffusion model,

$$RCJ(t, T) \leqslant \lambda p \frac{\eta_1}{\eta_1 - 1} K \cdot U(t, T),$$

$$U(t, T) = \int_t^T \mathsf{E}^* \left[\left(\frac{S_{s^-}^*}{S_{s^-}} \right)^{-(\eta_1 - 1)} \mathbf{1}_{\{S_{s^-} \leqslant S_{s^-}^*\}} \middle| S_t \right] \mathrm{d}s.$$

Thus, we can conclude that the term $RCJ(t, T)$ may be neglected when we have small upside jump rate λp or when the parameter η_1 is large [in which case the integrand inside $U(t, T)$ will be small], and we can ignore the term $RCJ(t, T)$ in Eq. (43) for practical usage.

Observing that at the optimal exercise boundary S_t^*, $P_A(S_t^*, t, T) = K - S_t^*$, we obtain an integral equation for S_t^*

$$K - S_t^* = P_E(S_t^*, t, T) + \int_t^T e^{-r(s-t)} rK \mathsf{E}^*[\mathbf{1}_{\{S_s \leqslant S_s^*\}} \mid S_t = S_t^*] \, \mathrm{d}s$$

$$- \int_t^T e^{-r(s-t)} \delta \mathsf{E}^*[S_s \mathbf{1}_{\{S_s \leqslant S_s^*\}} \mid S_t = S_t^*] \, \mathrm{d}s,$$

ignoring the term $RCJ(t, T)$. To solve this integral equation, we shall use a piecewise exponential function representation for the early exercise boundary as in Ju (1998).

More precisely, with n intervals of size $\Delta T = T/n$ we approximate the optimal boundary S_t^* by an n-piece exponential function $\tilde{S}_t = \exp(s_i^* + \alpha_i t)$ for $t \in [(i-1)\Delta T, i\Delta T)$ with $i = 1, \ldots, n$. In our numerical experiments, even $n = 3$ or 5 will give sufficient accuracy in most cases. To determine the value of the constants s_i^* and α_i in each interval, we make use of the "value-matching" and "smoothing-pasting" conditions (requiring the slope at the contacting point to be -1 to make the curve smooth). Thus, starting from $i = n$ going backwards to $i = 1$ we solve recursively at $t_i = (i-1)\Delta T$ the two unknowns s_i^* and α_i in terms of the system of two equations, i.e., the value matching equation

$$K - \tilde{S}_i = P_E(\tilde{S}_i, t_i, T) + \sum_{j=i}^{n} IA_j(\tilde{S}_i, t_j) - \sum_{j=i}^{n} DL_j(\tilde{S}_i, t_j), \tag{44}$$

and the smoothing pasting equation

$$-1 = \frac{\partial}{\partial \tilde{S}_i} P_E(\tilde{S}_i, t_i, T) + \sum_{j=i}^{n} \frac{\partial}{\partial \tilde{S}_i} IA_j(\tilde{S}_i, t_j) - \sum_{j=i}^{n} \frac{\partial}{\partial \tilde{S}_i} DL_j(\tilde{S}_i, t_j), \tag{45}$$

where $\tilde{S}_i \equiv \tilde{S}_{t_i} = \exp\{s_i^* + \alpha_i t_i\}$,

$$IA_j(S_t, u) = rK \int_u^{t_{j+1}} e^{-r(s-t)} \mathsf{E}^*[\mathbf{1}_{\{S_s \leqslant \tilde{S}_s\}} \mid S_t] \, \mathrm{d}s,$$

$$t \leqslant u, \ u \in [t_j, t_{j+1}),$$

$$DL_j(S_t, u) = \delta \int_u^{t_{j+1}} e^{-r(s-t)} \mathsf{E}^*[S_s \mathbf{1}_{\{S_s \leqslant \tilde{S}_s\}} \mid S_t] \, \mathrm{d}s,$$

$$t \leqslant u, \ u \in [t_j, t_{j+1}).$$

This system of equations can be solved numerically via an iterative procedure to be specified shortly, if the right-hand sides of (44) and (45) can be computed. To this end, Kou et al. (2005) derive the Laplace transforms with respect to s_i^* of IA_j and DL_j, and the Laplace transforms of $\frac{\partial}{\partial S_t} IA_j$ and $\frac{\partial}{\partial S_t} DL_j$.

In summary, we have the following algorithm.

The Algorithm.

1. *Compute the approximation exercise boundary $\tilde{S}$ by letting i going backwards from n to* 1 *while, at each time point t_i one solves the system of two equations in* (44) *and* (45) *to get s_i^* and α_i, with the right-hand side of* (44)

and (45) *being obtained by inverting their Laplace transforms. The system of two equations can be solved, for example, by using the multi-dimensional secant method by Broydn* (*as implemented in* Press et al., 1993).

2. *After the boundary* $\tilde{S}$ *is obtained, at any time* $t \in [t_i, t_{i+1})$*, the value of the American put option is given by*

$$P_E(S_t, t, T) + IA_i(S_t, t) + \sum_{j=i+1}^{n} IA_j(S_t, t_j) - DL_i(S_t, t) - \sum_{j=i+1}^{n} DL_j(S_t, t_j).$$

In the numerical implementation, one can use the two-sided Euler algorithm in Petrella (2004) to do inversion in Step 1. The initial values for the secant method is obtained by setting $\alpha_i = 0$ and using the critical value in the approximation given by Kou and Wang (2004) as an initial value of S_i^*.

Kou et al. (2005) report the prices using a 3- and 5-piece exponential approximation of the boundary (3EXP and 5EXP, respectively), and compare the results with (i) the "true" values computed using the tree method as in Amin (1993), and (ii) the prices obtained by the analytic approximation in Kou and Wang (2004). The running time of the new algorithm is less than 2 s for 3EXP and 4 s for 5EXP, compared to more than an hour required by the Amin's tree method. In most cases 3EXP provides an estimate of the option price more accurate than the quadratic approximation in Kou and Wang (2004), and, as we would expect, 5EXP has even better accuracy.

In summary, the quadratic approximation is easier in terms of programming effort, as it is an analytical approximation, and is faster in terms of computation time. However, the piecewise exponential approximation is more accurate.

10 Extension of the jump-diffusion models to multivariate cases

Many options traded in exchanges and in over-the-counter markets, such as two-dimensional barrier options and exchange options, have payoffs depending on more than one assets. An exchange option gives the holder the right to exchange one asset to another asset. More precisely, the payoff of an exchange option is $(S_1(T) - e^{-k}S_2(T))^+$, where e^{-k} is the ratio of the shares to be exchanged. A two-dimensional barrier option has a regular call or put payoff from one asset while the barrier crossing is determined by another asset. For example, in late 1993 Bankers Trust issued a call option on a basket of Belgian stocks which would be knocked out if the Belgian franc appreciated by more than 30% (Zhang, 1998); in this case we have a up-and-out call option. There are eight types of two-dimensional barrier options: up (down)-and-in (out) call (put) options. Mathematically, the payoff of a two-dimensional up-an-in put

barrier option is $(K - S_1(T))^+ 1_{\{\max_{0\leqslant t\leqslant T} S_2(t)\geqslant H\}}$, where $S_i(t)$, $i = 1, 2$, are prices of two assets, $K > 0$ is the strike price of the put option and H is the barrier level. To price this option, it is crucial to compute the joint distribution of the first passage time

$$\mathsf{P}\Big(X_T^{(1)} \leqslant a, \max_{0\leqslant s\leqslant T} X_s^{(2)} \geqslant b\Big) = \mathsf{P}\big(X_T^{(1)} \leqslant a, \tau_b \leqslant T\big),$$

where the first passage time τ_b is defined to be $\tau_b \equiv \tau_b^{(2)} := \inf\{t \geqslant 0: X_t^{(2)} \geqslant b\}$, $b > 0$. Here $X_T^{(i)} = \log(S_i(T)/S_i(0))$ is the return process for the ith asset, $i = 1, 2$.

Analytical solutions for these options are available under the classical Brownian models; see, e.g., the books by Hull (2005) and Zhang (1998). However, it becomes difficult to retain analytical tractability after jumps being introduced, partly because of the "overshoot" problem due to the possibility of jumping over the barrier. For example, it is difficult to get analytical solutions for two-dimensional barrier options under Merton's normal jump-diffusion model.

Huang and Kou (2006) extends the previous one-dimensional double exponential jump-diffusion models by providing a multivariate jump-diffusion model with both correlated common jumps and individual jumps is proposed. The jump sizes have a multivariate asymmetric Laplace distribution (which is related but not equal to the double exponential distribution). The model not only provides a flexible framework to study correlated jumps but also is amenable for computation, especially for barrier options. Analytical solutions for the first passage time problem in two dimension are given, and analytical solutions for barrier and exchange options and other related options are also given. Compared to the one-dimensional case the two-dimensional problem poses some technical challenges. First, with both common jumps and individual jumps, the generator of the two-dimensional process becomes more involved. Second, because the joint density of the asymmetric Laplace distribution has no analytical expression, the calculation related to the joint density and generator becomes complicated. Third, one has to use several uniform integrability arguments to substantiate a martingale argument, as Itô's formula cannot be applied directly due to discontinuity.

10.1 Asymmetric Laplace distribution

The common jumps in the multivariate jump-diffusion model to be introduced next will have a multivariate asymmetric Laplace distribution. An n-dimensional asymmetric Laplace random vector Y, denoted by $Y \sim \mathcal{AL}_n(m, J)$, is defined via its characteristic function

$$\Psi_Y(\theta) = \mathsf{E}\big[e^{i\theta' Y}\big] = \frac{1}{1 + \frac{1}{2}\theta' J\theta - \mathrm{i}m'\theta}, \tag{46}$$

where $m \in R^n$ and J is an $n \times n$ positive definite symmetric matrix. The requirement of the matrix J being positive definite is postulated to guarantee that the n-dimensional distribution is nondegenerate; otherwise, the dimension of the distribution may be less than n. The vector m is the mean $\mathsf{E}[Y] = m$ and the matrix J plays a role similar to that of the variance and covariance matrix.

In the case of the univariate Laplace distribution, the characteristic function in (46) becomes

$$\Psi_Y(\theta) = \frac{1}{1 + \frac{1}{2}v^2\theta^2 - \mathrm{im}\,\theta}, \tag{47}$$

where v^2 is the equivalence of J in (46). For further information about the asymmetric Laplace distribution, see Kotz et al. (2001).

The asymmetric Laplace distribution has many properties similar to those of the multivariate normal distribution. This can be easily seen from the fact that

$$Y \stackrel{d}{=} mB + B^{1/2}Z, \tag{48}$$

where $Z \sim N_n(0, J)$ is a multivariate normal distribution with mean 0 and covariance matrix J, and B is a one-dimensional exponential random variable with mean 1, independent of Z. For example, for the kth component of Y we have $Y^{(k)} \stackrel{d}{=} m_k B + B^{1/2}Z_k$ with $B \sim \exp(1)$ and $Z_k \sim N(0, J_{kk})$, which implies that the marginal distribution of $Y^{(k)}$ has a univariate asymmetric Laplace distribution. Furthermore, the difference between any two components,

$$Y^{(k)} - Y^{(j)} \stackrel{d}{=} (m_k - m_j)B + B^{\frac{1}{2}}(Z_k - Z_j), \quad 1 \leqslant k,\ j \leqslant n, \tag{49}$$

is again a univariate Laplace distribution. However, it is worth mentioning $Y + a$ does not have the asymmetric Laplace distribution, for $a \neq 0$.

The univariate asymmetric Laplace distribution defined by its characteristic function in (47) is a special case of the double exponential distribution, because the univariate asymmetric Laplace distribution has the density function

$$f_Y(y) = p \cdot \eta_1 e^{-\eta_1 y} 1_{\{y \geqslant 0\}} + q \cdot \eta_2 e^{\eta_2 y} 1_{\{y<0\}},$$
$$p > 0,\ q > 0,\ p + q = 1,$$

but with $p\eta_1 = q\eta_2$ and the parameters given by

$$\eta_1 = \frac{2}{\sqrt{m^2 + 2v^2} + m}, \qquad \eta_2 = \frac{2}{\sqrt{m^2 + 2v^2} - m},$$
$$p = \frac{\sqrt{m^2 + 2v^2} + m}{2\sqrt{m^2 + 2v^2}}. \tag{50}$$

Asymmetric Laplace distribution can also be viewed as a special case of the generalized hyperbolic distribution introduced by Barndorff-Nielsen (1977). In

fact, a generalized hyperbolic random variable X is defined as $X \stackrel{d}{=} \mu + m\zeta + \zeta^{1/2}Z$, where Z is a multivariate normal distribution, ζ is a generalized inverse Gaussian distribution. Since the exponential random variable belongs to generalized inverse Gaussian distribution, the asymmetric Laplace distribution is a special case of the generalized hyperbolic distribution. For more details on applications of the generalized hyperbolic distribution in finance, see Eberlein and Prause (2002).

10.2 A multivariate jump-diffusion model

We propose a multivariate jump-diffusion model in which the asset prices $S(t)$ have two parts, a continuous part driven by a multivariate geometric Brownian motion, and a jump part with jump events modeled by a Poisson process. In the model, there are both common jumps and individual jumps. More precisely, if a Poisson event corresponds to a common jump, then all the asset prices will jump according to the multivariate asymmetric Laplace distribution; otherwise, if a Poisson event corresponds to an individual jump of the jth asset, then only the jth asset will jump. In other words, the model attempts to capture various ways of correlated jumps in asset prices.

Mathematically, under the physical measure P the following stochastic differential equation is proposed to model the asset prices $S(t)$:

$$\frac{\mathrm{d}S(t)}{S(t-)} = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) + \mathrm{d}\left(\sum_{i=1}^{N(t)}(V_i - 1)\right), \tag{51}$$

where $W(t)$ is an n-dimensional standard Brownian motion, $\sigma \in R^{n\times n}$ with the covariance matrix $\Sigma = \sigma\sigma^T$. The rate of the Poisson process $N(t)$ process is $\lambda = \lambda_c + \sum_{k=1}^{n}\lambda_k$; in other words, there are two types of jumps, common jumps for all assets with jump rate λ_c and individual jumps with rate λ_k, $1 \leqslant k \leqslant n$, only for the kth asset.

The logarithms of the common jumps have an m-dimensional asymmetric Laplace distribution $\mathcal{AL}_n(m_c, J_c)$, where $m_c = (m_{1,c}, \ldots, m_{n,c})' \in R^n$ and $J_c \in R^{n\times n}$ is positive definite. For the individual jumps of the kth asset, the logarithms of the jump sizes follow a one-dimensional asymmetric Laplace distribution, $\mathcal{AL}_1(m_k, v_k^2)$. In summary

$$Y = \log(V) \sim \begin{cases} \mathcal{AL}_n(m_c, J_c), & \text{with prob. } \lambda_c/\lambda, \\ (\underbrace{0, \ldots, 0}_{k-1}, \mathcal{AL}_1(m_k, v_k^2), \underbrace{0, \ldots, 0}_{n-k})', \\ \quad \text{with prob. } \lambda_k/\lambda,\ 1 \leqslant k \leqslant n. \end{cases}$$

The sources of randomness, $N(t)$, $W(t)$ are assumed to be independent of the jump sizes V_is. Jumps at different times are assumed to be independent.

Note that in the univariate case the above model degenerates to the double exponential jump-diffusion model (Kou, 2002) but with $p\eta_1 = q\eta_2$.

Solving the stochastic differential equation in (51) gives the dynamic of the asset prices:

$$S(t) = S(0)\exp\left[\left(\mu - \frac{1}{2}\Sigma_{\text{diag}}\right)t + \sigma W(t)\right]\prod_{i=1}^{N(t)} V_i, \tag{52}$$

where Σ_{diag} denotes the diagonal vector of Σ. Note that $\forall 1 \leqslant k \leqslant n$,

$$\mathsf{E}\big(V^{(k)}\big) = \mathsf{E}\big(e^{Y^{(k)}}\big) = \frac{\lambda_c/\lambda}{1 - m_{k,c} - J_{c,kk}/2} + \frac{\lambda_k/\lambda}{1 - m_k - v_k^2/2}. \tag{53}$$

The requirements $m_{k,c} + J_{c,kk}/2 < 1$ and $m_k + v_k^2/2 < 1$ are needed to ensure $\mathsf{E}(V^{(k)}) < \infty$ and $\mathsf{E}(S_k(t)) < \infty$, i.e. the stock price has finite expectation.

In the special case of two-dimension, the asset prices can be written as

$$\begin{aligned} S_1(t) &= S_1(0)\exp\left[\mu_1 t + \sigma_1 W_1(t) + \sum_{i=1}^{N(t)} Y_i^{(1)}\right], \\ S_2(t) &= S_2(0)\exp\left[\mu_2 t + \sigma_2\big[\rho W_1(t) + \sqrt{1-\rho^2}W_2(t)\big] + \sum_{i=1}^{N(t)} Y_i^{(2)}\right]. \end{aligned} \tag{54}$$

Here all the parameters are risk-neutral parameters, $W_1(t)$ and $W_2(t)$ are two independent standard Brownian motions, and $N(t)$ is a Poisson process with rate $\lambda = \lambda_c + \lambda_1 + \lambda_2$. The distribution of the logarithm of the jump sizes Y_i is given by

$$Y_i = \big(Y_i^{(1)}, Y_i^{(2)}\big)' \sim \begin{cases} \mathcal{AL}_2(m_c, J_c), & \text{with prob. } \lambda_c/\lambda, \\ (\mathcal{AL}_1(m_1, v_1^2), 0)', & \text{with prob. } \lambda_1/\lambda, \\ (0, \mathcal{AL}_1(m_2, v_2^2))', & \text{with prob. } \lambda_2/\lambda, \end{cases} \tag{55}$$

where the parameters for the common jumps are

$$m_c = \begin{pmatrix} m_{1,c} \\ m_{2,c} \end{pmatrix}, \qquad J_c = \begin{pmatrix} v_{1,c}^2 & cv_{1,c}v_{2,c} \\ cv_{1,c}v_{2,c} & v_{2,c}^2 \end{pmatrix}.$$

Since $S(t)$ is a Markov process, an alternative characterization of $S(t)$ is to use the generator of $X(t) = \log S(t)/S(0)$. The two-dimensional jump-diffusion return process $(X_1(t), X_2(t))$ in (54) is given by

$$\begin{aligned} X_1(t) &= \mu_1 t + \sigma_1 W_1(t) + \sum_{i=1}^{N(t)} Y_i^{(1)}, \\ X_2(t) &= \mu_2 t + \sigma_2\big[\rho W_1(t) + \sqrt{1-\rho^2}W_2(t)\big] + \sum_{i=1}^{N(t)} Y_i^{(2)}, \end{aligned}$$

with the infinitesimal generator

$$\begin{aligned}\mathcal{L}u = {} & \mu_1 \frac{\partial u}{\partial x_1} + \mu_2 \frac{\partial u}{\partial x_2} + \frac{1}{2}\sigma_1^2 \frac{\partial^2 u}{\partial x_1^2} + \frac{1}{2}\sigma_2^2 \frac{\partial^2 u}{\partial x_2^2} + \rho\sigma_1\sigma_2 \frac{\partial^2 u}{\partial x_1 \partial x_2} \\ & + \lambda_c \int_{y_2=-\infty}^{\infty} \int_{y_1=-\infty}^{\infty} \big[u(x_1 + y_1, x_2 + y_2) - u(x_1, x_2)\big] \\ & \times f^c_{(Y^{(1)}, Y^{(2)})}(y_1, y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2 \\ & + \lambda_1 \int_{y_1=-\infty}^{\infty} \big[u(x_1 + y_1, x_2) - u(x_1, x_2)\big] f_{Y^{(1)}}(y_1)\, \mathrm{d}y_1 \\ & + \lambda_2 \int_{y_2=-\infty}^{\infty} \big[u(x_1, x_2 + y_2) - u(x_1, x_2)\big] f_{Y^{(2)}}(y_2)\, \mathrm{d}y_2, \end{aligned} \tag{56}$$

for all continuously twice differentiable function $u(x_1, x_2)$, where $f^c_{(Y^{(1)}, Y^{(2)})}(y_1, y_2)$ is the joint density of correlated common jumps $\mathcal{AL}_2(m_c, J_c)$, and $f_{Y^{(i)}}(y_i)$ is the individual jump density of $\mathcal{AL}_1(m_i, J_i)$, $i = 1, 2$.

One difficulty in studying the generator is that the joint density of the asymmetric Laplace distribution has no analytical expression. Therefore, the calculation related to the joint density and generator becomes complicated. See Huang and Kou (2006) for change of measures from a physical measure to a risk-neutral measure, analytical solutions for the first passage times, and pricing formulae for barrier options and exchange options.

References

Abate, J., Whitt, W. (1992). The Fourier series method for inverting transforms of probability distributions. *Queueing Systems* 10, 5–88.

Amin, K. (1993). Jump-diffusion option valuation in discrete time. *Journal of Finance* 48, 1833–1863.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9, 203–228.

Asmussen, S., Avram, F., Pistorius, M.R. (2004). Russian and American put options under exponential phase-type Lévy models. *Stochastic Processes and Their Applications* 109, 79–111.

Avram, F., Kyprianou, A.E., Pistorius, M.R. (2004). Exit problems for spectrally negative Lévy processes and applications to (Canadized) Russian options. *Annals of Applied Probability* 14, 215–238.

Barberis, N., Shleifer, A., Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics* 49, 307–343.

Barone-Adesi, G., Whaley, R.E. (1987). Efficient analytic approximation of American option values. *Journal of Finance* 42, 301–320.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London Series A* 353, 409–419.

Barndorff-Nielsen, O.E., Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion). *Journal of Royal Statistical Society Series B* 63, 167–241.

Blattberg, R.C., Gonedes, N.J. (1974). A comparison of the stable and Student distributions as statistical models for stock prices. *Journal of Business* 47, 244–280.
Boyarchenko, S., Levendorskiĭ, S. (2002). Barrier options and touch-and-out options under regular Lévy processes of exponential type. *Annals of Applied Probability* 12, 1261–1298.
Boyle, P., Broadie, M., Glasserman, P. (1997). Simulation methods for security pricing. *Journal of Economic Dynamics and Control* 21, 1267–1321.
Carr, P., Madan, D.B. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance* 2, 61–73.
Carr, P., Jarrow, R., Myneni, R. (1992). Alternative characterizations of American puts. *Mathematical Finance* 2, 87–106.
Carr, P., Geman, H., Madan, D., Yor, M. (2003). Stochastic volatility for Lévy processes. *Mathematical Finance* 13, 345–382.
Chen, N., Kou, S.G., 2005. Credit spreads, optimal capital structure, and implied volatility with endogenous default and jump risk. *Mathematical Finance*, in press.
Choudhury, G.L., Lucantoni, D.M., Whitt, W. (1994). Multidimensional transform inversion with applications to the transient M/M/1 queue. *Annals of Applied Probability* 4, 719–740.
Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 135–155.
Cont, R., Tankov, P. (2004). *Financial Modelling with Jump Processes*, second printing. Chapman and Hall/CRC Press, London.
Cont, R., Voltchkova, E. (2005). Finite difference methods for option pricing in jump-diffusion and exponential Lévy models. *SIAM Journal of Numerical Analysis* 43, 1596–1626.
Cox, J.T., Ross, S.A. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3, 145–166.
Craddock, M., Heath, D., Platen, E. (2000). Numerical inversion of Laplace transforms: A survey of techniques with applications to derivative pricing. *Journal of Computational Finance* 4, 57–81.
Das, S.R., Foresi, S. (1996). Exact solutions for bond and option prices with systematic jump risk. *Review of Derivatives Research* 1, 7–24.
Davydov, D., Linetsky, V. (2001). Pricing and hedging path-dependent options under the CEV process. *Management Science* 47, 949–965.
Derman, E., Kani, I. (1994). Riding on a smile. *RISK* Feb. 32–39.
d'Halluin, Y., Forsyth, P.A., Vetzal, K.R. (2003). Robust numerical methods for contingent claims under jump-diffusion processes. *Working paper*, University of Waterloo.
Duffie, D., Pan, J., Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68, 1343–1376.
Dupire, B. (1994). Pricing with a smile. *RISK* Feb. 18–20.
Eberlein, E., Prause, K. (2002). The generalized hyperbolic model: Financial derivatives and risk measures. In: Geman, H., Madan, D., Pliska, S., Vorst, T. (Eds.), *Mathematical Finance-Bachelier Congress 2000*. Springer-Verlag, New York, pp. 245–267.
Engle, R. (1995). *ARCH, Selected Readings*. Oxford Univ. Press, Oxford, UK.
Fama, E. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49, 283–306.
Feng, L., Linetsky, V. (2005). Pricing options in jump-diffusion models: An extrapolation approach. *Operations Research*, in press.
Feng, L., Linetsky, V., Marcozzi, M. (2004). On the valuation of options in jump-diffusion models by variational methods. *Preprint*, Northwestern University.
Fouque, J.-P., Papanicolaou, G., Sircar, K.R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge Univ. Press, Cambridge, UK.
Fu, M., Madan, D., Wang, T. (1999). Pricing continuous Asian options: A comparison of Monte Carlo and Laplace transform inversion methods. *Journal of Computational Finance* 2, 49–74.
Gatto, M., Goldman, M.B., Sosin, H. (1979). Path dependent options: "Buy at the low, sell at the high". *Journal of Finance* 34, 1111–1127.
Geman, H., Yor, M. (1993). Bessel processes, Asian options and perpetuities. *Mathematical Finance* 3, 349–375.

Glasserman, P., Kou, S.G. (2003). The term structure of simple forward rates with jump risk. *Mathematical Finance* 13, 383–410.
Gukhal, C.R. (2001). Analytical valuation of American options on jump-diffusion processes. *Mathematical Finance* 11, 97–115.
Haug, E.G. (1999). Barrier put–call transformations. *Working paper*, Tempus Financial Engineering.
Heston, S. (1993). A closed-form solution of options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6, 327–343.
Heyde, C.C. (2000). A risky asset model with strong dependence through fractal activity time. *Journal of Applied Probability* 36, 1234–1239.
Heyde, C.C., Kou, S.G. (2004). On the controversy over tailweight of distributions. *Operations Research Letters* 32, 399–408.
Heyde, C.C., Kou, S.G., Peng, X.H. (2006). What is a good risk measure: Bridging the gaps between robustness, subadditivity, prospect theory, and insurance risk measures. *Preprint*, Columbia University.
Huang, Z., Kou, S.G. (2006). First passage times and analytical solutions for options on two assets with jump risk. *Preprint*, Columbia University.
Hull, J. (2005). *Options, Futures, and Other Derivatives Securities*, sixth ed. Prentice Hall, New Jersey.
Hull, J., White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 281–300.
Johnson, N., Kotz, S., Balakrishnan, N. (1995). *Continuous Univariate Distribution*, vol. 2, second ed. Wiley, New York.
Ju, N. (1998). Pricing an American option by approximating its early exercise boundary as a multipiece exponential function. *Review of Financial Studies* 11, 627–646.
Karatzas, I., Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.
Karlin, S., Taylor, H. (1975). *A First Course in Stochastic Processes*, second ed. Academic Press, New York.
Kijima, M. (2002). *Stochastic Processes with Applications to Finance*. Chapman and Hall, London.
Kotz, S., Kozubowski, T.J., Podgórski, K. (2001). *The Laplace Distribution and Generalization: A Revisit with Applications to Communications, Economics, Engineering and Finance*. Birkhäuser, Boston.
Kou, S.G. (2002). A jump-diffusion model for option pricing. *Management Science* 48, 1086–1101.
Kou, S.G., Wang, H. (2003). First passage time of a jump diffusion process. *Advances in Applied Probability* 35, 504–531.
Kou, S.G., Wang, H. (2004). Option pricing under a double exponential jump-diffusion model. *Management Science* 50, 1178–1192.
Kou, S.G., Petrella, G., Wang, H. (2005). Pricing path-dependent options with jump risk via Laplace transforms. *Kyoto Economic Review* 74, 1–23.
Kyprianou, A., Pistorius, M. (2003). Perpetual options and Canadization through fluctuation theory. *Annals of Applied Probability* 13, 1077–1098.
Lucas, R.E. (1978). Asset prices in an exchange economy. *Econometrica* 46, 1429–1445.
MacBeth, J., Merville, L. (1979). An empirical examination of the Black–Scholes call option pricing model. *Journal of Finance* 34, 1173–1186.
Madan, D.B., Seneta, E. (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business* 63, 511–524.
Madan, D.B., Carr, P., Chang, E.C. (1998). The variance gamma process and option pricing. *European Finance Review* 2, 79–105.
Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36, 394–419.
Merton, R.C. (1973). The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–183.
Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.
Naik, V., Lee, M. (1990). General equilibrium pricing of options on the market portfolio with discontinuous returns. *Review of Financial Studies* 3, 493–521.
Petrella, G. (2004). An extension of the Euler Laplace transform inversion algorithm with applications in option pricing. *Operations Research Letters* 32, 380–389.

Petrella, G., Kou, S.G. (2004). Numerical pricing of discrete barrier and lookback options via Laplace transforms. *Journal of Computational Finance* 8, 1–37.

Pham, H. (1997). Optimal stopping, free boundary and American option in a jump-diffusion model. *Applied Mathematics and Optimization* 35, 145–164.

Press, W., Teukolsky, S., Vetterling, W., Flannery, B. (1993). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge Univ. Press, New York.

Ramezani, C.A., Zeng, Y. (2002). Maximum likelihood estimation of asymmetric jump-diffusion process: Application to security prices. *Working paper*, University of Missouri at Kansas City, Department of Mathematics and Statistics.

Rogers, L.C.G. (1997). Arbitrage from fractional Brownian motion. *Mathematical Finance* 7, 95–105.

Rogers, L.C.G. (2000). Evaluating first-passage probabilities for spectrally one-sided Lévy processes. *Journal of Applied Probability* 37, 1173–1180.

Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978. *Journal of Finance* 40, 455–480.

Samorodnitsky, G., Taqqu, M.S. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York.

Schröder, M. (1999). Changes of numéraire for pricing futures, forwards and options. *Review of Financial Studies* 12, 1143–1163.

Sepp, A. (2004). Analytical pricing of double-barrier options under a double exponential jump diffusion process: Applications of Laplace transform. *International Journal of Theoretical and Applied Finance* 7, 151–175.

Singleton, K. (2006). *Empirical Dynamic Asset Pricing*. Princeton Univ. Press, Princeton, NJ.

Stokey, N.L., Lucas, R.E. (1989). *Recursive Methods in Economic Dynamics*. Harvard Univ. Press, Cambridge, MA.

Zhang, P.G. (1998). *Exotic Options*, second ed. World Scientific, Singapore.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15003-9

Chapter 3

Modeling Financial Security Returns Using Lévy Processes

Liuren Wu

Zicklin School of Business, Baruch College, City University of New York
E-mail: Liuren_Wu@baruch.cuny.edu
url: http://faculty.baruch.cuny.edu/lwu/

Abstract

Lévy processes can capture the behaviors of return innovations on a full range of financial securities. Applying stochastic time changes to the Lévy processes randomizes the clock on which the processes run, thus generating stochastic volatilities and stochastic higher return moments. Therefore, with appropriate choices of Lévy processes and stochastic time changes, we can capture the return dynamics of virtually all financial securities. Furthermore, in contrast to the hidden factor approach, we can readily assign explicit economic meanings to each Lévy process component and its associated time change in the return dynamics. The economic mapping not only facilitates the interpretation of existing models and their structural parameters, but also adds economic intuition and direction for designing new models capturing new economic behaviors. Finally, under this framework, the analytical tractability of a model for derivative pricing and model estimation originates from the tractability of the Lévy process specification and the tractability of the activity rate dynamics underlying the time change. Thus, we can design tractable models using any combination of tractable Lévy specifications and tractable activity rate dynamics. I elaborate through examples on the generality of the framework in capturing the return behavior of virtually all financial securities, the explicit economic mapping that facilitates the interpretation and creation of new models, and the tractability embedded in the framework for derivative pricing and model estimation.

Keywords: Lévy processes; Return innovations; Stochastic time changes; Stochastic volatility; Characteristic functions; Exponential martingales; Measure change; Option pricing; Fourier inversion

1 Introduction

Since Black and Scholes (1973), Brownian motion has emerged as the benchmark process for describing asset returns in continuous time. Brownian

motion generates normally distributed return innovations. Merton (1976) augments the Brownian motion with a compound Poisson process with normally distributed jump sizes in the asset return. As a result, the return innovation distribution becomes a mixture of normals weighted by Poisson probabilities. These two innovation specifications have dominated the continuous-time finance literature for several decades, drawing criticisms that the continuous-time framework is not as flexible as the discrete-time framework: One can assume any arbitrary distribution for the return innovation in discrete time, but only normals or mixtures of normals could be generated from continuous-time specifications.

The recent introduction of Lévy processes into financial modeling exonerates continuous-time finance from such criticisms. A Lévy process can generate a much wider spectrum of distributions at a fixed time horizon. While the Brownian motion component in a Lévy process generates a normal distribution, non-normal distributions can be generated via the appropriate specification of the Lévy density for a Lévy jump process, which determines the arrival rate of jumps of all possible sizes.

Financial security returns can be driven by several economic forces. The impact of each force can vary stochastically over time. Accordingly, we can model the return innovation using several Lévy processes as building blocks matching the distributional behavior of shocks from different economic forces. Furthermore, applying stochastic time change to each Lévy component randomizes the clock on which the process runs, thus capturing the stochastically varying impacts from different economic forces. Statistically, applying stochastic time changes to different Lévy components can generate both stochastic volatility and stochastic higher return moments, both of which are well-documented features for financial securities. Therefore, with appropriate choices of Lévy processes and stochastic time changes, we can capture the return dynamics of virtually all financial securities.

Generality is not the only virtue of Lévy processes. By modeling return dynamics using different combinations of Lévy components with time changes, we can readily assign explicit economic meanings to each Lévy component and its associated time change in the return dynamics. The explicit economic mapping not only facilitates the interpretation of existing models and their structural parameters, but also adds economic intuition and direction for designing new models that are parsimonious and yet adequate in capturing the requisite economic behaviors.

A common approach in the literature is to model returns by a set of hidden statistical factors. Factor rotations make it inherently difficult to assign economic meanings to the statistical factors. The absence of economic mapping also makes the model design process opaque. One often finds that a generic hidden-factor model cannot match the requisite target behaviors of the financial securities returns, and yet many parameters of the model are difficult to identify empirically. The issue of being both "too little" in performance and "too much" in model identification can only be solved by exhaustive econometr-

ric analysis. In contrast, by mapping each Lévy process to an economic force, and using random time change to capture its intensity variation, we can readily construct parsimonious models that generate the target behavior.

The generality of the framework does not hinders its analytical tractability for derivative pricing and model estimation, either. When modeling return dynamics using Lévy processes with time changes, tractability of the return dynamics originates from tractability of the Lévy component specification and tractability of the activity rate dynamics underlying the time change. Thus, we can design tractable models using any combinations of tractable Lévy processes and tractable activity rate dynamics. In this regard, we can incorporate and hence *encompass* all tractable models in the literature as building blocks. Examples of tractable Lévy specifications include Brownian motions, compound Poisson jumps, and other tractable jump specifications like variance gamma, dampened power law, normal inverse Gaussian, and so on. Examples of tractable activity rate dynamics include the affine class of Duffie et al. (2000), the quadratic class of Leippold and Wu (2002), and the 3/2 process of Heston (1997) and Lewis (2001). By modeling financial securities returns with time-changed Lévy processes, we encompass all these models into one general and yet tractable framework.

Through examples, I elaborate the three key virtues of Lévy processes with stochastic time changes:

(i) the generality of the framework in capturing the return behavior of virtually all financial securities,
(ii) the explicit economic mapping that facilitates the interpretation and creation of new models capturing specific economic behaviors, and
(iii) the tractability embedded in the framework for derivative pricing and model estimation.

In designing models for a financial security return, the literature often starts by specifying a very general process with a set of hidden factors and then testing different restrictions on this general process. Here I take the opposite approach. First, I look at the data and identify stylized features that a reasonable model needs to capture. Second, I design different components of the model to match different features of the data and capture the impacts from different economic forces. The final step is to assemble all the parts together. Using time-changed Lévy processes matches this procedure well. First, we can choose Lévy components to match the properties of return innovations generated from different economic forces. Statistically, we ask the following set of questions: Do we need a continuous component? Do we need a jump component? Do the jumps arrive frequently or are they rare but large events? Do up and down movements show different behaviors?

Once we have chosen the appropriate Lévy components, we can use time changes to capture the intensity variation for the different components and generate stochastic volatilities and stochastic higher return moments from different economic sources. We use time changes to address the following ques-

tions: Is stochastic volatility driven by intensity variations of small movements (Brownian motion) or large movements (jumps)? Do the intensities of different types of movements vary synchronously or separately? Do they show any dynamic interactions? Based on answers to these questions, we can apply different time changes to different Lévy components and model their intensity dynamics in a way matching their observed dynamic interactions.

The final step involves assembling the different Lévy components with or without time changes into the asset return dynamics. When the dynamics are specified under the risk-neutral measure for derivative pricing, adjustments are necessary to guarantee the martingale property.

When designing models, tractability requirement often comes from derivative pricing since we need to take expectations of future contingent payoffs under the risk-neutral measure to obtain its present value. Thus, it is often convenient to start by specifying a tractable return dynamics under the risk-neutral measure. Then the statistical dynamics can be derived based on market price of risk specifications. The less stringent requirement for tractability for the statistical dynamics often allows us to specify very flexible market price of risk specifications, with the only constraints coming from reasonability for investor behaviors and parsimony for econometric identification.

In designing the different Lévy components and applying the time changes, I quote Albert Einstein as the guiding principle: "*Everything should be made as simple as possible, but not simpler*." The explicit economic purpose for each Lévy component and its time change allows us to abide by this guiding principle much more easily than in a general hidden statistical factor framework.

The rest of the article is organized as follows. The next section discusses Lévy processes and how they can be used to model return innovations. Section 3 discusses how to use time changes to generate stochastic volatility and stochastic higher moments from different sources. Section 4 discusses how to assemble different pieces together, how to make appropriate adjustments to satisfy the martingale condition under the risk-neutral measure, and how to derive the statistical dynamics based on market price of risk specifications. Section 5 discusses option pricing under time-changed Lévy processes. Section 6 addresses the estimation issues using time-series returns and/or option prices. Section 7 concludes.

2 Modeling return innovation distribution using Lévy processes

A Lévy process is a continuous time stochastic process with stationary independent increments, analogous to i.i.d. innovations in a discrete-time setting. Until very recently, the finance literature narrowly focuses on two examples of Lévy processes: the Brownian motion underlying the Black and Scholes (1973) model and the compound Poisson jump process with normal jump sizes underlying the jump diffusion model of Merton (1976). A Brownian motion generates normal innovations. The compound Poisson process in the Merton model

generates return non-normality through a mixture of normal distributions with Poisson probability weightings. A general Lévy process can generate a much wider range of distributional behaviors through different types of jump specifications. The compound Poisson process used in the Merton model generates a finite number of jumps within a finite time interval. Such a jump process is suitable to capture rare and large events such as market crashes and corporate defaults. Nevertheless, many observe that asset prices can also display many small jumps on a fine time scale. A general Lévy process can not only generate continuous movements via a Brownian motion and rare and large events via a compound Poisson process, but it can also generate frequent jumps of different sizes.

2.1 Lévy characteristics

We start with a one-dimensional real-valued stochastic process $\{X_t | t \geqslant 0\}$ with $X_0 = 0$ defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ endowed with a standard complete filtration $\mathbf{F} = \{\mathcal{F}_t | t \geqslant 0\}$. We assume that X is a Lévy process with respect to the filtration $\mathbf{F}$, that is, X_t is adapted to $\mathcal{F}_t$, the sample paths of X are right-continuous with left limits, and $X_u - X_t$ is independent of $\mathcal{F}_t$ and distributed as X_{u-t} for $0 \leqslant t < u$. By the Lévy–Khintchine Theorem, the characteristic function of X_t has the form,

$$\phi_{X_t}(u) \equiv \mathbb{E}^{\mathbb{P}}\left[e^{iuX_t}\right] = e^{-t\psi_x(u)}, \quad t \geqslant 0, \tag{1}$$

where the *characteristic exponent* $\psi_x(u)$ is given by,

$$\psi_x(u) = -iu\mu + \frac{1}{2}u^2\sigma^2 + \int_{\mathbb{R}_0} \left(1 - e^{iux} + iux1_{|x|<1}\right)\pi(x)\,\mathrm{d}x, \tag{2}$$

where $\mu \in \mathbb{R}$ describes the constant drift, $\sigma^2 \in \mathbb{R}^+$ describes the constant variance of the continuous component of the Lévy process, and the *Lévy density* $\pi(x)$ describes the arrival rates for jumps of every possible size x. The triplet (μ, σ^2, π) fully specifies the Lévy process X_t and is referred to as the *Lévy characteristics* (Bertoin, 1996).

With a fixed time horizon, any return distribution can be represented uniquely by its characteristic function and hence its characteristic exponent. Equation (2) illustrates that a Lévy process can generate a wide range of characteristic exponent behaviors through a flexible specification of the Lévy density $\pi(x)$.

The Lévy density $\pi(x)$ is defined on the real line excluding zero, $\mathbb{R}_0$. The truncation function $x1_{|x|<1}$ equals x when $|x| < 1$ and zero otherwise. Other truncation functions are also used in the literature as long as they are bounded, with compact support, and satisfy $h(x) = x$ in a neighborhood of zero (Jacod

and Shiryaev, 1987).[1] The purpose of the truncation function is to analyze the jump properties around the singular point of zero jump size.

The characteristic function in (1) is defined on the real line $u \in \mathbb{R}$. In many applications, it is convenient to extend the definition to the complex plane, $u \in \mathcal{D} \subseteq \mathcal{C}$, where the characteristic exponent is well-defined. When $\phi_{X_t}(u)$ is defined on the complex plane, it is referred to as the *generalized* Fourier transform (Titchmarsh, 1986). It is also helpful to define the cumulant exponent of a Lévy process X_t,

$$\begin{aligned}\varphi_x(s) &\equiv \frac{1}{t}\ln \mathbb{E}\big[e^{sX_t}\big] \\ &= s\mu + \frac{1}{2}s^2\sigma^2 + \int_{\mathbb{R}_0}\big(e^{sx} - 1 - sx1_{|x|<1}\big)\pi(x)\,\mathrm{d}x, \quad s \in \mathcal{D}_s \subseteq \mathcal{C},\end{aligned} \tag{3}$$

where $\mathcal{D}_s$ denotes the subsect of the complex plain under which the cumulant exponent is well-defined. Our extensions on the domains of the characteristic coefficient u and cumulant coefficient s implies that $\psi_x(u) = -\varphi_x(iu)$ whenever the two are both well-defined. Option pricing and likelihood estimation for Lévy processes often rely on the tractability of the characteristic exponent and specifically, analytical solutions to the integral in Eq. (2) or (3).

The sample paths of a pure jump Lévy process exhibit *finite activity* when the integral of the Lévy density is finite:

$$\int_{\mathbb{R}^0} \pi(x)\,\mathrm{d}x = \lambda < \infty, \tag{4}$$

where λ measures the mean arrival rate of jumps. A finite activity jump process generates a finite number of jumps within any finite time interval.

When the integral in (4) is infinite, the sample paths exhibit *infinite activity*, and generate an infinite number of jumps within any finite interval. Nevertheless, the sample paths show *finite variation* if the following integral is finite:

$$\int_{\mathbb{R}_0} \big(|x| \wedge 1\big)\pi(x)\,\mathrm{d}x < \infty. \tag{5}$$

When the integral in (5) is infinite, the jump process exhibit *infinite variation*, a property also shared by the Brownian motion. The truncation function in the definition of characteristic exponent is needed only for infinite variation jumps. When the integral in (5) is not finite, the sum of small jumps does not converge, but the sum of the jumps compensated by their mean converges. This special behavior generates the necessity for the truncation term in (2).

[1] Commonly used truncation functions include $h(x) = x/(1 + x^2)$, and $h(x) = 1 \wedge |x|$ (the minimum of 1 and $|x|$).

For all jump specifications, we require that the process exhibit *finite quadratic variation*:

$$\int_{\mathbb{R}_0} \left(1 \wedge x^2\right) \pi(x)\, \mathrm{d}x < \infty, \tag{6}$$

a necessary condition for the jump process to be a semimartingale. Lévy processes are within a subclass of semimartingales.

2.2 Lévy examples

Black and Scholes (1973) model the asset return by a purely continuous Lévy process and hence with $\pi(x) = 0$ for all x. The characteristic exponent is simply:

$$\psi(u) = -iu\mu + \frac{1}{2}u^2\sigma^2. \tag{7}$$

The associated normal probability density function is also well known.

Merton (1976) incorporates an additional compound Poisson jump component with mean arrival rate λ. The jump size in the log asset return is normally distributed with mean μ_J and variance v_J, conditional on one jump occurring. The Lévy density of this jump component can be written as,

$$\pi(x) = \lambda \frac{1}{\sqrt{2\pi v_J}} \exp\left(-\frac{(x-\mu_J)^2}{2v_J}\right). \tag{8}$$

The characteristic exponent for this compound Poisson jump is:

$$\psi(u) = \lambda\left(1 - e^{iu\mu_J - \frac{1}{2}u^2 v_J}\right). \tag{9}$$

A key property of compound Poisson jumps is that the sample paths exhibit finite activity. Finite-activity jumps are useful in capturing large but rare events. For example, the credit-risk literature has used Poisson process extensively to model the random arrival of default events (Lando, 1998; Duffie and Singleton, 1999, 2003; and Duffie et al., 2003b). More recently, Carr and Wu (2005) use a Poisson jump with zero recovery to model the impact of corporate default on the stock price. Upon arrival, the stock price jumps to zero. Carr and Wu (2007b) use a Poisson jump with random recovery to model the impact of sovereign default on its home currency price. Upon arrival, the currency price jumps down by a random amount.

Within the compound Poisson jump type, Kou (2002) proposes a double-exponential conditional distribution for the jump size. The Lévy density is given by,

$$\pi(x) = \begin{cases} \lambda \exp(-\beta_+ x), & x > 0, \\ \lambda \exp(-\beta_- |x|), & x < 0, \end{cases} \quad \lambda, \beta_+, \beta_- > 0. \tag{10}$$

Under this specification, the jump arrival rate increases monotonically with decreasing jump size. Asymmetry between up and down jumps are induced by the different exponential coefficients β_+ and β_-. The characteristic exponent for this pure jump process is,

$$\psi(u) = -\lambda\big[(\beta_+ - iu)^{-1} - \beta_+^{-1} + (\beta_- + iu)^{-1} - \beta_-^{-1}\big]. \tag{11}$$

Kou and Wang (2004) show that the double-exponential jump specification allows tractable pricing for American and some path-dependent options.

Although it is appropriate to use compound Poisson jumps to capture rare and large events such as market crashes and corporate defaults, many observe that asset prices actually display many small jumps. These types of behaviors are better captured by infinite-activity jumps, which generate infinite number of jumps within any finite time interval. A popular example that can generate different jump types is the CGMY model of Carr et al. (2002), with the following Lévy density,

$$\pi(x) = \begin{cases} \lambda \exp(-\beta_+ x) x^{-\alpha-1}, & x > 0, \\ \lambda \exp(-\beta_- |x|) |x|^{-\alpha-1}, & x < 0, \end{cases}$$
$$\lambda, \beta_+, \beta_- > 0, \alpha \in [-1, 2]. \tag{12}$$

In this specification, the power coefficient α controls the arrival frequency of small jumps and hence the jump type. With the power coefficient $\alpha = -1$, the Lévy density becomes the double-exponential specification in (10), the sample paths of which show finite activity. The model generates finite-activity jumps as long as $\alpha < 0$. When $\alpha \in [0, 1)$, the model generates jumps with infinite activity but finite variation. The jump process exhibits infinite variation when $\alpha \in [1, 2]$. The condition $\alpha \leqslant 2$ is necessary to guarantee finite quadratic variation. With $\alpha < 0$, the power term makes the jump arrival approaches infinity as the jump size approaches zero. The larger the power coefficient, the higher the frequency of small jumps. The two exponential coefficients β_+ and β_- control the arrival of large jumps. The difference in the two coefficients generates asymmetry in the tails of the distribution.

The physics literature often refers to the specification in (12) as truncated Lévy flights. The CGMY terminology comes from the initials of the four authors in Carr et al. (2002), who regard the model as an extension of the variance gamma (VG) model of Madan and Seneta (1990) and Madan et al. (1998). Under the VG model, $\alpha = 0$. Wu (2006) labels the specification in (12) as *exponentially dampened power law* (DPL), regarding it as the Lévy density of an α-stable Lévy process with exponential dampening. Wu shows that applying measure changes using exponential martingales to an α-stable Lévy process generates the exponentially dampened power law. Hence, the whole class of α-stable processes, made popular to the finance field by Mandelbrot (1963) and Fama (1965), can be regarded as a special class of the dampened power law.

When $\alpha \neq 0$ and $\alpha \neq 1$, the characteristic exponent associated with the dampened power law Lévy density specification takes the following form:

$$\psi(u) = -\Gamma(-\alpha)\lambda\big[(\beta_+ - iu)^\alpha - \beta_+^\alpha + (\beta_- + iu)^\alpha - \beta_-^\alpha\big] - iuC(h), \tag{13}$$

where $\Gamma(a) \equiv \int_0^\infty x^{a-1}e^{-x}\,\mathrm{d}x$ is the gamma function and the linear term $C(h)$ is induced by the inclusion of a truncation function $h(x)$ for infinite-variation jumps when $\alpha > 1$. As I will make clear in later sections, in modeling return dynamics, any linear drift term in X_t will be canceled out by the corresponding term in its concavity adjustment. Hence, the exact form of the truncation function and the resultant coefficient $C(h)$ are immaterial for modeling and estimation. Wu (2006) explicitly carries out the integral in (3) through an expansion method and solves the truncation-induced term $C(h)$ under the truncation function $h(x) = x1_{|x|<1}$:

$$\begin{aligned} C(h) = \lambda\big(&\beta_+\big(\Gamma(-\alpha)\alpha + \Gamma(1-\alpha, \beta_+)\big) \\ &- \beta_-\big(\Gamma(-\alpha)\alpha + \Gamma(1-\alpha, \beta_-)\big)\big), \quad \alpha > 1, \end{aligned} \tag{14}$$

where $\Gamma(a, b) \equiv \int_b^\infty x^{a-1}e^{-x}\,\mathrm{d}x$ is the incomplete gamma function.

The dampened power law specification has two singular points at $\alpha = 0$ and $\alpha = 1$, under which the characteristic exponent takes different forms. The case of $\alpha = 0$ corresponds to the variance gamma model. Its characteristic exponent is,

$$\begin{aligned} \psi(u) = \lambda \ln(1 - iu/\beta_+)(1 + iu/\beta_-) = \lambda\big(&\ln(\beta_+ - iu) - \ln\beta \\ &+ \ln(\beta_- + iu) - \ln\beta_-\big). \end{aligned} \tag{15}$$

Since this process exhibits finite variation, we can perform the integral in (2) without the truncation function. When $\alpha = 1$, the characteristic exponent is (Wu, 2006),

$$\begin{aligned} \psi(u) = -\lambda\big(&(\beta_+ - iu)\ln(\beta_+ - iu)/\beta_+ \\ &+ \lambda(\beta_- + iu)\ln(\beta_- + iu)/\beta_-\big) - iuC(h), \end{aligned} \tag{16}$$

where the truncation-induced term is given by $C(h) = \lambda(\Gamma(0, \beta_+) - \Gamma(0, \beta_-))$ under the truncation function $h(x) = x1_{|x|<1}$.

Other popular pure jump Lévy processes include the normal inverse Gaussian (NIG) process (Barndorff-Nielsen, 1998), the generalized hyperbolic process (Eberlein et al., 1998), and the Meixner process (Schoutens, 2003). These processes all have tractable characteristic exponents.

2.3 *Empirical evidence*

Merton's (1976) compound Poisson jump specification is suitable to capture large and rare events such as market crashes and corporate defaults. Nevertheless, recent empirical evidence suggests that infinite-activity jump specifications that generate frequent jumps of all sizes are better suited to capture

the daily market movements of many financial securities such as stocks, stock indexes, and exchange rates. Furthermore, in reality the distinction between continuous and discontinuous market movements is not at all clear cut. Instead, we observe movements of all sizes, with small movements arriving more frequently than large movements. This type of behavior asks for a Lévy density that is monotone in the absolute jump magnitude. The dampened power law specification in (12) has this monotonic behavior. When the power coefficient $\alpha \geqslant 1$, the arrival rate of small jumps is so frequent that the specification generates sample paths with infinite variation, a property also shared by the Brownian motion. Hence, a Lévy process with infinite-variation jump provides a smooth transition from large jumps to small jumps and then to the continuous movements captured by a Brownian motion.

Several studies show that infinite-activity jumps perform better than finite-activity jumps in describing the statistical behavior of stock and stock index returns. Likelihood estimation of the dampened power law in Carr et al. (2002) on individual stocks and stock indexes shows that the estimates for the power coefficient α are mostly greater than zero. Li et al. (2007) use Markov Chain Monte Carlo method to estimate three Lévy specifications with stochastic time changes on the stock index. They find that infinite-activity jump specifications perform better in capturing the index behavior than finite-activity jumps do. Their simulation analysis further shows that an infinite-activity jump process cannot be adequately approximated by a finite-activity jump process, irrespective of the parameter choices.

Empirical studies using options show that using infinite-activity jumps also generate better option pricing performance. Carr and Wu (2003a) test the option pricing performance on the Merton jump-diffusion model, the variance gamma model, and their infinite-variation finite-moment log stable model. The pricing performance of the log stable model is the best among the three jump specifications. Huang and Wu (2004) apply various time changes on the three jump specifications to generate stochastic volatilities. They find that under all stochastic volatility specification, infinite-activity jumps perform significantly better than finite-activity jumps in pricing options.

Wu (2006) estimates the dampened power law using both the time-series returns and option prices on S&P 500 index. He obtains an estimate of the power coefficient at about 1.5. He also finds that although the exponential coefficient on down jumps β_- is large under the statistical measure, the estimate on its risk-neutral counterpart is not significantly different from zero. Without exponential dampening on down jumps, the return variance is infinite under the risk-neutral measure, even though it is finite under the statistical measure. As a result, the classic central limit theorem does not apply under the risk-neutral measure although it is applicable under the statistical measure. The difference under the two measures explains the empirical observation that the non-normality in the time-series index returns dissipates rapidly with time aggregation, but the risk-neutral return non-normality inferred from the options data persists to long option maturities.

When earlier studies use the compound Poisson jump to capture rare and large price movements, it is imperative to add a diffusion component to fill the gaps in between the arrival of the jumps. However, if we start with an infinite-activity jump that can generate an infinite number of small and large movements within any finite interval, it is not clear that we still need a diffusion component to fill the gaps. Carr et al. (2002) conclude from their empirical study that a diffusion component is no longer necessary as long as they adopt an infinite activity pure jump process. Carr and Wu (2003a) arrive at similar conclusions in their infinite variation log stable model. Huang and Wu (2004) find that a diffusion return component is useful in their time-changed Lévy process setting in generating correlations with the diffusive activity rate process. Nevertheless, it is not clear whether the diffusion return component is still needed if the activity rate also follows a pure jump process and correlations are constructed through synchronous jumps in return and the activity rate.

Carr and Wu (2003b) identify the presence of jump and diffusion components in the underlying asset price process by investigating the short-maturity behavior of at-the-money and out-of-the-money options written on this asset. They prove that a jump component, if present, dominates the short-maturity behavior of out-of-the-money options and hence can readily be identified. A diffusion component, if present, usually dominates the short-maturity behavior of at-the-money options. However, an infinite-variation jump component can generate short-maturity behavior for at-the-money options that are similar to those generated from a diffusion process. The similar behavior makes the identification of a diffusion component more difficult when an infinite-variation jump component is present.[2]

Aït-Sahalia (2004) proves in a simple Lévy setting that when a diffusion component is present, the diffusion variance can be effectively identified from discretely sampled time-series data using maximum likelihood method even in the presence of infinite-variation jumps, as long as the power coefficient of the jump component is not too close to 2. Aït-Sahalia and Jacod (2005) further show that the maximum likelihood method can also separately identify two jump components as long as the power coefficients for the jump components are sufficient apart from each other.

3 Generating stochastic volatility by applying stochastic time changes

It is well documented that asset return volatilities are stochastic (Engle, 2004). Recent evidence from the derivatives market suggests that higher return moments such as skewness also vary significantly over time (David and

[2] For pure jump α-stable Lévy processes with $\alpha \in [1, 2)$, at-the-money option prices converge to zero with declining maturity T at the rate of $O(T^{1/\alpha})$. The convergence rate is $O(T^{1/2})$ when there is a diffusion component. Hence, identifying the diffusion component becomes difficult when the power coefficient of the jump component is close to 2.

Veronesi, 1999; Johnson, 2002; and Carr and Wu, 2007a). A convenient approach to generating stochastic volatility on non-normal return innovations is to apply stochastic time changes to a Lévy process. A tractable way of generating stochastic skewness and other higher order moments is to apply separate time changes to multiple Lévy components with different degrees of skewness and higher order moments. The random time change amounts to stochastically altering the clock on which the Lévy process is run. Intuitively, a time change can be used to regulate the number of order arrivals that occur in a given time interval. More order arrivals generate higher return volatility (Ané and Geman, 2000). It can also be used to randomize the shocks from different economic sources. Separate time changes on different Lévy components can capture separate variations of different economic shocks.

3.1 Time changes and activity rates

Let X_t denote a Lévy process and let $t \to \mathcal{Z}_t (t \geqslant 0)$ be an increasing right-continuous process with left limits that satisfy the usual technical conditions, we can define a new process Y obtained by evaluating X at $\mathcal{Z}$, i.e.,

$$Y_t \equiv X_{\mathcal{Z}_t}, \quad t \geqslant 0. \tag{17}$$

Monroe (1978) proves that every semimartingale can be written as a time-changed Brownian motion. Hence, equation in (17) is a very general specification. In principle, the random time $\mathcal{Z}_t$ can be modeled as a nondecreasing semimartingale,

$$\mathcal{Z}_t = \mathcal{T}_t + \int_0^t \int_0^\infty x\mu(\mathrm{d}t, \mathrm{d}x), \tag{18}$$

where $\mathcal{T}_t$ is the locally deterministic and continuous component and $\mu(\mathrm{d}t, \mathrm{d}y)$ denotes the counting measure of the possible positive jumps of the semimartingale. The two components can be used to play different roles. Applying a time change defined by the positive jump component $\int_0^t \int_0^\infty x\mu(\mathrm{d}t, \mathrm{d}x)$ to a Brownian motion generates a new discontinuous process. If we model the positive jump component by a Lévy process, it is often referred to as a *Lévy subordinator*. A Lévy process subordinated by a Lévy subordinator yields a new Lévy process (Sato, 1999). Therefore, this component can be used to randomize the original return innovation defined by X to generate a refined return innovation distribution. For example, Madan and Seneta (1990) generate the variance-gamma pure jump Lévy process by applying a gamma time change to a Brownian motion.

To generate stochastic volatility on non-normal return innovations, I start directly with a Lévy process that *already* captures the non-normal return innovation distribution, and then apply a locally deterministic time change $\mathcal{T}_t$ purely for the purpose of generating stochastic volatilities and stochastic higher return

moments. We can characterize the locally deterministic time change in terms of its local intensity $v(t)$:

$$\mathcal{T}_t = \int_0^t v(u_-)\,\mathrm{d}u,$$

Carr and Wu (2004) label $v(t)$ as the *instantaneous activity rate*, with $v(u_-)$ denoting the activity rate at time u just prior to a jump. When X_t is a standard Brownian motion, v_t becomes the instantaneous variance of the Brownian motion. When X_t is a pure jump Lévy process, such as the compound Poisson jump process of Merton (1976), $v(t)$ is proportional to the jump arrival rate.

Although $\mathcal{T}_t$ is locally deterministic and continuous, the instantaneous activity rate process $v(t)$ can be fully stochastic and can jump. Given any continuous or discontinuous dynamics for $v(t)$, the integration over its sample path makes $\mathcal{T}_t$ locally predictable and continuous. Nevertheless, for $\mathcal{T}_t$ to be nondecreasing, the activity rate needs to be nonnegative, a natural requirement for diffusion variance and jump arrival rates.

3.2 Generating stochastic volatility from different economic sources

By applying stochastic time changes to Lévy processes, it becomes obvious that stochastic volatility can come from multiple sources. It can come from the instantaneous variance of a diffusion return component, or the arrival rate of a jump component, or both. Huang and Wu (2004) design and estimate a class of models for S&P 500 index returns based on the time-changed Lévy process framework. They allow the return innovation to contain both a diffusion component and a jump component. Then, they consider several cases where they apply stochastic time changes to

(1) the diffusion component only (SV1),
(2) the jump component only (SV2),
(3) both components with one joint activity rate (SV3), and
(4) both components with separate activity rates for each component (SV4).

They find that by allowing the diffusion variance rate and the jump arrival rate to follow separate dynamic processes, the SV4 specification outperforms all the other single activity rate specifications in pricing the index options.

Applying separate stochastic time changes to different Lévy components also proves to be a tractable way of generating stochastic higher return moments such as skewness. In the SV4 specification of Huang and Wu (2004), one activity rate controls the intensity of a diffusion and hence a normal innovation component and the other activity rate controls the intensity of a negatively skewed pure jump component. The variation of the two activity rates over time generates variation in the relative proportion of the diffusion versus

the negatively-skewed jump return innovation component. As a result, the degree of the negative skewness for the index return varies over time (David and Veronesi, 1999).

Carr and Wu (2005) apply the time-changed Lévy process framework to jointly price stock options and credit default swaps written on the same company. They assume that corporate default arrives via a Poisson process with stochastic arrival rate. Upon default, the stock price jumps to zero. Prior to default, the stock price follows a purely continuous process with stochastic volatility. Hence, the model decomposes the stock return into two Lévy components:

(i) the continuous component that captures the market risk, and
(ii) the jump component that captures the impact of credit risk.

Separate time changes on the two components generate stochastic volatility for market movements and stochastic arrival for corporate default, respectively. Carr and Wu (2007b) use a similar specification to capture the correlation between sovereign credit default swap spreads and currency options. They assume that sovereign default induces a negative but random jump in the price of the home currency.

For stock indexes and the dollar (or euro) prices of emerging market currencies, the risk-neutral return distribution skewness is time-varying, but the sign stays negative across most of the sample period.[3] In contrast, for the exchange rates between two relatively symmetric economies, Carr and Wu (2007a) find that the risk-neutral currency return distribution inferred from option prices shows skewness that not only varies significantly over time in magnitudes, but also switches signs. To capture the stochastic skewness with possible sign switches, they decompose the currency return into two Lévy components that generate positive and negative skewness, respectively. Then, they apply separate stochastic time changes to the two Lévy components so that the relative proportion of the two components and hence the relative degree and direction of the return skewness can vary over time. They model the positively-skewed Lévy process with a jump component that only jumps upward and the negatively-skewed Lévy process with a jump component that only jumps downward. Furthermore, each process contains a diffusion component that is correlated with their respective activity rate process. The correlation is positive for the positive-skewed Lévy process and negative for the negative-skewed Lévy component. Thus, the up and down jumps generate short-term positive and negative skewness for the two Lévy components, and the different correlations between the two Lévy components and their respective activity rates generate long-term skewness.

[3] See the evidence in David and Veronesi (1999) and Foresi and Wu (2005) on stock index options and Carr and Wu (2007b) on currency options.

In contrast to modeling returns by a set of hidden factors, our modeling approach of applying stochastic time changes to different Lévy processes makes explicit the purpose of each modeling component. Under this framework, we use different Lévy processes as building blocks representing different economic forces. Applying stochastic time changes on each component randomizes the intensity of the impact from each economic force. The clear economic mapping makes the model design more intuitive and concise. Each component is added for a specific economic purpose. Using this approach is more likely to create models that are parsimonious and yet capable of delivering the target properties.

3.3 Theory and evidence on activity rate dynamics

Exploiting information in variance swap rates and various realized variance estimators constructed from high-frequency returns, Wu (2005) empirically studies the activity rate dynamics for the S&P 500 index returns under a generalized affine framework. He finds that the activity rate for the index return contains an infinite-activity jump component, with its arrival rate proportional to the activity rate level. The Markov Chain Monte Carlo estimation in Eraker et al. (2003) on long histories of index returns also suggests the presence of a jump component in the activity rate dynamics.

The impact of a jump component in the activity rate dynamics is usually small on the pricing of stock (index) options (Broadie et al., 2002) and the term structure of variance swaps (Wu, 2005). Hence, many specifications for option pricing assume pure continuous activity rate dynamics for parsimony. Nevertheless, jumps are an integral part of the statistical variance dynamics. Furthermore, their pricing impacts can become more significant for derivative contracts that are sensitive to the tails of the variance distribution, e.g., options on variance swaps or realized variance.

When separate time changes are applied to different innovation components, the underlying activity rates can be modeled independently or with dynamic interactions. For example, Carr and Wu (2007a) assume that the two activity rates that govern the positive and negative Lévy components are independent of each other. Independent assumptions are also applied in the SV4 specification in Huang and Wu (2004). In contrast, Carr and Wu (2005) find that stock return volatilities and corporate default arrival intensities co-move with each other. To capture the co-movements, they model the joint dynamics of the stock return diffusion variance rate v_t and the default arrival rate λ_t as

$$\begin{aligned} \mathrm{d}v_t &= (u_v - \kappa_v v_t)\,\mathrm{d}t + \sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v, \\ \lambda_t &= \xi v_t + z_t, \\ \mathrm{d}z_t &= (u_z - \kappa_z z_t - \kappa_{vz} v_t)\,\mathrm{d}t + \sigma_z \sqrt{z_t}\,\mathrm{d}W_t^z, \end{aligned} \tag{19}$$

where W_t^v and W_t^z denote two independent Brownian motions. The interactions between the diffusion variance and default arrival are captured by both

the contemporaneous loading coefficient ξ and the dynamic predictive coefficient κ_{vz}.

When the purpose is to capture the option price behavior at a narrow range of maturities, a one-factor activity rate specification is often adequate in generating stochastic volatilities. However, if the purpose is to capture the term structure of at-the-money implied volatilities or variance swap rates across a wide range of maturities, a one-factor activity rate process is often found inadequate. In most options markets, the persistence of the implied volatilities increases with the option maturity. This feature calls for multi-factor activity rate dynamics with different degrees of persistence for the different factors. One example is to allow the activity rate to revert to a stochastic mean level:

$$\begin{aligned} \mathrm{d}v_t &= \kappa(m_t - \kappa_v v_t)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v, \\ \mathrm{d}m_t &= \kappa_m(\theta - m_t)\,\mathrm{d}t + \sigma_m\sqrt{m_t}\,\mathrm{d}W_t^m, \end{aligned}$$

where the mean-reversion speed of m, κ_m, is usually much smaller than the mean-reversion speed of the activity rate itself κ_v. Balduzzi et al. (1998) use a similar specification for the instantaneous interest rate dynamics and label m as the stochastic central tendency factor. Intuitively, the activity rate $v(t)$ affects short-term option implied volatilities more heavily whereas the central tendency factor m_t dominates the variation of long-term options. Thus, the persistence of the option implied volatility or variance swap rate can increase with the option maturities. Carr and Wu (2007a) consider a similar extension to their stochastic skew model, where the activity rates of both the positive and the negative Lévy components are allowed to revert to a common stochastic central tendency factor. Their estimation shows that the extension significantly improves the option pricing performance along the maturity dimension. Carr and Wu (2007b) also consider a similar extension on the default arrival dynamics to better capture the term structure of credit default swap spreads.

Most applications in option pricing use affine specifications for the activity rate dynamics, under which the activity rate is an affine function of a set of state variables and both the drift and variance of the state variables are affine in the state variables. When upward jumps are allowed in these state variables, their arrival rate are also affine in the state variables. Carr and Wu (2004) show that both affine and quadratic specifications can be used to model the activity rate while retaining the analytical tractability for option pricing. Santa-Clara and Yan (2005) estimate a model with quadratic activity rates on S&P 500 index options. In their model, the return innovation consists of both a diffusion component and a compound Poisson jump component, and each component is time changed separately, with the underlying activity rate being a quadratic function of an Ornstein–Ulenbeck process. They show that they can incorporate more intricate correlation structures under their quadratic specification than under the affine specification while maintaining tractability.

Lewis (2000) and Heston (1997) show that option pricing is also reasonably tractable when the activity rate is governed by the 3/2 dynamics:

$$\mathrm{d}v_t = \kappa v_t(\theta - v_t)\,\mathrm{d}t + \sigma_v v_t^{3/2}\,\mathrm{d}W_t. \tag{20}$$

Carr and Sun (2005) show that under a pure diffusion model for the asset return with a 3/2 variance rate dynamics, European option values can be written as a function of the asset price level and the level of the variance swap rate of the same maturity, with no separate dependence on calendar time or time-to-maturity. Furthermore, the pricing function depends only on the volatility of volatility coefficient σ_v, but not on the drift parameters (θ, κ). Therefore, if we observe the underlying asset's price and its variance swap rate quotes, we can price options with merely one model parameter σ_v, without the need to estimate the drift function of the variance rate dynamics.

Within the one-factor diffusion context, several empirical studies find that a 3/2 specification on the variance rate dynamics performs better than the square-root specification. Favorable evidence based on time-series returns includes Chacko and Viceira (2003), Ishida and Engle (2002), Javaheri (2005), and Jones (2003). Jones (2003), Medvedev, Scaillet (2003), and Bakshi et al. (2006) also find supporting evidence for the 3/2 specification from equity index options implied volatilities.

4 Modeling financial security returns with time-changed Lévy processes

Once we have a clear understanding on the different roles played by Lévy innovations and random time changes, we can assemble the pieces together and write a complete model for the financial security return. The traditional literature often starts with the specification of the return dynamics under the statistical measure $\mathbb{P}$, and derive the return dynamics under the risk-neutral measure $\mathbb{Q}$ for option pricing based on market price of risk specifications. However, since the requirement for analytical tractability mainly comes from the expectation operation under the risk-neutral measure in pricing contingent claims, it is often convenient to start directly with a tractable risk-neutral dynamics. Then, since we do not have as much concern for the tractability of the statistical dynamics, we can accommodate very flexible market price of risk specifications, with the only practical constraints coming from reasonability and identification considerations.

4.1 Constructing risk-neutral return dynamics

Let S_t denote the time-t price of a financial security. Let $\{X^k_{\mathcal{T}^k_t}\}_{=1}^K$ denote a series of independent time-changed Lévy processes, which are specified under a risk-neutral measure $\mathbb{Q}$. We use these processes as building blocks for the return dynamics. The independence assumption between different components

is for convenience only, although interactions can be added when necessary as in Eq. (19). We model the risk-neutral return dynamics over the time period $[0, t]$ as

$$\ln S_t/S_0 = (r-q)t + \sum_{k=1}^{K}\left(b^k X^k_{\mathcal{T}^k_t} - \varphi_{x^k}\left(b^k\right)\mathcal{T}^k_t\right), \tag{21}$$

where r denotes the instantaneous interest rate, q denotes the dividend yield for stocks or the foreign instantaneous interest rate for currencies, and b^k denotes a constant loading coefficient on the kth component. For notational clarity, I assume both r and q constant. If we allow both to be stochastic, the first term should be replaced by an integral $\int_0^t (r(u)-q(u))\,\mathrm{d}u$. If they vary deterministically over time, we can also replace the integral with the continuously compounded yields over the horizon $[0, t]$.

Equation (21) models the risks in the asset return using K components of time-changed Lévy processes. The cumulant exponent $\varphi_{x^k}(b^k)$ represents a concavity adjustment so that the return dynamics satisfy the martingale condition under the risk-neutral measure:

$$\mathbb{E}^{\mathbb{Q}}_0[S_t/S_0] = e^{(r-q)t}. \tag{22}$$

Since by definition

$$\mathbb{E}^{\mathbb{Q}}_0\left[e^{b^k X_t}\right] = e^{\varphi_{x^k}(b^k)t}, \tag{23}$$

the following expectation is a martingale:

$$\mathbb{E}^{\mathbb{Q}}_0\left[e^{b^k X_t - \varphi_{x^k}(b^k)t}\right] = 1. \tag{24}$$

The martingale condition retains when we replace t with a locally predictable and continuous time change $\mathcal{T}_t$ (Küchler and Sørensen, 1997):

$$\mathbb{E}^{\mathbb{Q}}_0\left[e^{b^k X_{\mathcal{T}^k_t} - \varphi_{x^k}(b^k)\mathcal{T}^k_t}\right] = 1. \tag{25}$$

Thus, we have

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}_0[S_t/S_0] &= \mathbb{E}^{\mathbb{Q}}_0\left[e^{(r-q)t+\sum_{k=1}^{K}\left(b^k X^k_{\mathcal{T}^k_t} - \varphi_{x^k}(b^k)\mathcal{T}^k_t\right)}\right] \\ &= e^{(r-q)t}\prod_{k=1}^{K}\mathbb{E}^{\mathbb{Q}}_0\left[e^{b^k X^k_{\mathcal{T}^k_t} - \varphi_{x^k}(b^k)\mathcal{T}^k_t}\right] = e^{(r-q)t}.\end{aligned} \tag{26}$$

The independence assumption between different Lévy components enables us to move the expectation operation inside the product.

Each Lévy process X^k_t can have a drift component of its own, but it is irrelevant in our return specification (21) because any drift will be canceled out with

a corresponding term in the concavity adjustment. Hence, for each Lévy component, we only specify the diffusion volatility σ if the security price is allowed to move continuously and the Lévy density $\pi(x)$ if the price is allowed to jump.

The above return dynamics are defined over the horizon $[0, t]$ with time 0 referring to today and t being some future time corresponding to the maturity date of the contingent claim being valued. The time change $\mathcal{T}_t$ represents the integral of the activity rates over the same time period $[0, t]$. Sometimes it is more convenient to use t to denote the current date and T for future date, with $\tau = T - t$ denoting the time to maturity of the contingent claim. Then, the time change can be defined accordingly between this time period,

$$\mathcal{T}_{t,T} \equiv \int_t^T v_-(u)\,\mathrm{d}u. \tag{27}$$

The log return between $[t, T]$ can be written as

$$\ln S_T/S_t = (r-q)(T-t) + \sum_{k=1}^{K}\bigl(b^k X^k_{\mathcal{T}^k_{t,T}} - \varphi_{x^k}\bigl(b^k\bigr)\mathcal{T}^k_{t,T}\bigr). \tag{28}$$

To illustrate the construction of the risk-neutral dynamics, I start with the simplest case where the return innovation is driven by one diffusion component without time change: $X^1_t = \sigma W_t$, $\mathcal{T}_t = t$, $K = 1$, $b^1 = 1$, with W_t denoting a standard Brownian motion. The return process becomes

$$\ln S_t/S_0 = (r-q)t + \sigma W_t - \frac{1}{2}\sigma^2 t. \tag{29}$$

The cumulant exponent of σW_t evaluated at $s = 1$ is $\varphi(1) = \frac{1}{2}\sigma^2$. Equation (29) is essentially the classic Black and Scholes (1973) model.

Applying random time change to the diffusion component, we have

$$\ln S_t/S_0 = (r-q)t + \sigma W_{\mathcal{T}_t} - \frac{1}{2}\sigma^2\mathcal{T}_t, \tag{30}$$

where we simply replace t with $\mathcal{T}_t$ on terms related to the Lévy component. If we model the activity rate v_t underlying the time change by the square-root process of Cox et al. (1985), we will generate the stochastic volatility model of Heston (1993):

$$\mathrm{d}v_t = \kappa(1 - v_t)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W^v_t. \tag{31}$$

The long-run mean of the activity rate is normalized to one for identification purpose, since we already have a free volatility parameter σ in (30) that captures the mean level of volatility. In the original Heston model, σ is normalized to one and the long-run mean of the activity rate is left as a free parameter. To match the activity rate specification with the time change notation, we can

rewrite the activity rate in integral forms

$$\begin{aligned} v_t &= v_0 + \int_0^t \kappa(1 - v_t)\,\mathrm{d}t + \int_0^t \sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v \\ &= v_0 + \kappa t - \kappa \mathcal{T}_t + \sigma_v W_{\mathcal{T}_t}^v. \end{aligned} \tag{32}$$

Technically, the second equality in (32) holds only *in distribution* and the W^v in $\int_0^t \sqrt{v_t}\,\mathrm{d}W_t^v$ denotes a different Brownian motion from the W^v in $W_{\mathcal{T}_t}^v$. Hence, a more technically correct way of writing the equality is:

$$\int_0^t \sqrt{v_t}\,\mathrm{d}W_t^v =^d \widetilde{W}_{\mathcal{T}_t}^v, \tag{33}$$

where $=^d$ denotes "equality in distribution," and $(W^v, \widetilde{W}^v)$ denote two different Brownian motions. To avoid notation clustering, we use W^v to represent two different Brownian motions in the two different representations. We also use the same equality sign "=" to represent both the traditional mathematical equality and the equality in distribution. Analogously, the equalities between $\int_0^t \sqrt{v_t}\,\mathrm{d}W_t$ and $W_{\mathcal{T}_t}$, between $\sqrt{v_t}\,\mathrm{d}W_t$ and $\mathrm{d}W_{\mathcal{T}_t}$, and between $\sqrt{v_t}\,\mathrm{d}W_t^v$ and $\mathrm{d}W_{\mathcal{T}_t}^v$ are all in distribution, and the two "W"s in each pair represent two different Brownian motions. Heston (1993) allows correlation between the activity rate innovation and the return innovation $\mathbb{E}[\mathrm{d}W_t\,\mathrm{d}W_t^v] = \rho\,\mathrm{d}t$, or equivalently under the time-change notation, $\mathbb{E}[\mathrm{d}W_{\mathcal{T}_t}\,\mathrm{d}W_{\mathcal{T}_t}^v] = \rho\,\mathrm{d}\mathcal{T}_t = \rho v_t\,\mathrm{d}t$.

Technicality aside, I regard the time-change notation as simply a convenient way of rewriting the traditional stochastic differential equation. Using the Heston (1993) model as an example. The traditional representation in terms of the stochastic differential equation is:

$$\begin{aligned} \mathrm{d}\ln S_t &= (r - q)\,\mathrm{d}t + \sigma\sqrt{v_t}\,\mathrm{d}W_t - \frac{1}{2}\sigma^2 v_t\,\mathrm{d}t, \\ \mathrm{d}v_t &= \kappa(1 - v_t)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v. \end{aligned} \tag{34}$$

The following time-changed Lévy process generates the same return distribution:

$$\begin{aligned} \ln S_t/S_0 &= (r - q)t + \sigma W_{\mathcal{T}_t} - \frac{1}{2}\sigma^2 \mathcal{T}_t, \\ v_t &= v_0 + \kappa t - \kappa \mathcal{T}_t + \sigma_v W_{\mathcal{T}_t}^v, \end{aligned} \tag{35}$$

with the technical caveat that (34) and (35) represent different processes and (W, W^v) in the two sets of equations represent completely different Brownian motions.

Now consider an example where the return innovation is driven by a pure jump Lévy component without time change and the jump arrival is governed

by the dampened power law specification in (12) with $\alpha \neq 0$ and $\alpha \neq 1$. The return dynamics can be written as

$$\ln S_t/S_0 = (r-q)t + J_t - \varphi_J(1)t, \tag{36}$$

where J_t denotes this Lévy jump component, and the cumulant exponent is

$$\varphi_J(s) = \Gamma(-\alpha)\lambda\left[(\beta_+ - s)^\alpha - \beta_+^\alpha + (\beta_- + s)^\alpha - \beta_-^\alpha\right] + sC(h), \tag{37}$$

with $C(h)$ given in (14). Since any linear drift terms in J_t will be canceled out by the corresponding term in the concavity adjustment, it becomes obvious that the exact form of the truncation function and the resultant linear coefficient $C(h)$ are immaterial for modeling and estimation. Given the cumulant exponent in (13), the concavity adjustment in Eq. (36) becomes

$$\varphi_J(1) = \Gamma(-\alpha)\lambda\left[(\beta_+ - 1)^\alpha - \beta_+^\alpha + (\beta_- + 1)^\alpha - \beta_-^\alpha\right] + C(h). \tag{38}$$

If we apply random time change to the pure jump Lévy component, we can simply replace J_t with $J_{\mathcal{T}_t}$ and $\varphi_J(1)t$ with $\varphi_J(1)\mathcal{T}_t$:

$$\ln S_t/S_0 = (r-q)t + J_{\mathcal{T}_t} - \varphi_J(1)\mathcal{T}_t, \tag{39}$$

which is a pure jump process with stochastic volatility generated purely from the stochastic arrival of jumps.

When we use one Lévy component in the return dynamics, it is natural to set the loading coefficient b to unity as it can always be absorbed into the scaling specification of the Lévy process. To show an example where the loading coefficient plays a more explicit role, we consider a market model for stock returns, where the return on each stock is decomposed into two orthogonal components: a market risk component and an idiosyncratic risk component. We use a Lévy process X_t^m to model the market risk and another Lévy process X_t^j to model the idiosyncratic risk for stock j. Then, the return on stock j can be written under the risk-neutral measure $\mathbb{Q}$ as

$$\ln S_t^j/S_0^j = (r-q)t + \left(b^j X_t^m - \varphi_{x^m}\left(b^j\right)t\right) + \left(X_t^j - \varphi_{x^j}(1)t\right), \tag{40}$$

where the first component $(r-q)t$ captures the instantaneous drift under the risk-neutral measure, the second component $(b^j X_t^m - \varphi_{x^m}(b^j)t)$ represents the concavity-adjusted market risk component, with b^j capturing the linear loading of the return on the market risk factor X_t^m, and the last component $(X_t^j - \varphi_{x^j}(1)t)$ is the concavity-adjusted idiosyncratic risk component for the stock return.

Under the Lévy specification in (40), stock returns are i.i.d. We can apply random time changes to the two Lévy processes to generate stochastic volatility:

$$\ln S_t^j/S_0^j = (r-q)t + \left(b^j X_{\mathcal{T}_t^m}^m - \varphi_{x^m}\left(b^j\right)\mathcal{T}_t^m\right) + \left(X_{\mathcal{T}_t^j}^j - \varphi_{x^j}(1)\mathcal{T}_t^j\right), \tag{41}$$

where stochastic volatility can come either from the market risk via $\mathcal{T}_t^m$ or from the idiosyncratic risk via $\mathcal{T}_t^j$. Mo and Wu (2007) propose an international capital asset pricing model with a structure similar to Eq. (41), where X^m represents a global risk factor and X^j a country-specific risk factor. They specify the dynamics under both the risk-neutral measure and the statistical measure, and estimate the joint dynamics of three economies (US, UK, and Japan) using the time-series returns and option prices on the S&P 500 index, the FTSE 100 Index, and the Nikkei-225 Stock Average.

4.2 Market price of risks and statistical dynamics

Once we have specified the return dynamics under the risk-neutral measure $\mathbb{Q}$, we can derive the dynamics under the statistical measure $\mathbb{P}$ if we know whether and how different sources of risks are priced. Take the generic return specification in (21) as an example. We have K sources of return risks as captured by the K Lévy processes $\{X_t^k\}_{k=1}^K$. We also have K sources of volatility risks corresponding to each return component. Furthermore, each Lévy process X_t^k can have a diffusion component and a jump component. The two components can be priced differently. Upside and downside jumps can also be priced differently (Wu, 2006 and Bakshi and Wu, 2005). The activity rate that underlies each time change can also have a diffusion and a jump component that can be priced differently. Depending on the market price of risk specification, the statistical return dynamics can look dramatically different from the risk-neutral dynamics.

In this subsection, we consider a simple class of market price of risk specifications, which in most cases generates statistical return dynamics that stay in the same class as the risk-neutral dynamics. The pricing kernel that defines the market price of all sources of risks can be written as

$$\begin{aligned}\mathcal{M}_t = e^{-rt}\prod_{k=1}^{K}\exp\big(&-\gamma_k X^k_{\mathcal{T}_t^k} - \varphi_{x^k}(-\gamma_k) - \gamma_{kv} X^{kv}_{\mathcal{T}_t^k}\\ &- \varphi_{x^{kv}}(-\gamma_{kv})\big)\cdot\zeta,\end{aligned} \tag{42}$$

where $X^k_{\mathcal{T}_t^k}$ denotes the return risk as in (21), $X^{kv}_{\mathcal{T}_t^k}$ denotes another set of time-changed Lévy processes that characterize the activity rate risk, and ζ denotes an orthogonal martingale component that prices other sources of risks independent of the security return under consideration. We maintain the constant interest rate assumption in the pricing kernel specification. The exponential martingale components in the pricing kernel determines the measure change from $\mathbb{P}$ to $\mathbb{Q}$. The simplicity of the specification comes from the constant assumption on the market price coefficients γ_k and γ_{kv}.

Given the pricing kernel in (42) and the risk-neutral return dynamics in (21), we can infer the statistical return dynamics. We use examples to illustrate the

procedure, starting with the simplest case where the return is driven by one diffusion component without time change as in (29). According to the above exponential martingale assumption, the measure change from $\mathbb{P}$ to $\mathbb{Q}$ is defined by

$$\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_t = \exp\left(-\gamma\sigma W_t - \frac{1}{2}\gamma^2\sigma^2 t\right) \tag{43}$$

with $\varphi_{\sigma W}(-\gamma) = \frac{1}{2}\gamma^2\sigma^2$. The literature has taken different approaches in arriving at the dynamics under a measure change. For measure changes defined by exponential martingales of a Lévy processes X, it is convenient to remember that $\varphi_X^{\mathbb{P}}(s) = \varphi_X^{\mathbb{Q}}(s+\gamma) - \varphi_X^{\mathbb{Q}}(\gamma)$ and that the drift adjustment of X is captured by $\eta = \varphi^{\mathbb{P}}(1) - \varphi^{\mathbb{Q}}(1)$ (Küchler and Sørensen, 1997). For the simple case in (43) with $X = \sigma W$, we have $\varphi_{\sigma W}^{\mathbb{Q}}(1) = \frac{1}{2}\sigma^2$, and

$$\begin{aligned}\varphi_{\sigma W}^{\mathbb{P}}(1) &= \varphi_{\sigma W}^{\mathbb{Q}}(1+\gamma) - \varphi_{\sigma W}^{\mathbb{Q}}(\gamma) = \frac{1}{2}(1+\gamma)^2\sigma^2 - \frac{1}{2}\gamma^2\sigma^2 \\ &= \frac{1}{2}\sigma^2 + \gamma\sigma^2. \end{aligned} \tag{44}$$

Thus, the drift adjustment, or the instantaneous expected excess return, is $\eta = \gamma\sigma^2$. The statistical ($\mathbb{P}$) return dynamics becomes[4]

$$\ln S_t/S_0 = (r-q)t + \gamma\sigma^2 t + \sigma W_t - \frac{1}{2}\sigma^2 t. \tag{45}$$

For the Heston (1993) model, which has the risk-neutral dynamics specified in (34) or equivalently (35), the associated exponential martingale that defines the measure change from $\mathbb{P}$ and $\mathbb{Q}$ becomes

$$\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_t = \exp\left(-\gamma\sigma W_{\mathcal{T}_t} - \frac{1}{2}\gamma^2\sigma^2\mathcal{T}_t - \gamma_v\sigma_v W^v_{\mathcal{T}_t} - \frac{1}{2}\gamma_v^2\sigma_v^2\mathcal{T}_t\right). \tag{46}$$

The cumulant exponent of the return innovation σW_t under measure $\mathbb{P}$ becomes $\varphi_{\sigma W}^{\mathbb{P}}(s) = \varphi_{\sigma W}^{\mathbb{Q}}(s+\gamma+\gamma_v\sigma_v\rho/\sigma) - \varphi_{\sigma W}^{\mathbb{Q}}(\gamma+\gamma_v\sigma_v\rho/\sigma)$. Hence, the drift adjustment induced by the measure change is $\eta = \gamma\sigma^2 + \gamma_v\sigma_v\sigma\rho$. The first term is induced by the pricing of the return risk W and the second term is induced by the pricing of the part of volatility risk W^v that is correlated with the return risk. Given the stochastic time change and hence stochastic activity rate, the risk premium over the horizon $[0, t]$ is $\eta\mathcal{T}_t$, and the instantaneous risk

[4] To be technically correct, we should also differentiate between $W_t^{\mathbb{P}}$ and $W_t^{\mathbb{Q}}$. Under our constant market price of risk specifications, we have: $\sigma W_t^{\mathbb{Q}} = \sigma W_t^{\mathbb{P}} + \gamma\sigma^2 t$. To maintain notational clarity, we use the same W notation without the superscript to represent a standard Brownian motion under all measures as no confusion shall occur.

premium at time t is ηv_t. The statistical return dynamics becomes

$$\ln S_t/S_0 = (r-q)t + \big(\gamma\sigma^2 + \gamma_v\sigma_v\sigma\rho\big)\mathcal{T}_t + \sigma W_{\mathcal{T}_t} - \frac{1}{2}\sigma^2\mathcal{T}_t. \tag{47}$$

To derive the statistical dynamics for the activity rate, we note that the cumulant exponent of the activity rate innovation $\sigma_v W^v$ under measure $\mathbb{P}$ becomes $\varphi^{\mathbb{P}}_{\sigma_v W^v}(s) = \varphi^{\mathbb{Q}}_{\sigma_v W^v}(s+\gamma_v+\gamma\sigma\rho/\sigma_v) - \varphi^{\mathbb{Q}}_{\sigma_v W^v}(\gamma_v+\gamma\sigma\rho/\sigma_v)$. Hence, the measure change induces an instantaneous drift change captured by $\eta^v = \varphi^{\mathbb{P}}_{\sigma_v W^v}(1) - \varphi^{\mathbb{Q}}_{\sigma_v W^v}(1) = \gamma_v\sigma_v^2 + \gamma\sigma\sigma_v\rho$, where the first term is induced by the pricing of the activity rate innovation W^v and the second term is induced by the pricing of the part of return risk W that is correlated with the activity rate. Since we apply the same time change $\mathcal{T}_t$ to the two sources of risks W and W^v, the actual drift adjustment over calendar time $[0, t]$ becomes $\eta^v\mathcal{T}_t$, and the instantaneous adjustment is $\eta^v v_t$. The statistical activity rate dynamics becomes

$$v_t = v_0 + at - \kappa\mathcal{T}_t + \eta^v\mathcal{T}_t + \sigma_v W^v_{\mathcal{T}_t}, \tag{48}$$

or in the form of the stochastic differential equation,

$$\mathrm{d}v_t = \big(a - \big(\kappa - \eta^v\big)v_t\big)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W^v_t, \tag{49}$$

where the measure change induces a change in the mean reversion speed from κ under $\mathbb{Q}$ to $\kappa^{\mathbb{P}} = \kappa - \eta^v = \kappa - \gamma_v\sigma_v^2 - \gamma\sigma\sigma_v\rho$ under $\mathbb{P}$. Estimation on stock indexes and stock index options often find that the market price of return risk (γ) is positive and the market price of variance risk (γ_v) is negative. Given the well-documented negative correlation (ρ) between the return and variance innovations, both sources of market prices make the activity rate more persistent under the risk-neutral measure than the activity rate is under the statistical measure: $\kappa < \kappa^{\mathbb{P}}$.[5]

For the pure jump Lévy process example as in (36), the measure change from $\mathbb{P}$ to $\mathbb{Q}$ is defined by the exponential martingale,

$$\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_t = \exp\big(-\gamma J_t - \varphi_J(-\gamma)t\big). \tag{50}$$

The Lévy density under the two measures are linked by $\pi^{\mathbb{P}}(x) = e^{\gamma x}\pi^{\mathbb{Q}}(x)$. If the Lévy density under $\mathbb{Q}$ is given by Eq. (12), its corresponding Lévy density under $\mathbb{P}$ becomes

$$\pi^{\mathbb{P}}(x) = \begin{cases} \lambda\exp(-(\beta_+ - \gamma)x)x^{-\alpha-1}, & x > 0, \\ \lambda\exp(-(\beta_- + \gamma)|x|)|x|^{-\alpha-1}, & x < 0. \end{cases} \tag{51}$$

[5] Since the constant part of drift remains the same as a, the long-run mean of the activity rate changes from a/κ under $\mathbb{Q}$ to $a/(\kappa-\eta^v)$ under $\mathbb{P}$. The smaller mean reversion under $\mathbb{Q}$ implies a higher long-run mean.

Therefore, the Lévy density is still controlled by a dampened power law under the statistical measure $\mathbb{P}$, only with the exponential dampening coefficients changed from (β_+, β_-) under $\mathbb{Q}$ to $\beta_+^{\mathbb{P}} = \beta_+ - \gamma$ and $\beta_-^{\mathbb{P}} = \beta_- + \gamma$ under $\mathbb{P}$. The dampening coefficients should be nonnegative under both measures. This condition limits the range of values that the market price of risk γ can take. Given the risk-neutral dampening coefficients (β_+, β_-), we need $\gamma \in [-\beta_-, \beta_+]$. Given the statistical coefficients $(\beta_+^{\mathbb{P}}, \beta_-^{\mathbb{P}})$, we need $\gamma \in [-\beta_+^{\mathbb{P}}, \beta_-^{\mathbb{P}}]$.

Wu (2006) and Bakshi and Wu (2005) allow the downside and upside jumps to have different market prices (γ_+, γ_-). In this case, we can directly specify the dampening coefficients under the two measures (β_+, β_-) and $(\beta_+^{\mathbb{P}}, \beta_-^{\mathbb{P}})$ as free parameters with positivity constraints. Then, the market prices of positive and negative jump risks can be derived as $\gamma_+ = \beta_+ - \beta_+^{\mathbb{P}}$ and $\gamma_- = \beta_-^{\mathbb{P}} - \beta_-$. By estimating this pure jump Lévy model to S&P 500 index time-series returns and option prices, Wu finds that there is zero dampening on downside jumps under the risk-neutral measure ($\beta_- = 0$). Thus, the market price of downside jump risk reaches its upper limit at $\gamma_- = \beta_-^{\mathbb{P}}$. This extremely high market price of downside risk is needed to capture the much higher prices for out-of-the-money put options than for the corresponding out-of-the-money call options on the index and the corresponding implied volatility smirk at both short and long maturities.

Given the measure change defined in (50), the cumulant exponent under measure $\mathbb{P}$ is linked to the cumulant exponent under measure $\mathbb{Q}$ by $\varphi^{\mathbb{P}}(s) = \varphi^{\mathbb{Q}}(s+\gamma) - \varphi^{\mathbb{Q}}(\gamma)$. The instantaneous expected excess return is given by $\eta = \varphi_J^{\mathbb{P}}(1) - \varphi_J^{\mathbb{Q}}(1) = \varphi_J^{\mathbb{Q}}(1+\gamma) - \varphi_J^{\mathbb{Q}}(\gamma) - \varphi_J^{\mathbb{Q}}(1)$. It is obvious that any term in the cumulant exponent $\varphi_J^{\mathbb{Q}}(s)$ that is linear in s does not contribute to the expected excess return η. Hence, the truncation-induced linear term $sC(h)$, or the choice of the truncation function $h(x)$, does not affect the computation of the expected excess return η.

Under the jump specification in (12) and when $\alpha \neq 0$ and $\alpha \neq 1$, the instantaneous expected excess return is:

$$\begin{aligned}\eta = \Gamma(-\alpha)\lambda\big[\big((\beta_+ - \gamma) - 1\big)^\alpha - (\beta_+ - \gamma)^\alpha + \big((\beta_- + \gamma) + 1\big)^\alpha \\ - (\beta_- + \gamma)^\alpha\big] - \Gamma(-\alpha)\lambda\big[(\beta_+ - 1)^\alpha - \beta_+^\alpha + (\beta_- + 1)^\alpha - \beta_-^\alpha\big],\end{aligned} \tag{52}$$

where the first line is the cumulant exponent under measure $\mathbb{P}$ evaluated at $s = 1$ and the second line is the cumulant exponent under measure $\mathbb{Q}$ evaluated at $s = 1$, with the term $C(h)$ in both cumulant exponents dropping out. Nevertheless, sometimes the measure change itself can induce an additional linear term that contributes to the expected excess return. Hence, it is safer to always evaluate η according to the equation $\eta = \varphi_J^{\mathbb{Q}}(1+\gamma) - \varphi_J^{\mathbb{Q}}(\gamma) - \varphi_J^{\mathbb{Q}}(1)$.

If we apply random time changes to the Lévy jump process and if the underlying activity rate risk is not correlated with the Lévy jump risk, we can simply

replace ηt with $\eta \mathcal{T}_t$ as the excess return over time period $[0, t]$. If the risk-neutral activity rate follows a square root process, the statistical dynamics for the activity rate can be derived analogous to (49) with $\rho = 0$ due to the orthogonality between the jump innovation in return and the diffusion innovation in the activity rate.

4.3 More flexible market price of risk specifications

Under the exponential martingale specification embedded in the pricing kernel in (42), return and volatility risks are both captured by a vector of time-changed Lévy processes, $(X^k_{\mathcal{T}_t^k}, X^{kv}_{\mathcal{T}_t^k})_{k=1}^K$ and the market prices on these risks, $(\gamma_k, \gamma_{kv})_{k=1}^K$ are assumed to be constant. The specification is parsimonious, under which the return (and activity rate) dynamics often stay within the same class under the two measures $\mathbb{P}$ and $\mathbb{Q}$. However, since tractability requirement mainly comes from option pricing due to the associated expectation operation under the risk-neutral measure, a more flexible market price of risk specification poses little problems if we start the modeling with a tractable risk-neutral dynamics. Complex market price of risk specifications only lead to complex statistical dynamics, which are irrelevant for option pricing. The complication does affect the derivation of the likelihood functions for time-series estimation. Yet, when the return series can be sampled frequently, an Euler approximation of the statistical dynamics often works well for estimation and it avoids the complication of taking expectations under the statistical measure for the conditional density derivation. Hence, we can specify arbitrarily complex market price of risks without incurring much difficulty for asset pricing. Beside the usual technical conditions that a pricing kernel needs to satisfy, the only practical constraints for the market price of risk specification come from reasonability and identification considerations. Even if a specification is mathematically allowed, we may discard it if it does not make economic sense and does not represent common investor behavior. Furthermore, a more flexible specification on the market price of risk gives us more degrees of freedom, but it can also cause difficulties in identification. Hence, it is always prudent to start with a parsimonious assumption on the market price of risk and consider extension only when the data ask for it.

Take the Black–Scholes model as a simple example, where the stock return under the risk-neutral measure $\mathbb{Q}$ is normally distributed with constant volatility σ as described in (29). Now we consider a flexible, but not necessarily reasonable, market price of risk specification that defines the measure change from $\mathbb{P}$ to $\mathbb{Q}$ as

$$\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_t = \exp\Big(-\big(\gamma_0 + \gamma_1 Z_t + \gamma_2 Z_t^2 + \gamma_3 Z_t^3\big)\sigma W_t - \frac{1}{2}\big(\gamma_0 + \gamma_1 Z_t + \gamma_2 Z_t^2 + \gamma_3 Z_t^3\big)^2 \sigma^2 t\Big), \tag{53}$$

where the market price of risk is given by a polynomial function of Z_t, $\gamma = \gamma_0 + \gamma_1 Z_t + \gamma_2 Z_t^2 + \gamma_3 Z_t^3$ with Z_t being some state variable whose dynamics is left unspecified. The order of three is purely arbitrary and for illustration only. Under this specification, the instantaneous expected excess return at time t is $\eta_t = (\gamma_0 + \gamma_1 Z_t + \gamma_2 Z_t^2 + \gamma_3 Z_t^3)\sigma^2$, and the $\mathbb{P}$-dynamics of the security price becomes

$$\mathrm{d}S_t/S_t = \big(r - q + \big(\gamma_0 + \gamma_1 Z_t + \gamma_2 Z_t^2 + \gamma_3 Z_t^3\big)\sigma^2\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t. \quad (54)$$

Many empirical studies identify dividend yield, default spread, interest rate, and lagged return as variables that can predict expected excess returns. If the evidence is robust, we can use them as the state variable Z_t, which can either be a scalar or a vector.

Regardless of the complexity of the statistical dynamics, option pricing still follows the Black–Scholes formula. The return distribution under the statistical measure $\mathbb{P}$ depends on the dynamics of Z_t. Nevertheless, with an Euler approximation, we can still assume that the conditional return distribution over a short time interval $[t, +\Delta t]$ is normally distributed, with mean $(r - q + \eta_t - \frac{1}{2}\sigma^2)\Delta t$ and variance $\sigma^2 \Delta t$, and then construct the conditional likelihood function of the return accordingly.

Consider another example where the activity rate follows a square-root dynamics under the risk-neutral measure,

$$\mathrm{d}v_t = (a - \kappa v_t)\,\mathrm{d}t + \sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v. \quad (55)$$

For simplicity, we assume that the return Lévy innovation is not correlated with the Brownian motion W_t^v in the activity rate process. As shown in a later section, the affine structure of the activity rate dynamics under the risk-neutral measure makes option pricing tractable. The previous section assumes a constant market price γ_v on $\sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v$,[6] which induces a drift change of $\gamma_v \sigma_v^2 v_t$. Hence, it amounts to change the mean-reversion coefficient from κ to $\kappa - \gamma_v \sigma_v^2$ under $\mathbb{P}$. Now we consider a more general specification,

$$\gamma_t^v = \gamma_0 / v_t + \gamma_1 + \gamma_2 v_t + \cdots + \gamma_k v_t^{k-1}, \quad (56)$$

for any order k. The induced drift change becomes: $\gamma_0 \sigma_v^2 + \gamma_1 \sigma_v^2 v_t + \gamma_2 \sigma_v^2 v_t^2 + \cdots + \gamma_k \sigma_v^2 v_t^k$. Thus, the drift of the activity rate process is no longer affine under the statistical measure, but the complication does not affect option pricing and we can resort to Euler approximation for the likelihood construction.

Nevertheless, the specification in (56) is not completely innocuous. When v_t approaches zero, its risk (innovation) $\sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v$ also approaches zero, yet the risk premium does not approach zero, but approaches a non-zero constant $\gamma_0 \sigma_v$. A riskless security cannot earn a non-zero risk premium. Hence,

[6] The literature often regards $\mathrm{d}W_t^v$ instead of $\sigma_v \sqrt{v_t}\,\mathrm{d}W_t^v$ as the risk. Then, our specification generates "proportional market price of risk" $\gamma_v \sigma_v \sqrt{v_t}$ on W_t^v, a language more commonly used in the literature.

the specification violates the no-arbitrage condition if the activity rate can stay at zero. Recently, Cheridito et al. (2003) and Pan and Singleton (2005) apply a restricted version of (56) with $\gamma_k = 0$ for $k \geqslant 2$. Then, the risk premium is affine in v_t and the activity rate dynamics remain affine under the statistical measure. To guarantee no-arbitrage, they add further technical conditions on the statistical dynamics so that zero is not an absorbing barrier, but a reflecting barrier of v_t. The technical condition guarantees no arbitrage. Nevertheless, the specification with γ_0 strictly nonzero still implies that investors charge a risk premium no smaller than $\gamma_0\sigma_v^2$ no matter how small the risk becomes. A flexible market rice of risk specification does not hinder option pricing as long as we start with the risk-neutral dynamics, but it remains important to apply our economic sense and the rule of parsimony and discipline in specifying them.

In the fixed income literature, an enormous amount of studies exploit various forms of the expectation hypothesis to predict future exchange rate movements using current interest rate differentials between the two economies, and predict short-term interest rate movements with the current term structure information. Several recent studies explore whether affine models can explain the regression slope coefficients.[7] Affine models ask that bond yields of all maturities are affine functions of a set of state variable. This cross-sectional relation has bearings on the risk-neutral dynamics: The risk-neutral drift and variance of the state vector are both affine functions of the state vector. However, it has no direct bearings on the statistical dynamics, nor on the expectation hypothesis. The above studies all require that the statistical drift of the state vector be also affine. This self-imposed requirement limits the market price of risk specification to an affine form $\gamma(X_t) = a + bX_t$ when the state variable has a constant diffusion, and of the form $\gamma(X_t) = a/X_t + b$ when the state variable follows a square root process.

5 Option pricing under time-changed Lévy processes

To price options when the underlying asset return is driven by Lévy processes with stochastic time changes, we first derive the generalized Fourier transform of the asset return under the risk-neutral measure and then use Fourier inversion methods to compute option prices numerically.

5.1 Deriving the Fourier transform

Carr and Wu (2004) propose a theorem that significantly enhances the tractability of option pricing under time-changed Lévy processes. They convert

[7] Examples include Backus et al. (2001a), Duffee (2002), Dai and Singleton (2002), and Roberds and Whiteman (1999) for expectation hypotheses on the term structure in a single economy, and Backus et al. (2001b) for international term structure and currency pricing.

the problem of finding the Fourier transform of a time-changed Lévy process into the problem of finding the Laplace transform of the random time under a new complex-valued measure,

$$\phi_Y(u) \equiv \mathbb{E}^{\mathbb{Q}}\big[e^{iuX_{\mathcal{T}_t}}\big] = \mathbb{E}^{\mathbb{M}}\big[e^{-\psi_x(u)\mathcal{T}_t}\big], \tag{57}$$

where $\psi_x(u)$ denotes the characteristic exponent of the underlying Lévy process X_t, and the second expectation is under a new measure $\mathbb{M}$, defined by the following complex-valued exponential martingale:

$$\left.\frac{d\mathbb{M}}{d\mathbb{Q}}\right|_t = \exp\big(iuX_{\mathcal{T}_t} + \mathcal{T}_t\psi_x(u)\big). \tag{58}$$

When the activity rate v_t underlying the time change is independent of the Lévy innovation X_t, the measure change is not necessary and the result in (57) can be obtained via the law of iterated expectations. When the two processes are correlated, the proposed measure change simplifies the calculation by absorbing the effect of correlation into the new measure.

According to (57), tractable Fourier transforms for the time-changed Lévy process, $\phi_Y(u)$, can be obtained if we can obtain tractable forms for the characteristic exponent of the Lévy process, $\psi_x(u)$, and the Laplace transform of the time change. The three most widely used Lévy jump specifications include the Merton (1976) compound Poisson models with normally distributed jump sizes, the dampened power law specification and its various special cases, and the normal inverse gamma model and its extensions. All these models have analytical solutions for the characteristic exponents.

To solve for the Laplace transform, it is important to note that if we write the time change $\mathcal{T}_t$ in terms of the activity rate $\mathcal{T}_t = \int_0^t v(s_-)\,ds$, the same form of expectation appears in the bond pricing literature with the analogous term for the instantaneous activity rate being the instantaneous interest rate. Furthermore, since both nominal interest rates and the activity rate are required to be positive, they can be modeled using similar dynamics. Therefore, any interest rate dynamics that generate tractable bond pricing formulas can be borrowed to model the activity rate dynamics under measure $\mathbb{M}$ with tractable solutions to the Laplace transform in Eq. (57). In particular, the affine class of Duffie and Kan (1996), Duffie et al. (2000, 2003a) and the quadratic class of Leippold and Wu (2002) for interest rates can be borrowed to model the activity rate dynamics with tractable exponential affine and exponential quadratic solutions for the Laplace transform, respectively. Carr and Wu (2004) discuss these models in their general forms. Of all these specifications, the most popular is the square root process used in Heston (1993) and its various extensions to multiple factors and to include positive jumps. The 3/2 activity rate dynamics also generate tractable solutions for the Laplace transform in (57), but the solution contains a confluent hypergeometric function $M(\alpha, \beta; z)$, where the two coefficients (α, β) are complex valued and are functions of the characteristic coefficient u, and the argument z is a function of the activity rate level

and option maturity. It remains a numerical challenge to compute this function efficiently over the wide range of complex-valued coefficients necessary for option pricing.

I illustrate the valuation procedure using the simple examples discussed in the previous sections, starting with the Black–Scholes model with the risk-neutral return dynamics given in (29):

$$\begin{aligned}\phi_s(u) &\equiv \mathbb{E}^{\mathbb{Q}}\left[e^{iu\ln S_t/S_0}\right] = e^{iu(r-q)t}\mathbb{E}^{\mathbb{Q}}\left[e^{iu(\sigma W_t-\frac{1}{2}\sigma^2 t)}\right]\\ &= e^{iu(r-q)t-\frac{1}{2}(iu+u^2)\sigma^2 t}.\end{aligned} \tag{59}$$

Given the constant interest rate and dividend yield assumption, we can factor them out before taking the expectation. In this case, the concavity adjustment term $iu\frac{1}{2}\sigma^2 t$ can also be factored out. Nevertheless, with time changes in mind, I leave it inside the expectation and write $\psi_x(u) = \frac{1}{2}(iu + u^2)\sigma^2$ as the characteristic exponent of the *concavity-adjusted return innovation* term: $X_t = \sigma W - \frac{1}{2}\sigma^2 t$.

The Black–Scholes option pricing formula is well known, deriving the generalized Fourier transform under the Black–Scholes model merely serves as a benchmark for more complicated examples. The first extension is to apply random time changes to the Black–Scholes specification,

$$\ln S_t/S_0 = (r-q)t + \sigma W_{\mathcal{T}_t} - \frac{1}{2}\sigma^2\mathcal{T}_t. \tag{60}$$

Here, we can apply Carr and Wu's theorem to find the generalized Fourier transform:

$$\phi_s(u) = e^{iu(r-q)t}\mathbb{E}^{\mathbb{Q}}\left[e^{iu(\sigma W_{\mathcal{T}_t}-\frac{1}{2}\sigma^2\mathcal{T}_t)}\right] = e^{iu(r-q)t}\mathbb{E}^{\mathbb{M}}\left[e^{-\psi_x(u)\mathcal{T}_t}\right], \tag{61}$$

where $\psi_x(u) = \frac{1}{2}(iu + u^2)\sigma^2$ is the same as for the concavity-adjusted return innovation for the Black–Scholes model. The construction of the new measure $\mathbb{M}$ and the Laplace transform under this new measure depend on the specification of the activity rate dynamics.

Take the Heston (1993) model as an example, where the activity rate dynamics under measure $\mathbb{Q}$ is, in stochastic differential equation form,

$$\mathrm{d}v_t = \kappa(1-v_t)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v, \quad \rho\,\mathrm{d}t = \mathbb{E}[\mathrm{d}W_t\,\mathrm{d}W_t^v]. \tag{62}$$

The measure change is defined by

$$\left.\frac{\mathrm{d}\mathbb{M}}{\mathrm{d}\mathbb{Q}}\right|_t = \exp\left(iu\left(\sigma W_{\mathcal{T}_t} - \frac{1}{2}\sigma^2\mathcal{T}_t\right) + \mathcal{T}_t\psi_x(u)\right). \tag{63}$$

The probabilistically equivalent writing under more traditional notation is

$$\left.\frac{\mathrm{d}\mathbb{M}}{\mathrm{d}\mathbb{Q}}\right|_t = \exp\left(iu\sigma\int_0^t\sqrt{v_s}\,\mathrm{d}W_s + \frac{1}{2}u^2\sigma^2\int_0^t v_s\,\mathrm{d}s\right), \tag{64}$$

where I plug in $\psi_x(u)$ and cancel out the concavity-adjustment term. This measure change induces a drift change in the activity rate dynamics given by the covariance term[8]:

$$\mu(v)^{\mathbb{M}}\,\mathrm{d}t - \mu(v)^{\mathbb{Q}}\,\mathrm{d}t = \langle iu\sigma\sqrt{v_t}\,\mathrm{d}W_t,\, \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v\rangle = iu\sigma\sigma_v v_t\rho\,\mathrm{d}t. \tag{66}$$

Hence, under measure $\mathbb{M}$, the activity rate dynamics become

$$\mathrm{d}v_t = \big(\kappa - \kappa^{\mathbb{M}}v_t\big)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v, \quad \kappa^{\mathbb{M}} = \kappa - iu\sigma\sigma_v\rho. \tag{67}$$

Both the drift and the instantaneous variance are affine in v_t under measure $\mathbb{M}$. The Laplace transform in (61) is exponential affine in the current level of the activity rate:

$$\phi_s(u) = e^{iu(r-q)t}\mathbb{E}^{\mathbb{M}}\big[e^{-\psi_x(u)\mathcal{T}_t}\big] = e^{iu(r-q)t - b(t)v_0 - c(t)}, \tag{68}$$

with the coefficients $b(t)$ and $c(t)$ given by

$$\begin{aligned} b(t) &= \frac{2\psi_x(u)(1-e^{-\xi t})}{2\xi - (\xi - \kappa^{\mathbb{M}})(1-e^{-\xi t})}, \\ c(t) &= \frac{\kappa}{\sigma_v^2}\left[2\ln\left(1 - \frac{\xi - \kappa^{\mathbb{M}}}{2\xi}\big(1-e^{-\xi t}\big)\right) + \big(\xi - \kappa^{\mathbb{M}}\big)t\right], \end{aligned} \tag{69}$$

with $\xi = \sqrt{(\kappa^{\mathbb{M}})^2 + 2\sigma_v^2\psi_x(u)}$.

Suppose we further allow the activity rate to revert to a stochastic central tendency factor in generating a two-factor activity rate dynamics under measure $\mathbb{Q}$:

$$\begin{aligned} \mathrm{d}v_t &= \kappa(m_t - v_t)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v, \\ \mathrm{d}m_t &= \kappa_m(1-m_t)\,\mathrm{d}t + \sigma_m\sqrt{m_t}\,\mathrm{d}W_t^m, \end{aligned} \tag{70}$$

with W_t^m being an independent Brownian motion. The dynamics under measure $\mathbb{M}$ becomes

$$\begin{aligned} \mathrm{d}v_t &= \big(\kappa m_t - \kappa^{\mathbb{M}}v_t\big)\,\mathrm{d}t + \sigma_v\sqrt{v_t}\,\mathrm{d}W_t^v, \quad \kappa^{\mathbb{M}} = \kappa - iu\sigma\sigma_v\rho, \\ \mathrm{d}m_t &= \kappa_m(1-m_t)\,\mathrm{d}t + \sigma_m\sqrt{m_t}\,\mathrm{d}W_t^m. \end{aligned} \tag{71}$$

Writing the dynamics in a matrix notation with $V_t \equiv [v_t, m_t]^\top$, we have

$$\mathrm{d}V_t = \big(a - \kappa_V^{\mathbb{M}}V_t\big)\,\mathrm{d}t + \sqrt{\Sigma V_t}\mathrm{d}W_t^V, \tag{72}$$

[8] In the integral form, the covariance is

$$\int_0^t \big(\mu(v_s)^{\mathbb{M}} - \mu(v_s)^{\mathbb{Q}}\big)\,\mathrm{d}s = \langle iu\sigma\,\mathrm{d}W_{\mathcal{T}_t},\, \sigma_v\,\mathrm{d}W_{\mathcal{T}_t}^v\rangle = iu\sigma\sigma_v v_t\rho\mathcal{T}_t. \tag{65}$$

with

$$a = \begin{bmatrix} 0 \\ \kappa_m \end{bmatrix}, \qquad \kappa_V^{\mathbb{M}} = \begin{bmatrix} \kappa^{\mathbb{M}} & -\kappa \\ 0 & \kappa_m \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_m^2 \end{bmatrix}.$$

Given the two-factor affine structure for the activity rate dynamics, the Laplace transform is exponential affine in the current level of the two factors $V_0 = [v_0, m_0]^\top$:

$$\phi_s(u) = e^{iu(r-q)t - b(t)^\top V_0 - c(t)}, \tag{73}$$

where the coefficients $b(t)$ and $c(t)$ can be solved from a set of ordinary differential equations:

$$\begin{aligned} b'(t) &= \psi_x(u) b_V - \left(\kappa^{\mathbb{M}}\right)^\top b(t) - \frac{1}{2}\Sigma\big[b(t) \odot b(t)\big], \\ c'(t) &= a^\top b(t), \end{aligned} \tag{74}$$

starting at $b(0) = 0$ and $c(0) = 0$, with $b_V = [1, 0]^\top$ denoting the instantaneous loading of the activity rate on the two factors and $\odot$ denoting the element-by-element product operation. The ordinary differential equations can be solved using standard numerical routines, such as an Euler approximation or the fourth-order Runge–Kutta method.

When the return innovation is not driven by a diffusion, but by a pure jump Lévy process such as the one governed by the dampened power law in (12), we simply need to replace the characteristic exponent of the concavity-adjusted diffusion component $\psi_x(u) = \frac{1}{2}(iu + u^2)\sigma^2$ by that of the concavity-adjusted jump component. With $\alpha \neq 0$ and $\alpha \neq 1$, we have:

$$\begin{aligned} \psi_x(u) = {} & -\Gamma(-\alpha)\lambda\big[(\beta_+ - iu)^\alpha - \beta_+^\alpha + (\beta_- + iu)^\alpha - \beta_-^\alpha\big] \\ & + iu\Gamma(-\alpha)\lambda\big[(\beta_+ - 1)^\alpha - \beta_+^\alpha + (\beta_- + 1)^\alpha - \beta_-^\alpha\big]. \end{aligned} \tag{75}$$

We can also include both a diffusion and a jump component, in which case the characteristic exponent becomes the sum of the two. Most importantly, we can treat the specification of the Lévy process and the time change separately, and hence derive the characteristic exponent $\psi_x(u)$ and the Laplace transform separately. Therefore, we can combine any tractable Lévy specifications with any tractable activity rate dynamics, and the generalized Fourier transform for the resultant return dynamics is tractable.

5.2 Computing the Fourier inversions

With tractable solutions to the generalized Fourier transform of the return distribution, European option prices can be computed by inverting the Fourier transform. The literature considers two broad ways of Fourier inversions. The first approach treat options analogous to a cumulative distribution function. Standard statistics books show how to invert the characteristic function to obtain a cumulative function. The inversion formula for option prices can be

analogously proved. The second approach treats the option price analogous to a probability density function. In this case, for the transform to be well-defined, the characteristic coefficient u in (57) needs to contain an imaginary component, the domain of which depends on the payoff structure. Based on this analogy, option prices across the whole spectrum of strikes can be obtained via fast Fourier transform (FFT).

For both cases, the Fourier transforms for a wide variety of European payoffs can be obtained. Their values can then be obtained by inverting the corresponding transforms. I use a European call option as an example to illustrate the transform methods. Indeed, in most situations, a call option value is all we need because most European payoff functions can be replicated by a portfolio of European call options across different strikes but at the same maturity.

The terminal payoff of a European call option at maturity t and strike K is,

$$\Pi_t = (S_t - K)1_{S_t \geqslant K}. \tag{76}$$

Since we have derived the Fourier transform of the asset returns, it is convenient to represent the payoff in log return terms,

$$\Pi_t = S_0\big(e^{\ln S_t/S_0} - e^{\ln K/S_0}\big)1_{S \geqslant K} = S_0\big(e^{s_t} - e^{k}\big)1_{s_t \geqslant k}, \tag{77}$$

with $s_t = \ln S_t/S_0$ and $k = \ln K/S_0$. The time-0 value of the call option is

$$C(K,t) = S_0 e^{-rt}\mathbb{E}_0^{\mathbb{Q}}\big[\big(e^{s_t} - e^{k}\big)1_{s_t \geqslant k}\big]. \tag{78}$$

Let $C(k) = C(K,t)/S_0$ denote the call option value in percentages of the current spot price level as a function of moneyness k and maturity t. In what follows, I focus on computing the relative call value $C(k)$. We can simply multiply it by the spot price to obtain the absolute call option value $C(K,t)$.[9] For notational clarity, we henceforth drop the maturity argument when no confusion shall occur.

5.2.1 *The cumulative distribution analogy*

We rewrite the call option value in terms of $x = -k$,

$$C(x) = C(k = -x) = e^{-rt}\mathbb{E}_0^{\mathbb{Q}}\big[\big(e^{s_t} - e^{-x}\big)1_{-s_t \leqslant x}\big]. \tag{79}$$

Treating the call option value $C(x)$ analogous to a cumulative distribution, we define its Fourier transform as

$$\chi_c(z) \equiv \int_{-\infty}^{\infty} e^{izk}\,\mathrm{d}C(x), \quad z \in \mathbb{R}. \tag{80}$$

[9] Some broker dealers provide the relative percentage quote $C(k)$ instead of the absolute quote $C(K,t)$ to achieve quote stability by excluding the impact of spot price fluctuation.

We can derive this transform in terms of the Fourier transform of the return $\phi_s(u)$:

$$\begin{aligned}\chi_c(z) &= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\int_{-\infty}^{\infty} e^{izx}\left(e^{s}\delta_{-s\leqslant x} - e^{-x}\delta_{-s\leqslant x} + e^{-x}1_{-s\leqslant x}\right)\mathrm{d}x\right] \\ &= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[e^{(1-iz)s} - e^{(1-iz)s} + \int_{-s}^{\infty} e^{(iz-1)x}\,\mathrm{d}x\right] \\ &= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\frac{e^{(1-iz)s}}{(1-iz)}\right] = e^{-rt}\frac{\phi_s(-i-z)}{1-iz}, \end{aligned} \tag{81}$$

which is solved by first applying Fubini's theorem and then applying the result on the Fourier transform of a Dirac function $\delta_{-s\leqslant x}$. Thus, tractable forms for the return transform $\phi_s(u)$ also means tractable forms for the option transform $\chi_c(z)$.

Given this transform, the option value can be computed via the following Fourier inversion formula:

$$C(x) = \frac{1}{2}\chi_c(0) + \frac{1}{2\pi}\int_0^{\infty}\frac{e^{izx}\chi_c(-z) - e^{-izx}\chi_c(z)}{iz}\,\mathrm{d}z. \tag{82}$$

The inversion formula and its proof are very much analogous to the inversion formula for a cumulative distribution (Alan and Ord, 1987). The only difference is at the boundary: For a cumulative distribution, the transform evaluated at $z = 0$ is one; for the option transform, it is $\chi_c(0) = e^{-rt}\phi_s(-i) = e^{-qt}$. Given $C(x)$, we obtain $C(k) = C(k = -x)$. We can also directly define the inversion formula as

$$C(k) = \frac{1}{2}\chi_c(0) + \frac{1}{2\pi}\int_0^{\infty}\frac{e^{-izk}\chi_c(-z) - e^{izk}\chi_c(z)}{iz}\,\mathrm{d}z, \tag{83}$$

$$\begin{aligned} &= e^{-rt}\left[\frac{1}{2}\phi_s(-i) - \frac{1}{2\pi}\int_0^{\infty}\left(e^{-izk}\frac{\phi_s(z-i)}{z^2-iz}\right.\right. \\ &\quad \left.\left. + e^{izk}\frac{\phi_s(-z-i)}{z^2+iz}\right)\mathrm{d}z\right]. \end{aligned} \tag{84}$$

To compute the option value from the transform, the inversion formula in (82) asks for a numerical integration of an oscillating function. Fortunately, being a weighted average of cosines, the integrand exhibits much less oscillatory behavior than the transform $\psi(u)$ itself. The integral can numerically be evaluated using quadrature methods (Singleton, 2001).

Duffie et al. (2000) and Leippold and Wu (2002) discuss the application of this approach for the valuation of general European-type state-contingent claims in the context of affine and quadratic models, respectively. In earlier works, e.g., Chen and Scott (1992), Heston (1993), Bates (1996), and Bakshi et al. (1997), the call option value is often written as the portfolio of two contingent claims:

$$\begin{aligned} C(x) &= e^{-rt}\mathbb{E}^{\mathbb{Q}}\big[e^{s}\big]\frac{\mathbb{E}^{\mathbb{Q}}[e^{s}1_{-s\leqslant x}]}{\mathbb{E}^{\mathbb{Q}}[e^{s}]} - e^{-rt}e^{-x}\mathbb{E}^{\mathbb{Q}}[1_{-s\leqslant x}] \\ &= e^{-qt}Q_1(x) - e^{-rt}e^{-x}Q_2(x), \end{aligned} \tag{85}$$

with $Q_1(x)$ and $Q_2(x)$ being the values of two contingent claims defined by

$$Q_1(x) = \frac{\mathbb{E}^{\mathbb{Q}}[e^{s}1_{-s\leqslant x}]}{\phi_s(-i)}, \qquad Q_2(x) = \mathbb{E}^{\mathbb{Q}}[1_{-s\leqslant x}]. \tag{86}$$

Q_2 is simply the cumulative distribution of $-s$. Its transform is

$$\begin{aligned} \chi_2(z) &= \int_{-\infty}^{\infty} e^{izx}\,\mathrm{d}Q_2(x) = \mathbb{E}^{\mathbb{Q}}\left[\int_{-\infty}^{\infty} e^{izx}\delta_{-s\leqslant x}\,\mathrm{d}x\right] \\ &= \mathbb{E}^{\mathbb{Q}}\big[e^{-izs}\big] = \phi_s(-z). \end{aligned} \tag{87}$$

The transform of $Q_1(x)$ is

$$\begin{aligned} \chi_1(z) &= \frac{1}{\phi_s(-i)}\mathbb{E}^{\mathbb{Q}}\left[\int_{-\infty}^{\infty} e^{izx}e^{s}\delta_{-s\leqslant x}\,\mathrm{d}x\right] = \frac{1}{\phi_s(-i)}\mathbb{E}^{\mathbb{Q}}\big[e^{(1-iz)s}\big] \\ &= \frac{\phi_s(-z-i)}{\phi_s(-i)}. \end{aligned} \tag{88}$$

Applying the inversion formula in (83), we have the values for the two contingent claims as

$$Q_1(k) = \frac{1}{2} + \frac{1}{2\pi\phi_s(-i)}\int_0^{\infty} \frac{e^{-izk}\phi_s(z-i) - e^{izk}\phi(-z-i)}{iz}\,\mathrm{d}z, \tag{89}$$

$$Q_2(k) = \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{e^{-izk}\phi_s(z) - e^{izk}\phi_s(-z)}{iz}\,\mathrm{d}z. \tag{90}$$

Nevertheless, doing one numerical integration according to my proposed transform in (84) is more efficient than doing two numerical integrations according to (89) and (90).

5.2.2 The probability density analogy

The second approach treats the option price analogous to a probability density and defines the Fourier transform of the option value as

$$\chi_p(z) \equiv \int_{-\infty}^{\infty} e^{izk} C(k)\,\mathrm{d}k, \quad z = z_r - iz_i,\ z_r \in \mathbb{R}, z_i \in \mathcal{D} \subseteq \mathbb{R}^+. \quad (91)$$

The transform coefficient z is extended to the complex plane to guarantee the finiteness of the transform. For the call option value, the transform is,

$$\begin{aligned}
\chi_p(z) &= \int_{-\infty}^{\infty} e^{izk}\mathbb{E}^{\mathbb{Q}}\left[e^{-rt}\left(e^s - e^k\right)1_{s\geqslant k}\right]\mathrm{d}k \\
&= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\int_{-\infty}^{\infty} e^{izk}\left(e^s - e^k\right)1_{s\geqslant k}\,\mathrm{d}k\right] \\
&= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\int_{-\infty}^{s} e^{izk}\left(e^s - e^k\right)\mathrm{d}k\right] \\
&= e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\left(\frac{e^{izk}e^s}{iz} - \frac{e^{(iz+1)k}}{iz+1}\right)\Bigg|_{k=-\infty}^{k=s}\right]. \qquad (92)
\end{aligned}$$

For $e^{izk} = e^{iz_rk+z_ik}$ to be convergent (to zero) at $k = -\infty$, we need $z_i > 0$, under which $e^{(iz+1)k}$ also converges to zero.[10] With $z_i > 0$, the transform for the call option value becomes

$$\chi_p(z) = e^{-rt}\mathbb{E}^{\mathbb{Q}}\left[\frac{e^{(1+iz)s}}{iz} - \frac{e^{(iz+1)s}}{iz+1}\right] = e^{-rt}\frac{\phi_s(z-i)}{(iz)(iz+1)}. \qquad (93)$$

For some return distributions, the return transform $\phi_s(z-i) = \mathbb{E}^{\mathbb{Q}}[e^{(1+iz)s}]$ is well-defined only when z_i is in a subset of the real line. In Eq. (91), we use $\mathcal{D} \in \mathbb{R}^+$ to denote the subset that both guarantees the convergence of e^{izk} and $e^{(iz+1)k}$ at $k = -\infty$, and assures the finiteness of the transform $\phi_s(z-i)$.

Given a finite transform $\chi_p(z)$ for the call option, the option value can be computed from the following Fourier inversion formula:

$$C(k) = \frac{1}{2}\int_{-iz_i-\infty}^{-iz_i+\infty} e^{-izk}\chi_p(z)\,\mathrm{d}z = \frac{e^{-z_ik}}{\pi}\int_0^{\infty} e^{-iz_rk}\chi_p(z_r - iz_i)\,\mathrm{d}z_r. \qquad (94)$$

[10] For other types of contingent claims, the transform will take different forms and the required domain for z_i that guarantees the finiteness of the transform varies accordingly.

We can approximate the integral using summations:

$$C(k) \approx \widehat{C}(k) = \frac{e^{-z_i k}}{\pi} \sum_{n=0}^{N-1} e^{-iz_r(n)k} \chi_p\big(z_r(n) - iz_i\big)\, \Delta z_r, \tag{95}$$

where $z_r(n)$ are the nodes of z_r and Δz_r is the spacing between nodes. The fast Fourier transform (FFT) is an efficient algorithm for computing the discrete Fourier coefficients. The discrete Fourier transform is a mapping of $\mathbf{f} = (f_0, \ldots, f_{N-1})^\top$ on the vector of Fourier coefficients $\mathbf{d} = (d_0, \ldots, d_{N-1})^\top$, such that

$$d_j = \sum_{n=0}^{N-1} f_n e^{-jn\frac{2\pi}{N}i}, \quad j = 0, 1, \ldots, N-1. \tag{96}$$

We use $\mathbf{d} = D(\mathbf{f})$ to denote the fast Fourier transform, which allows the efficient calculation of $\mathbf{d}$ if N is an even number, say $N = 2^m$, $m \in \mathbb{N}$. The algorithm reduces the number of multiplications in the required N summations from an order of 2^{2m} to that of $m2^{m-1}$, a very considerable reduction. By a suitable choice of Δz_r and a discretization scheme for k, we can cast the approximation in the form of (96) to take advantage of the computational efficiency of the FFT.

Following Carr and Madan (1999), we set $z_r(n) = \eta n$ and $k_j = -b + \lambda j$, and require $\eta\lambda = 2\pi/N$. Then, we can cast the option valuation approximation in (95) in the form of the FFT summation in (96):

$$\widehat{C}(k_j) = \sum_{n=0}^{N-1} f_n e^{-jn\frac{2\pi}{N}i} = D_j(\mathbf{f}), \quad j = 0, 1, \ldots, N-1, \tag{97}$$

with

$$f_n = \frac{1}{\pi} e^{-z_i k_j + ib\eta n} \eta \chi_p(\eta n - iz_i). \tag{98}$$

Under such a discretization scheme, the effective upper limit for the integration is $N\eta$, with a spacing of η. The range of log strike level is from $-b$ to $N\lambda - b$, with a uniform spacing of λ in the log strike. To put at-the-money ($k = 0$) option at the middle of the strike range, we can set $b = N\lambda/2$.

The restriction of $\eta\lambda = 2\pi/N$ reveals the trade-off between a fine grid in log strike and a fine grid in summation. With $N = 2^{12}$, Carr and Madan (1999) set $\eta = 0.25$ to price stock options. To price currency and interest-rate options, I often set $\eta = 1$ to generate a finer spacing of strikes and hence more option values within the relevant range. The choice of the imaginary part of the transform coefficient z_i also affects the numerical accuracy of the fast Fourier inversion. Lee (2004) provides detailed analysis on the error bounds and on the choice of the imaginary part of the transform coefficient z_i.

5.2.3 *Fractional FFT*

Recently, Chourdakis (2005) adopts the fractional Fourier transform (FRFT) method of Bailey and Swartztrauber (1991) in inverting the option transform $\chi_p(z)$. The method can efficiently compute

$$d_j = \sum_{n=0}^{N-1} f_n e^{-jn\alpha i}, \quad j = 0, 1, \ldots, N-1, \tag{99}$$

for any value of the parameter α. The standard FFT can be seen as a special case for $\alpha = 2\pi/N$. Therefore, we can use the FRFT method to compute

$$\widehat{C}(k, t) = \sum_{n=0}^{N-1} f_n e^{-jn\eta\lambda i}, \quad j = 0, 1, \ldots, N-1, \tag{100}$$

without the trade-off between the summation grid η and the strike spacing λ.

We use $\mathbf{d} = D(\mathbf{f}, \alpha)$ to denote the FRFT operation, with $D(\mathbf{f}) = D(\mathbf{f}, 2\pi/N)$ being the standard FFT as a special case. An N-point FRFT can be implemented by invoking three $2N$-point FFT procedures. Define the following $2N$-point vectors:

$$\mathbf{y} = \left(\left(f_n e^{i\pi n^2\alpha}\right)_{n=0}^{N-1}, (0)_{n=0}^{N-1}\right), \tag{101}$$

$$\mathbf{z} = \left(\left(e^{i\pi n^2\alpha}\right)_{n=0}^{N-1}, \left(e^{i\pi(N-n)^2\alpha}\right)_{n=0}^{N-1}\right). \tag{102}$$

The FRFT is given by

$$D_k(\mathbf{h}, \alpha) = \left(e^{i\pi k^2\alpha}\right)_{k=0}^{N-1} \odot D_k^{-1}\big(D_j(\mathbf{y}) \odot D_j(\mathbf{z})\big), \tag{103}$$

where $D_k^{-1}(\cdot)$ denotes the inverse FFT operation and $\odot$ denotes element-by-element vector multiplication. Due to the multiple application of the FFT operations, Chourdakis (2005) shows that an N-point FRFT procedure demands a similar number of elementary operations as a $4N$-point FFT procedure. However, given the free choices on λ and η, FRFT can be applied more efficiently. Using a smaller N with FRFT can achieve the same option pricing accuracy as using a much larger N with FFT. Numerical analysis shows that with similar computational time, the FRFT method can often achieve better computational accuracy than the FFT method. The accuracy improvement is larger when we have a better understanding of the model and model parameters so that we can set the boundaries more tightly. Nevertheless, the analysis also reveals a few cases of complete breakdown when the model takes extreme parameters and when the bounds are set too tight. Hence, the more freedom also asks for more discretion and caution in applying this method to generate robust results in all situations. This concern becomes especially important for model estimation, during which the trial model parameters can vary greatly.

6 Estimating Lévy processes with and without time changes

Model estimation can be classified into three categories: (1) estimating the statistical dynamics to capture the behavior of the time-series returns, (2) estimating the risk-neutral dynamics to match the option price behavior, and (3) estimating the statistical and risk-neutral dynamics jointly using both time-series returns and option prices and learning the behavior of market prices of various sources of risks.

6.1 Estimating statistical dynamic using time-series returns

Without time change, a Lévy process implies that the security returns are i.i.d. Thus, we can regard each day's return as random draws from the same distribution. This property makes the maximum likelihood method easy to implement. For the Lévy processes that I have discussed in this paper, only a few of them have analytical density functions, but virtually all of them have analytical characteristic functions. We can use fast Fourier transform (FFT) to numerically convert the characteristic function into density functions. Carr et al. (2002) use this method to estimate the CGMY model to stock returns. To implement this method, we normally need to use a large number N for the FFT so that we obtain numerical density values at a fine grid of realizations. Then, we can map the actual data to the grids by grouping the actual realizations into different bins that match the grids of the FFT and assign the same likelihood for realizations within the same bin. Alternatively, we can simply interpolate the density values from the FFT to match the actual realizations. Furthermore, to improve numerical stability and to generate enough points in the relevant FFT range, it is often helpful to standardize the return series (Wu, 2006).

The estimation becomes more involved when the model contains random time changes. Since the activity rates are not observable, some filtering technique is often necessary to determine the current level of the activity rates. Eraker et al. (2003) and Li et al. (2007) propose to estimate the dynamics using a Bayesian approach involving Markov Chain Monte Carlo (MCMC) simulation. They use MCMC to Bayesian update the distribution of both the state variables and model parameters. Javaheri (2005) propose a maximum likelihood method in estimating time-changed Lévy processes. Under this method, the distribution of the activity rates are predicted and updated according to Bayesian rules and using Markov Chain Monte Carlo simulation. Then, the model parameters are estimated by maximizing the likelihood of the time-series returns. Kretschmer and Pigorsch (2004) propose to use the efficient method of moments (EMM) of Gallant and Tauchen (1996).

6.2 Estimating risk-neutral dynamic to fit option prices

If the objective is to estimate a Lévy process for the risk-neutral return dynamics using option prices, nonlinear least square or some variant of it is

the most direct method to use. Since a Lévy process implies i.i.d. returns, the conditional return distribution over a fixed time horizon remains the same at different dates. Accordingly, the option price behavior across strikes and time-to-maturities, when scaled by the spot price, should remain the same across the different dates. In particular, the Black–Scholes implied volatility surface across moneyness and time-to-maturity should remain the same across different days. In reality, however, the option price behavior does change over time. For example, the implied volatility levels vary over time. The shape of the implied volatility smile also varies over time. A Lévy model without time change cannot capture these time variations. A common practice in the industry is to re-estimate the model daily, that is, to use different model parameters to match the different implied volatility levels and shapes at different days. This method is convenient and is also used in early academic works, e.g., Bakshi et al. (1997) and Carr and Wu (2003a).

In fact, even for one day, most Lévy processes have difficulties fitting the implied volatility surface across different maturities. The implied volatility smile observed from the market often persists as maturity increases, implying that the risk-neutral return distribution remains highly non-normal at long horizons. Yet, since Lévy models imply i.i.d. returns, if the return variance is finite under the model specification, the classic central limit theorem dictates that the skewness of the return distribution declines like the reciprocal of the square root of the horizon and the excess kurtosis declines like the reciprocal of horizon. Hence, return non-normality declines rapidly with increasing maturities. For these models, calibration is often forced to be done at each maturity. A different set of model parameters are used to fit the implied volatility smile at different maturities.

Carr and Wu (2003a) uses a maximum negatively skewed α-stable process to model the stock index return. Although the model-implied return distribution is i.i.d., the model-implied return variance is infinite and hence the central limit theorem does not apply. Thus, the model is capable of generating persistent implied volatility smiles across maturities. Wu (2006) use the dampened power law to model the index return innovation. With exponential dampening under the statistical measure, return variance is finite and the central limit theorem applies. The statistical return distribution is non-normal at high sampling frequencies but converges to normal rapidly with time aggregation. However, by applying a measure change using an exponential martingale, the dampening on the left tail can be made to disappear under the risk-neutral measure so that the return variance becomes infinite under the risk-neutral measure and the risk-neutral return non-normality no longer disappears with increasing option maturity.

Applying stochastic time change to Lévy processes not only generates time variation in the return distribution, but also generates cross-sectional option price behaviors that are more consistent with market observations. For example, a persistent activity rate process can generate non-normality out of a normal return innovation and can slow down the convergence of a non-normal

return distribution to normality. For daily calibration, the unobservable activity rates are treated the same as model parameters. They are all used as free inputs to make the model values fit market observations.

A dynamically consistent estimation is to keep the model parameters constant and only allow the activity rates to vary over time. Huang and Wu (2004) employ a nested nonlinear least square procedure for this purpose. Given parameter guesses, they minimize the pricing errors at each day to infer the activity rates at that day. Then, the parameters are chosen to minimize the aggregate pricing errors over the whole sample period. Carr and Wu (2007a) cast the models in a state-space form and estimate the model parameters using the maximum likelihood method. The state propagation equations are defined by the time-series dynamics of the activity rates and the measurement equations are defined on the option prices. Given parameter guesses, they use an extended version of the Kalman filter, the unscented Kalman filter (Wan and van der Merwe, 2001), to obtain the forecasts and filtering on the conditional mean and variance of the states and measurements. Then, they construct the likelihood of the option series assuming normally distributed forecasting errors. Using this approach, they identify both the statistical and the risk-neutral dynamics of the activity rates, and thus the market price of the activity rate risk. Nevertheless, by using only options data, they do not estimate the statistical return dynamics, nor the market price of return risk.

6.3 Static and dynamic consistency in model estimation

Daily re-calibration or re-calibration at each option maturity raises the issue of internal consistency. Option values generated from a no-arbitrage model are internally consistent with one another and do not generate arbitrage opportunities among themselves. When a model is re-calibrated at each maturity, the option values generated at different maturities are essentially from different models and hence the internal consistency between them is no longer guaranteed. When a model is re-calibrated daily, option values generated from the model at one day are not guaranteed to be consistent with option values generated at another day. One of the potential dangers of doing daily re-calibration is in risk management. A "fully" hedged option portfolio based on a model assuming constant model parameters is destined to generate hedging errors if the model parameters are altered on a daily basis.

Both the academia and practitioners appreciate the virtue of being both cross-sectionally and dynamically consistent. Nevertheless, building a dynamically consistent model that fits the market data well can be difficult. Hence, the daily re-calibration method can be regarded as a compromise to achieve static consistency cross-sectionally but not dynamic consistency over time. It remains true that a hedging strategy with constant parameter assumptions is bound to generate hedging errors when the model parameters are altered. One way to minimize the impact of varying parameters is to consider short investment horizons. For example, an investor that closes her position daily does not need

to worry about the dynamic inconsistency of daily re-calibration. Within her investment horizon of one day, the model parameters are fixed and the option values generated from the model are internally consistent. Market makers are often regarded as very short-term investors since they rarely hold long-term inventories. Therefore, dynamic consistency may not be as an overriding concern as it is to long-term investors. The more pressing concern for market makers is to achieve cross-sectional consistency across quotes for different contracts at a point in time. Furthermore, since they need to provide two-sided quotes, they often need a model that can match the current market quotes well.

On the other hand, for a hedge fund that bets on long-term convergence, a model that always fits the data well is not the key requirement. In fact, since their objective is to find market mispricings, it is important that their model can generate values that differ from the market. A good model produces pricing errors that are zero on average and transient in nature, so that if the model picks out a security that is over-valued, the over-valuation disappears in the near future. However, although they have a less stringent requirement on the model's fitting performance, they often have a more stringent requirement for dynamic consistency when they bet on long-term convergence. To them, it is important to keep the model parameters fixed over time and only allow state variables to vary, even if such a practice increases the model complexity and sometimes also increase the pricing errors of the model.

In a dynamically consistent model, the parameters that are allowed to vary daily should be converted into state variables, and their dynamics should be priced when valuing a contingent claim. Stochastic time change provides an intuitive and tractable way of turning a static model to a dynamic one. Under Lévy processes with stochastic time changes, we can build tractable models that generate reasonable pricing performance while maintaining dynamic consistency. Recent developments in econometrics further enable us to estimate these models with dynamic consistency constraints and within a reasonably short time framework. Once estimated, updating the activity rates based on newly arrived option quotes can be done almost instantaneously. Hence, it causes no delays in trading or market making.

6.4 Joint estimation of statistical and risk-neutral dynamics

One of the frontiers in the academic literature is to exploit the information in the derivatives market to infer the market prices on various sources of risks. While a long time series can be used to estimate the statistical dynamics of the security return, a large cross section of option prices across multiple strikes and maturities provide important information about the risk-neutral dynamics. The market prices of various sources of risks dictate the difference between the return dynamics under the two measures. Hence, estimation using both time-series and cross-sectional data can help us identify the dynamics under both measures and the market pricing on various sources of risks.

Pan (2002) uses the generalized methods of moments to estimate affine jump-diffusion stochastic volatility models under both probability measures and study the jump risk premia implicit in options. The moment conditions are constructed using both options and time-series returns. Eraker (2004) estimate similar dynamics under both measures using the MCMC approach. At each day, he uses the time-series returns and a few randomly sampled option prices. As a result, many available options data are thrown out in his estimation. Bakshi and Wu (2005) propose a maximum likelihood approach, where the likelihood on options and on time-series returns are constructed sequentially and the maximization is over the sum of the likelihoods on the two sets of data. First, they cast the activity rate dynamics into a state-propagation equation and the option prices into measurement equations. Second, they use the unscented Kalman filter to predict and update on the activity rates. Third, the likelihood on the options are constructed based on the forecasting errors on the options assuming normal forecasting errors. Fourth, they take the filtered activity rates as given and construct the likelihood of the returns conditional on the filtered activity rates. The conditional likelihood can be obtained using fast Fourier inversion of the conditional characteristic function. Finally, model parameters are chosen to maximize the sum of the likelihood of the time-series returns and option prices. They use this estimation procedure to analyze the variation of various sources of market prices around the Nasdaq bubble period.

7 Concluding remarks

Lévy processes with stochastic time changes have become the universal building blocks for financial security returns. Different Lévy components can be used to capture both continuous and discontinuous movements. Stochastic time changes can be applied to randomize the intensity of these different movements to generate stochastic time variation in volatility and higher return moments. I provide a summary on how different return behaviors can be captured by different Lévy components and different ways of applying time changes, under both the risk-neutral measure and the statistical measure. I also discuss how to compute European option values under these specifications using Fourier transform methods, and how to estimate the model parameters using time-series returns and option prices.

Acknowledgements

I thank Vadim Linetsky (the editor), Peter Carr, Xiong Chen, Roger Lee, Haitao Li, Hengyong Mo, and Yi Tang for comments. All remaining errors are mine.

References

Aït-Sahalia, Y. (2004). Disentangling diffusion from jumps. *Journal of Financial Economics* 74, 487–528.

Aït-Sahalia, Y., Jacod, J. (2005). Fisher's information for discretely sample Lévy processes. Working paper. Princeton University.

Alan, S., Ord, J.K. (1987). *Kendall's Advanced Theory of Statistics*, vol. 1. Oxford Univ. Press, New York.

Ané, T., Geman, H. (2000). Order flow, transaction clock and normality of asset returns. *Journal of Finance* 55, 2259–2284.

Backus, D., Foresi, S., Mozumdar, A., Wu, L. (2001a). Predictable changes in yields and forward rates. *Journal of Financial Economics* 59, 281–311.

Backus, D., Foresi, S., Telmer, C. (2001b). Affine term structure models and the forward premium anomaly. *Journal of Finance* 56, 279–304.

Bailey, D.H., Swartztrauber, P.N. (1991). The fractional Fourier transform and applications. *SIAM Review* 33, 389–404.

Bakshi, G., Wu, L. (2005). Investor irrationality and the Nasdaq bubble. Working paper. University of Maryland and Baruch College.

Bakshi, G., Cao, C., Chen, Z. (1997). Empirical performance of alternative option pricing models. *Journal of Finance* 52, 2003–2049.

Bakshi, G., Ju, N., Ou-Yang, H. (2006). Estimation of continuous-time models with an application to equity volatility. *Journal of Financial Economics* 82, 227–429.

Balduzzi, P., Das, S., Foresi, S. (1998). The central tendency: A second factor in bond yields. *Review of Economics and Statistics* 80, 62–72.

Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type. *Finance and Stochastics* 2, 41–68.

Bates, D. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies* 9, 69–107.

Bertoin, J. (1996). *Lévy Processes*. Cambridge Univ. Press, Cambridge.

Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–654.

Broadie, M., Chernov, M., Johannes, M. (2002). Jumps in volatility: The evidence of index options. Working paper. Columbia University.

Carr, P., Madan, D. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance* 2, 61–73.

Carr, P., Sun, J. (2005). A new approach for option pricing under stochastic volatility. Working paper. Bloomberg LP.

Carr, P., Wu, L. (2003a). Finite moment log stable process and option pricing. *Journal of Finance* 58, 753–777.

Carr, P., Wu, L. (2003b). What type of process underlies options? A simple robust test. *Journal of Finance* 58, 2581–2610.

Carr, P., Wu, L. (2004). Time-changed Lévy processes and option pricing. *Journal of Financial Economics* 71, 113–141.

Carr, P., Wu, L. (2005). Stock options and credit default swaps: A joint framework for valuation and estimation. Working paper. New York University and Baruch College.

Carr, P., Wu, L. (2007a). Stochastic skew in currency options. *Journal of Financial Economics*, in press.

Carr, P., Wu, L. (2007b). Theory and evidence on the dynamic interactions between sovereign credit default swaps and currency options. *Journal of Banking and Finance*, in press.

Carr, P., Geman, H., Madan, D., Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75, 305–332.

Chacko, G., Viceira, L. (2003). Spectral GMM estimation of continuous-time processes. *Journal of Econometrics* 116, 259–292.

Chen, R.-R., Scott, L. (1992). Pricing interest rate options in a two-factor Cox-Ingersoll-Ross model of the term structure. *Review of Derivatives Studies* 5, 613–636.

Cheridito, P., Filipović, D., Kimmel, R.L. (2003). Market price of risk specification for affine models: Theory and evidence. Working paper. Princeton University.

Chourdakis, K.M. (2005). Option pricing using the fractional FFT. *Journal of Computational Finance* 8, 1–18.
Cox, J.C., Ingersoll, J.E., Ross, S.R. (1985). A theory of the term structure of interest rates. *Econometrica* 53, 385–408.
Dai, Q., Singleton, K. (2002). Expectation puzzles, time-varying risk premia, and affine models of the term structure. *Journal of Financial Economics* 63, 415–441.
David, A., Veronesi, P. (1999). Option prices with uncertain fundamentals: Theory and evidence on the dynamics of implied volatilities. Working paper. Federal Reserve and University of Chicago.
Duffee, G.R. (2002). Term premia and interest rate forecasts in affine models. *Journal of Finance* 57, 405–443.
Duffie, D., Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance* 6, 379–406.
Duffie, D., Singleton, K. (1999). Modeling term structure of defaultable bonds. *Review of Financial Studies* 12, 687–720.
Duffie, D., Singleton, K. (2003). *Credit Risk: Pricing, Measurement and Management*. Princeton Univ. Press, Princeton, NJ.
Duffie, D., Pan, J., Singleton, K. (2000). Transform analysis and asset pricing for affine jump diffusions. *Econometrica* 68, 1343–1376.
Duffie, D., Filipović, D., Schachermayer, W. (2003a). Affine processes and applications in finance. *Annals of Applied Probability* 13, 984–1053.
Duffie, D., Pedersen, L.H., Singleton, K. (2003b). Modeling sovereign yield spreads: A case study of Russian debt. *Journal of Finance* 58, 119–160.
Eberlein, E., Keller, U., Prause, K. (1998). New insights into smile, mispricing, and value at risk: The hyperbolic model. *Journal of Business* 71, 371–406.
Engle, R. (2004). Risk and volatility: Econometric models and financial practice. *American Economic Review* 94, 405–420.
Eraker, B. (2004). Do stock prices and volatility jump? Reconciling evidence from spot and option prices. *Journal of Finance* 59, 1367–1404.
Eraker, B., Johannes, M., Polson, N. (2003). The impact of jumps in equity index volatility and returns. *Journal of Finance* 58, 1269–1300.
Fama, E.F. (1965). The behavior of stock market prices. *Journal of Business* 38, 34–105.
Foresi, S., Wu, L. (2005). Crash-o-phobia: A domestic fear or a worldwide concern? *Journal of Derivatives* 13, 8–21.
Gallant, A.R., Tauchen, G. (1996). Which moment to match. *Econometric Theory* 12, 657–681.
Heston, S. (1993). Closed-form solution for options with stochastic volatility, with application to bond and currency options. *Review of Financial Studies* 6, 327–343.
Heston, S. (1997). A simple new formula for options with stochastic volatility. Working paper. University of Maryland.
Huang, J., Wu, L. (2004). Specification analysis of option pricing models based on time-changed Lévy processes. *Journal of Finance* 59, 1405–1440.
Ishida, I., Engle, R.F. (2002). Modeling variance of variance: The square root, the affine, and the CEV GARCH models. Working paper. New York University.
Jacod, J., Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin.
Javaheri, A. (2005). *Inside Volatility Arbitrage; The Secrets of Skewness*. Wiley, London.
Johnson, T.C. (2002). Volatility, momentum, and time-varying skewness in foreign exchange returns. *Journal of Business and Economic Statistics* 20, 390–411.
Jones, C.S. (2003). The dynamics of stochastic volatility: Evidence from underlying and options markets. *Journal of Econometrics* 116, 181–224.
Kou, S.G. (2002). A jump-diffusion model for option pricing. *Management Science* 48, 1086–1101.
Kou, S.G., Wang, H. (2004). Option pricing under a double-exponential jump diffusion model. *Management Science* 50, 1178–1192.
Kretschmer, U., Pigorsch, C. (2004). EMM estimation of time-changed Lévy processes. Working paper. University of Bonn and University of Munich.
Küchler, U., Sørensen, M. (1997). *Exponential Families of Stochastic Processes*. Springer, New York.
Lando, D. (1998). On Cox processes and credit risky securities. *Review of Derivatives Research* 2, 99–120.

Lee, R.W. (2004). Option pricing by transform methods: Extensions, unification and error control. *Journal of Computational Finance* 7, 51–86.

Leippold, M., Wu, L. (2002). Asset pricing under the quadratic class. *Journal of Financial and Quantitative Analysis* 37, 271–295.

Lewis, A.L. (2000). *Option Valuation under Stochastic Volatility*. Finance Press, Newport Beach, CA, USA.

Lewis, A.L. (2001). A simple option formula for general jump-diffusion and other exponential Lévy processes. Working paper. Envision Financial Systems and OptionCity.net Newport Beach, California, USA.

Li, H., Wells, M., Yu, L. (2007). A Bayesian analysis of return dynamics with Lévy Jumps. *Review of Financial Studies*, in press.

Madan, D., Seneta, E. (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business* 63, 511–524.

Madan, D.B., Carr, P.P., Chang, E.C. (1998). The variance gamma process and option pricing. *European Finance Review* 2, 79–105.

Mandelbrot, B.B. (1963). The variation of certain speculative prices. *Journal of Business* 36, 394–419.

Medvedev, Scaillet, O. (2003). A simple calibration procedure of stochastic volatility models with jumps by short term asymptotics. Working paper. HEC, University of Geneva.

Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.

Mo, H., Wu, L. (2007). International capital asset pricing: Theory and evidence from index options. *Journal of Empirical Finance*, in press.

Monroe, I. (1978). Processes that can be embedded in Brownian motion. *Annals of Probability* 6, 42–56.

Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* 63, 3–50.

Pan, J., Singleton, K.J. (2005). Default and recovery implicit in the term structure of sovereign CDS spreads. Working paper. Stanford University and MIT.

Roberds, W., Whiteman, C. (1999). Endogenous term premia and anomalies in the term structure of interest rates: Explaining the predictability smile. *Journal of Monetary Economics* 44, 555–580.

Santa-Clara, P., Yan, S. (2005). Crashes, volatility, and the equity premium: Lessons from S&P 500 options. Working paper. UCLA.

Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Univ. Press, Cambridge.

Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives*. Wiley, London.

Singleton, K.J. (2001). Estimation of affine asset pricing models using the empirical characteristic function. *Journal of Econometrics* 102, 111–141.

Titchmarsh, E.C. (1986). *Introduction to the Theory of Fourier Integrals*. Chelsea Publication Co., New York.

Wan, E.A., van der Merwe, R. (2001). The unscented Kalman filter. In: Haykin, S. (Ed.), *Kalman Filtering and Neural Networks*. Wiley, New York.

Wu, L. (2005). Variance dynamics: Joint evidence from options and high-frequency returns. Working paper. Baruch College.

Wu, L. (2006). Dampened power law: Reconciling the tail behavior of financial asset returns. *Journal of Business* 79, 1445–1474.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15004-0

Chapter 4

Pricing with Wishart Risk Factors

Christian Gourieroux
CREST, CEPREMAP, and University of Toronto
E-mail: Christian.Gourieroux@ensal.fr

Razvan Sufana
York University
E-mail: rsufana@yorkn.ca

Abstract

This paper is a survey of asset pricing based on risk factors that follow a Wishart process. The general approach of pricing with Wishart risk factors is explained both in discrete time and continuous time. The approach is illustrated by an application to quadratic term structure, a multivariate extension of Heston model, an application to a structural model for credit risk, and a model for credit risk that takes into account default occurrence and loss-given-default.

1 Introduction

1.1 Factor models

Factor models are used in Finance to understand and represent the joint determination and evolution of prices (or returns) of a large number of assets. These assets may concern the stocks traded on a given stock exchange, the risk-free zero-coupon bonds at all times-to-maturity, the derivatives written on the same underlying asset, the credit derivatives corresponding to a given set of firms, or currencies and index markets in international studies. The basic models assume that the prices (or returns) depend mainly on a rather small number of driving variables, called factors, and focus on these underlying variables.

Factor models already have a long story in Finance, starting with the capital asset pricing model (Merton, 1973) and its multifactor extension by Sharpe (1964), Lintner (1965). This model is generally applied to liquid stocks, and highlights the role of the so-called market portfolio. This simple linear factor model is used for different purposes concerning stylized facts, prediction,

hedging, or creation of new financial assets. (i) This model is able to reproduce the large historical (unconditional) correlation observed between stock returns as a consequence of the common effect of the market portfolio return; (ii) By distinguishing between the market and idiosyncratic effects, and noting the weak serial dependence of the idiosyncratic terms, it allows to focus on market prediction for predicting future asset prices; (iii) The distinction between market and idiosyncratic effects is also important for hedging, since the idiosyncratic components are often diversifiable. Then, the risk on a rather diversified asset portfolio is driven by the risk on the market portfolio, that is, the risk on the factor; (iv) Finally, appropriate insurance products on the variability of market portfolio have been introduced. These are derivatives such as futures, or European calls written on market indexes.

A second generation of factor models were introduced in Finance at the end of the eighties. These models focus on the underlying risk, generally represented by a so-called underlying stochastic volatility. Factor ARCH models (Diebold and Nerlove, 1989; Engle et al., 1990) use a factor representation to capture the endogenous periods of low (respectively high) variability of asset returns, to robustify the prediction interval for future portfolio returns, and to provide dynamic Value-at-Risk (Engle and Manganelli, 2004; Gourieroux and Jasiak, 2005). As a byproduct, these models reproduce a large part of the fat tails observed on historical (unconditional) return distributions. The stochastic volatility models (Hull and White, 1987; Heston, 1993; Ball and Roma, 1994) have been introduced to get a better description of the relationship between the prices of derivatives written on the same asset. In particular, they are able to replicate a part of the smile and skewness effects observed on the implied Black–Scholes volatility associated with derivative prices.

Loosely speaking, a factor model relates asset prices (returns) to factors and idiosyncratic components. In a discrete-time framework, a dynamic factor model can be written as (Gourieroux and Jasiak, 2001):

$$p_{i,t} = g_i(F_t, F_{t-1}, \ldots; u_{i,t}, u_{i,t-1}, \ldots), \tag{1.1}$$

where i, $i = 1, \ldots, n$, indexes the asset, $p_{i,t}$ denotes the asset price, F_t the K-dimensional factor, $u_{i,t}$, $i = 1, \ldots, n$, the idiosyncratic components. The idiosyncratic errors $(u_{1,t}, \ldots, u_{n,t})$ are assumed independent, identically distributed, which ensures that the cross-sectional and serial dependence effects pass by means of the factors. The difficulty is of course to write a reliable and tractable specification of the factor model. In particular, we have to choose the number of factors, specify their dynamics, but also explain the relationships between functions g_i, $i = 1, \ldots, n$, especially when the set of assets includes several derivatives written on the same underlying product.

1.2 Observable or unobservable factors

The issue of factor observability has been addressed very early in the financial literature, when the market portfolio return was replaced by a proxy such

as a market index return (see e.g. Roll, 1977). It is important to understand the implications of factor observability, and to distinguish the information of the investors and the information of the econometrician.

Assuming optimal portfolio management by investors, their information will influence their asset demand and the asset prices at equilibrium. Financial theory generally assumes that the factor values are known by the informed investors up to current time t.

In contrast, it is preferable for the econometrician to assume a priori that the factors are not observed, for at least the following three reasons:

(i) First, if we introduce observable factors such as for instance a cycle indicator, the model is not easy to implement for prediction purposes. Indeed, the asset prices can be predicted only after having predicted the future value of this cycle indicator, which is a very difficult task.

(ii) Moreover, by specifying a priori the factors, there is a high risk of factor misspecification. It is preferable to introduce unobservable factors, in particular without any a priori interpretation as asset prices or returns (such factors are often called exogeneous factors), to try to reconstitute them from price data, and ex-post to give them an appropriate financial or physical interpretation. This is in the spirit of from-general-to-specific modeling approach.

(iii) Finally, unobservable factors are used in practice to replicate stylized facts observed with a much smaller information set. The unobserved factors in the Sharpe–Lintner specification are used to reproduce the high unconditional correlations between asset returns, that is, the correlation measured in the absence of information. The unobserved stochastic volatility is introduced in pricing models to reproduce the smile and skewness effects observed date by date, that is, with an information set which does not take into account dynamic effects. Similarly, unobservable factors are introduced in default intensity to reproduce the observed unconditional default correlation, and unobservable factors are introduced in default intensity and loss-given-default (LGD) to reproduce the unconditional dependence observed between default probabilities and expected LGD, and its evolution through the business cycle.

However, if the factors are unobservable a priori for the econometrician, some of them can be deduced ex-post from observed derivative prices up to some parameters to be estimated (see Section 3).

1.3 Factors representing risk

Factors are often introduced in Finance to capture the underlying risks and their effects on asset prices. The Wishart factor models are based on the following remark: *multidimensional risk is generally represented by means of a volatility–covolatility matrix*. Thus, factors representing the risk can be chosen

as the elements of such a matrix: $F_t = \text{vech}(Y_t)$, where Y_t denotes a (n, n) stochastic symmetric positive define matrix, and vech denotes the operator that stacks the different elements of Y_t into a vector.

When $n = 1$, we get a standard single factor model with a factor interpreted as a stochastic volatility. When $n = 2$, we get a three-factor model, where the factors are $F_{1,t} = Y_{11,t}$, $F_{2,t} = Y_{12,t}$, $F_{3,t} = Y_{22,t}$. $F_{1,t}$ and $F_{3,t}$ can be interpreted as underlying stochastic volatilities, whereas $F_{2,t}$ is an underlying stochastic covolatility. When $n = 3$, we get a six-factor model, and so on.

In this framework, the factor process (F_t) can be replaced by the matrix process (Y_t), which is a sequence of stochastic positive definite matrices. The positive definiteness condition implies nonlinear inequality restrictions on the factors. For instance, when $n = 2$, the factors are constrained by

$$Y_{11,t} > 0, \qquad Y_{11,t}Y_{22,t} - Y_{12,t}^2 > 0,$$

where the second inequality is the Cauchy–Schwarz inequality.

Generally, the effect of factors on prices is specified by means of indexes (also called scores in the credit literature), that is, linear combinations of factors. For a symmetric matrix Y_t, a linear combination of factors is easily written as $Tr(DY_t)$, where D denotes a (n, n) real symmetric matrix and Tr is the trace operator, which provides the sum of the diagonal elements of a square matrix. Indeed, we have

$$Tr(DY_t) = \sum_{i=1}^{n}(DY_t)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}Y_{ji,t} = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}Y_{ij,t}.$$

For instance, for $n = 2$, we get: $Tr(DY_t) = d_{11}Y_{11,t} + d_{22}Y_{22,t} + 2d_{12}Y_{12,t}$. In the sequel, we use this transformation on the matrix factor to express linear combinations.

Finally, we expect that these indexes, which summarize the multidimensional risk, increase with multidimensional risk. This is especially the case when such an index has to be interpreted as a risk premium. The condition can be written as

$$Tr(DY) \geqslant Tr(DY^*), \quad \text{if } Y \gg Y^*, \tag{1.2}$$

where $\gg$ denotes the standard ordering on symmetric matrices. We have the following property (see e.g. Gourieroux et al., 2007).

(i) If D is a symmetric positive semidefinite matrix, the condition (1.2) is satisfied.
(ii) In particular, if $D \gg 0$ and $Y \gg 0$, we have $Tr(DY) \geqslant 0$.

The condition (1.2) does not imply that the index reacts positively to any shock on the components of Y, when D is symmetric positive definite. Indeed, let us consider for illustration the case $n = 2$. The index is: $Tr(DY_t) = d_{11}Y_{11,t} + d_{22}Y_{22,t} + 2d_{12}Y_{12,t}$, where the matrix D satisfies the constraints: $d_{11} > 0$, $d_{22} > 0$, $d_{11}d_{22} - d_{12}^2 > 0$. A shock on $Y_{11,t}$ (respectively $Y_{22,t}$) has

a positive effect on the index, since $d_{11} > 0$ (respectively $d_{22} > 0$). However, the restrictions on D are compatible with both positive and negative values of the cross-term d_{12}. Thus, a shock on the covolatility $Y_{12,t}$ can have a positive or negative impact, depending on the sign of d_{12}. This property is useful to create positive as well as negative dependence, when some factors (here the underlying covolatilties) are integrated out.

1.4 Plan of the paper

Wishart processes are stochastic processes that represent a sequence of stochastic symmetric positive define matrices, and are good candidates for factors representing risk. They are defined in Section 2, by means of their conditional Laplace transform. They are special cases of Compound autoregressive (Car) and affine processes (for discrete and continuous time, respectively), which explains why it is easy to derive nonlinear prediction formulas at any horizon for these processes. They can be seen as multivariate extensions of the autoregressive gamma process (discrete time) (Gourieroux and Jasiak, 2006) and Cox–Ingersoll–Ross process (continuous time) (Cox et al., 1985). As for these standard processes, closed-form formulas can be derived for the conditional Laplace transform at any horizon of the future values of the process and integrated process.

In Section 3, we explain how Wishart factors can be introduced in pricing models. We first recall the pricing approach by stochastic discount factor under the historical probability, and its relationship with the equivalent approach under the risk-neutral probability. Then, we explain how to jointly introduce the Wishart factor components and the idiosyncratic components in the underlying asset price formulas and in the stochastic discount factor in order to get rather simple derivative prices. The approach is illustrated by an application to quadratic term structure, a multivariate extension of Heston model application to a structural model for credit risk, and a model for credit risk that takes into account default occurrence and loss-given-default. Section 4 concludes.

2 Wishart process

According to the problems of interest and conventions, derivative pricing can be considered in discrete, or continuous time. The pricing of European or American calls written on liquid assets is generally performed in continuous time, while for instance the determination of Credit VaR for a portfolio of retail credit or illiquid corporate bonds is done in discrete time. For this reason, the Wishart processes are presented in this section both in discrete and continuous time. Due to the time coherency condition, there exist much more Wishart processes in discrete time than in continuous time, and any discretized continuous-time Wishart process is a special case of discrete-time Wishart process.

2.1 Definition of the discrete-time Wishart process

The discrete-time Wishart process is a model for the dynamics of stochastic positive definite matrices, introduced by Gourieroux et al. (2007). The distribution of the process Y_t can be characterized by the conditional Laplace transform (moment generating function), which provides the conditional moments of exponential affine combinations of the elements of Y_t:

$$\Psi_t(\Gamma) = E\big[\exp Tr(\Gamma Y_{t+1}) \mid \underline{Y_t}\big],$$

where $\underline{Y_t}$ denotes the information (filtration) including the current and lagged values of Y, Γ is a (n, n) real symmetric matrix such that the above expectation exists. Indeed, the real Laplace transform characterizes the distribution due to the positivity of the process (Feller, 1971).

Definition 1. The Wishart autoregressive process of order one, denoted WAR(1), is a matrix Markov process (Y_t) with the following conditional Laplace transform:

$$\Psi_t(\Gamma) = \frac{\exp Tr[M'\Gamma(Id - 2\Sigma\Gamma)^{-1}MY_t]}{[\det(Id - 2\Sigma\Gamma)]^{K/2}}. \tag{2.1}$$

The transition density depends on the following parameters: K is the scalar degree of freedom strictly larger than $n-1$, M is the (n, n) matrix of autoregressive parameters, and Σ is a (n, n) symmetric positive definite matrix. The Laplace transform is defined for a matrix Γ such that $\|2\Sigma\Gamma\| < 1$, where the norm $\|\cdot\|$ is the maximal eigenvalue.

The transition density of this process is noncentered Wishart $W_n(K, M, \Sigma)$ (see Muirhead, 1982, p. 442):

$$\begin{aligned} f(Y_{t+1} \mid Y_t) = {} & \frac{1}{2^{Kn/2}} \frac{1}{\Gamma_n(K/2)} (\det \Sigma)^{-K/2} (\det Y_{t+1})^{(K-n-1)/2} \\ & \times \exp\big[-Tr\big[\Sigma^{-1}(Y_{t+1} + MY_tM')\big]/2\big] \\ & \times {}_0F_1\big(K/2, (1/4)MY_tM'Y_{t+1}\big), \end{aligned}$$

where $\Gamma_n(K/2) = \int_{A \gg 0} \exp[Tr(-A)](\det A)^{(K-n-1)/2}\, dA$ is the multidimensional gamma function, ${}_0F_1$ is the hypergeometric function of matrix argument, and the density is defined on positive definite matrices. The hypergeometric function has a series expansion:

$${}_0F_1\big(K/2, (1/4)MY_tM'Y_{t+1}\big) = \sum_{p=0}^{\infty} \sum_{l} \frac{C_l((1/4)MY_tM'Y_{t+1})}{(K/2)_l\, p!},$$

where $\sum_l$ denotes summation over all partitions $l = (p_1, \ldots, p_m)$, $p_1 \geqslant \cdots \geqslant p_m \geqslant 0$, of p into integers, $(K/2)_l$ is the generalized hypergeometric coefficient $(K/2)_l = \prod_{i=1}^{m} (K/2 - (i-1)/2)_{p_i}$, with $(a)_{p_i} = a(a+1)\cdots(a+p_i-1)$,

and $C_l((1/4)MY_tM'Y_{t+1})$ is the zonal polynomial associated with partition l. The zonal polynomials have no closed-form expressions, but can be computed recursively (see Muirhead, 1982, Chapter 7.2; and James, 1968).

The description above of the conditional Laplace transform and conditional density shows that computations based on the Laplace transform will likely be simpler than computations based on the density. In fact, nonlinear prediction and derivative pricing problems can be studied by means of the conditional Laplace transform.

The conditional log-Laplace transform is

$$\log \Psi_t(\Gamma) = -\frac{K}{2}\log\det(Id - 2\Sigma\Gamma) + Tr\big[M'\Gamma(Id - 2\Sigma\Gamma)^{-1}MY_t\big].$$

This is a linear affine function of the current value of the Wishart process. Thus, the discrete-time Wishart process is a special case of compound autoregressive (Car) processes, for which conditional moments, and more generally conditional distributions, are easily computed at any horizon (Darolles et al., 2006).

2.2 Definition of a continuous-time Wishart process

The continuous-time (n, n) Wishart process can be defined as the solution of the diffusion system (see Bru, 1989, 1991)

$$\begin{aligned} \mathrm{d}Y_t = {} & (KQQ' + AY_t + Y_tA')\,\mathrm{d}t \\ & + Y_t^{1/2}\,\mathrm{d}W_t(Q'Q)^{1/2} + (Q'Q)^{1/2}(\mathrm{d}W_t)'Y_t^{1/2}, \end{aligned} \tag{2.2}$$

where K is a scalar, A, Q are (n, n) matrices, W_t is a (n, n) matrix, whose components are independent Brownian motions, and $Y_t^{1/2}$ is the positive symmetric square root of matrix Y_t. The continuous-time Wishart process is an affine diffusion process, since both the drift and volatility are affine functions of Y (Duffie and Kan, 1996). The time discretization of the continuous-time process (2.2) is a discrete-time Wishart process with degree of freedom K, autoregressive matrix $M = \exp(Ah)$, and innovation variance $\Sigma = \int_0^h \exp(Au)QQ'[\exp(Au)]'\,\mathrm{d}u$, where h is the time step. In particular, the autoregressive matrix of the discretized process is constrained to be a matrix exponential. This shows that the class of discrete-time Wishart processes is larger than the class of continuous-time Wishart processes, which is a general result for Car and affine processes (Darolles et al., 2006; Gourieroux et al., 2007).

2.3 Conditional Laplace transform at any horizon

For expository purposes, we consider the discrete-time framework. The conditional distribution at horizon h is easily characterized, since it is a Wishart distribution with modified parameters.

Proposition 1. *Let us consider a WAR*(1) *process. The conditional distribution of* Y_{t+h} *given* Y_t *is a noncentered Wishart distribution* $W_n(K, M^h, \Sigma(h))$*, where*

$$\Sigma(h) = \Sigma + M\Sigma M' + \cdots + M^{h-1}\Sigma\big(M^{h-1}\big)'.$$

This result is the basis for an analysis of mixing properties of Wishart processes. Intuitively, the process is asymptotically stationary if $\lim_{n\to\infty} M^h = 0$, that is, if the eigenvalues of the autoregressive matrix M have a modulus strictly smaller than one. When this condition is satisfied, the stationary distribution of the process is obtained by setting $h \to \infty$, and is the centered Wishart distribution $W_n(K, 0, \Sigma(\infty))$, where $\Sigma(\infty)$ solves the equation:

$$\Sigma(\infty) = M\Sigma(\infty)M' + \Sigma.$$

2.4 Conditional Laplace transform of the integrated Wishart process

For derivative pricing, it is also useful to predict the value of integrated stochastic volatility as in the standard Hull–White formula. These nonlinear predictions are also easily derived by means of the conditional Laplace transform.

2.4.1 Discrete-time framework

Let $(Y_t) \sim W_n(K, M, \Sigma)$ be a discrete-time Wishart process. The conditional Laplace transform of the integrated process is defined by

$$\Psi_{t,h}(C, c_0, \widetilde{C}) = E_t \exp\left[\sum_{i=t+1}^{t+h} Tr(CY_i + c_0) + Tr(\widetilde{C}Y_{t+h})\right],$$

where the symmetric matrices C, $\widetilde{C}$ and the coefficient c_0 can be real or complex, whenever the expectation exists. It provides the conditional moments of exponential transforms of any future path of volatilities and integrated volatilities. Since (Y_t) is a Car process, the conditional Laplace transform of the integrated process has an exponential affine form

$$\Psi_{t,h}(C, c_0, \widetilde{C}) = \exp\big[Tr\big(B(h)Y_t\big) + b(h)\big], \tag{2.3}$$

where the symmetric matrix B and the scalar b satisfy the system of difference equations

$$B(h) = M'\big(B(h-1) + C\big)\big[Id - 2\Sigma\big(B(h-1) + C\big)\big]^{-1}M, \tag{2.4}$$

$$b(h) = b(h-1) + c_0 - 0.5K \log\det\big[Id - 2\Sigma\big(B(h-1) + C\big)\big], \tag{2.5}$$

with initial conditions: $B(0) = \widetilde{C}$, $b(0) = 0$. This system can be solved recursively.

2.4.2 Continuous-time framework

Let us now consider the continuous-time framework and assume that (Y_t) is a continuous-time Wishart process given by the solution of Eq. (2.2). The conditional Laplace transform of the integrated process is defined by

$$\Psi_{t,h}(C, c_0, \widetilde{C}) = E_t \exp\left[\int_t^{t+h} Tr(CY_u + c_0)\,\mathrm{d}u + Tr(\widetilde{C}Y_{t+h})\right] \tag{2.6}$$

where the symmetric matrices $C, \widetilde{C}$ and the coefficient c_0 can be real or complex, whenever the expectation exists. Since (Y_t) is an affine continuous-time process, the conditional Laplace transform has an exponential-affine form (Duffie et al., 2003)

$$\Psi_{t,h}(C, c_0, \widetilde{C}) = \exp\big[Tr\big(B(h)Y_t\big) + b(h)\big], \tag{2.7}$$

where the symmetric matrix B and the scalar b satisfy the system of Riccati equations

$$\frac{\mathrm{d}B(h)}{\mathrm{d}h} = B(h)A + A'B(h) + 2B(h)Q'QB(h) + C, \tag{2.8}$$

$$\frac{\mathrm{d}b(h)}{\mathrm{d}h} = KTr\big[B(h)QQ'\big] + c_0, \tag{2.9}$$

with initial conditions: $B(0) = \widetilde{C}$, $b(0) = 0$.

In general, such a system of Riccati equations cannot be solved explicitly. However, this is possible for Wishart processes (see Gourieroux and Sufana, 2005 and Fonseca et al., 2005).

Proposition 2. *The solution for coefficient $B(h)$ is*

$$\begin{aligned} B(h) = B^* &+ \exp\big[(A + 2Q'QB^*)h\big]' \\ &\times \left\{(\widetilde{C} - B^*)^{-1} - 2\int_0^h \exp\big[(A + 2Q'QB^*)u\big]Q'Q\right. \\ &\left.\times \exp\big[(A + 2Q'QB^*)u\big]'\,\mathrm{d}u\right\}^{-1} \exp\big[(A + 2Q'QB^*)h\big], \end{aligned}$$

where B^* is a symmetric matrix which satisfies

$$A'B^* + B^*A + 2B^*Q'QB^* + C = 0.$$

The solution for $b(h)$ is

$$b(h) = KTr\left[\int_0^h B(u)\,\mathrm{d}uQQ'\right] + c_0h.$$

The existence of a symmetric matrix solution B^* of the implicit equation in Proposition 2 implies restrictions on matrix C. For instance, matrix C has to be smaller than $A'(2Q'Q)^{-1}A$, according to the standard ordering on symmetric matrices. Note also that the expression of $B(h)$ admits a closed form. Indeed, up to an appropriate change of basis, the matrix $\exp[(A + 2Q'QB^*)u]$ can be written under a diagonal or triangular form, whose elements are simple functions of u (exponential or exponential times polynomial), which can be explicitly integrated between 0 and h.

3 Pricing

3.1 Pricing with stochastic discount factor

In this section, we briefly recall the approach of derivative pricing by stochastic discount factor, both in discrete and continuous time. Then, we compare both approaches. We focus on European derivatives.

3.1.1 Pricing in discrete time

Let us consider the pricing at time t of a European derivative with payoff $g(\underline{F_{t+h}}, u_{t+h})$ at time $t + h$, where $\underline{F_t}$ is the set of current and lagged values of the common risk factors, $\underline{F_t} = \{F_t, F_{t-1}, \ldots\}$, and u_t is the current value of the idiosyncratic risk factor. The approach specifies the dynamic properties of the underlying risk factors under the actual (historical) probability. The no-arbitrage assumption implies the existence of a strictly positive stochastic discount factor $M_{t,t+1}$, for period $(t, t + 1)$, which summarizes the one-period discounting with respect to both time and uncertainty.

In the sequel, we assume that:

Assumption A.1: the risk factor processes (F_t) and (u_t) are independent under the actual probability,

Assumption A.2: (u_t) is a m-dimensional white noise, and

Assumption A.3: the stochastic discount factor depends on the common risk factors only: $M_{t,t+1} = M_{t,t+1}(\underline{F_{t+1}})$.

Then, the price $P_t(g)$ at time t of the European derivative is

$$P_t(g) = E_t\big[M_{t,t+1}(\underline{F_{t+1}}) \cdots M_{t+h-1,t+h}(\underline{F_{t+h}})g(\underline{F_{t+h}}, u_{t+h})\big],$$

where E_t denotes the expectation conditional on the information set $I_t = \{\underline{F_t}, \underline{u_t}\}$ at time t. Under Assumptions A.1–A.3, the pricing formula can be simplified. More precisely, by applying the law of iterated expectations, we get, for $h \geqslant 1$,

$$\begin{aligned} P_t(g) &= E_t\big[M_{t,t+1}(\underline{F_{t+1}}) \cdots M_{t+h-1,t+h}(\underline{F_{t+h}}) \\ &\qquad \times E\big[g(\underline{F_{t+h}}, u_{t+h}) \mid \underline{F_{t+h}}, u_t\big]\big] \\ &= E_t\big[M_{t,t+1}(\underline{F_{t+1}}) \cdots M_{t+h-1,t+h}(\underline{F_{t+h}})\end{aligned}$$

$$\times E\big[g(\underline{F_{t+h}}, u_{t+h}) \mid \underline{F_{t+h}}\,\big]\big]$$
$$= E_t\big[M_{t,t+1}(\underline{F_{t+1}}) \cdots M_{t+h-1,t+h}(\underline{F_{t+h}})h(\underline{F_{t+h}})\big], \tag{3.1}$$

where $h(\underline{F_{t+h}}) = E[g(\underline{F_{t+h}}, u_{t+h}) \mid \underline{F_{t+h}}]$. From Assumption A.1, we finally deduce that

$$P_t(g) = E\big[M_{t,t+1}(\underline{F_{t+1}}) \cdots M_{t+h-1,t+h}(\underline{F_{t+h}})h(\underline{F_{t+h}}) \mid \underline{F_t}\,\big]. \tag{3.2}$$

Thus, pricing European derivatives written on $(\underline{F_{t+h}}, u_{t+h})$ is equivalent to pricing European derivatives written on $\underline{F_{t+h}}$ only, using only the information on common factors.

3.1.2 Pricing in continuous time

The pricing approach is similar in a continuous-time framework. $\underline{F_t}$ denotes the set of all current and past values of the common risk factors, and the horizon h is a real positive number. We again make Assumptions A.1, A.2 in Section 3.1.1, and assume that

Assumption A.3′ the stochastic discount factor for period $(t, t+h)$ is of the form

$$M_{t,t+h} = \exp\left(\int_t^{t+h} m(F_u)\,\mathrm{d}u\right),$$

where m is a real function of F_u.

Then, the price $P_t(g)$ at time t of the European derivative with payoff $g(\underline{F_{t+h}}, u_{t+h})$ at time $t+h$, is

$$P_t(g) = E_t\big[M_{t,t+h}g(\underline{F_{t+h}}, u_{t+h})\big].$$

Under Assumptions A.1, A.2, A.3′, computations similar to those in Section 3.1.1 lead to the same result that pricing European derivatives written on $(\underline{F_{t+h}}, u_{t+h})$ is equivalent to pricing European derivatives written on $\underline{F_{t+h}}$ only

$$P_t(g) = E\big[M_{t,t+h}h(\underline{F_{t+h}}) \mid \underline{F_t}\,\big].$$

3.2 Pricing with Wishart factors

3.2.1 Pricing in discrete time

Let us consider a discrete-time Wishart process (Y_t), and assume that the stochastic discount factor $M_{t,t+1}$ has an exponential-affine form in the elements of Y_{t+1}

$$M_{t,t+1} = \exp\big[Tr(DY_{t+1}) + d\big],$$

where D is a (n, n) symmetric matrix, and d is a scalar. Particularly interesting is the special case of a European derivative with an exponential-affine payoff

written on the factors:

$$h(\underline{Y_{t+h}}) = \exp Tr(GY_{t+h}), \tag{3.3}$$

where G is a (n, n) symmetric matrix.

The pricing formula (3.2) implies that the price of this European derivative is

$$P_t(h) = E_t \exp\left[\sum_{i=t+1}^{t+h} Tr(DY_i + d) + Tr(GY_{t+h})\right] = \Psi_{t,h}(D, d, G).$$

Thus, this price is equal to a conditional Laplace transform of the integrated Wishart process calculated in Section 2.4.1. Prices of other derivatives, like for example European call options, are obtained by using the transform analysis (Duffie et al., 2000), that is by inverting the conditional Laplace transform written on imaginary arguments (Fourier transform).

3.2.2 Pricing in continuous time

Let us now consider a continuous-time Wishart process (Y_t), and assume that the stochastic discount factor is affine

$$m(u) = Tr(DY_u) + d, \tag{3.4}$$

where D is a (n, n) symmetric matrix, and d is a scalar. The price $P_t(h)$ of a security with payoff $h(\underline{Y_{t+h}}) = \exp Tr(GY_{t+h})$ at time $t + h$ is

$$P_t(h) = E_t\left[\exp\left(\int_t^{t+h} \big(Tr(DY_u) + d\big)\,\mathrm{d}u + Tr(GY_{t+h})\right)\right]. \tag{3.5}$$

This is a conditional Laplace transform of the integrated process and, as shown in Section 2.4.2, this price can be computed in closed form.

3.2.3 The link between discrete- and continuous-time pricing

It is interesting to relate the pricing in continuous and discrete time. For expository purposes, let us consider the continuous-time pricing of a European call with time to maturity $h = 2$. The price of this derivative is

$$\begin{aligned} P_t(\tau) &= E_t \exp\left[\int_t^{t+2} Tr(DY_u + d)\,\mathrm{d}u + Tr(GY_{t+2})\right] \\ &= E_t\left[\exp\left(\int_t^{t+1} m(u)\,\mathrm{d}u\right)\exp\left(\int_{t+1}^{t+2} m(u)\,\mathrm{d}u\right)\right. \\ &\quad \left.\times \exp Tr(GY_{t+2})\right]. \end{aligned}$$

Let us denote $\overline{Y_t}$ the information generated by the current and future values of Y. We can write

$$P_t(\tau) = E_t\left[\exp\left(\int_t^{t+1} m(u)\,\mathrm{d}u\right)\right.$$
$$\left.\times E\left(\exp\left(\int_{t+1}^{t+2} m(u)\,\mathrm{d}u\right)\middle|\underline{Y_{t+1}}, \overline{Y_{t+2}}\right)\exp Tr(GY_{t+2})\right]$$
$$= E_t\big[\widetilde{M}_{t,t+1}\widetilde{M}_{t+1,t+2}\exp Tr(GY_{t+2})\big],$$

where

$$\widetilde{M}_{t,t+1} = E\left[\exp\left(\int_t^{t+1} m(u)\,\mathrm{d}u\right)\middle|\underline{Y_t}, \overline{Y_{t+1}}\right].$$

Thus, when pricing derivatives with discrete time to maturity, it is possible to replace the continuous-time stochastic discount factor m by the discrete-time stochastic discount factor $\widetilde{M}$. $\widetilde{M}_{t,t+1}$ is a conditional Laplace transform of the integrated Wishart process given both the past and future values of the process. It is easily checked that $\widetilde{M}_{t,t+1}$ is also an exponential-affine function of the closest past and future values

$$\widetilde{M}_{t,t+1} = \exp\big[Tr(D_0 Y_t) + Tr(D_0 Y_{t+1}) + d\big].$$

We deduce that continuous-time pricing based on a Wishart process is equivalent to discrete-time pricing based on the Wishart process, once lagged values of the Wishart process are introduced in the discrete-time stochastic discount factor. Simple pricing formulas also exist with this introduction of lagged volatility in the stochastic discount factor $\widetilde{M}$.

4 Examples

The aim of this section is to discuss various examples of derivative pricing in order to illustrate the flexibility and importance of Wishart risk factor models. The presentation can be equivalently performed in discrete or continuous time.

4.1 Wishart quadratic term structure

Let $D(t, h)$ denote the price at time t of a zero-coupon bond that pays \$1 at time $t + h$. We assume that bond prices depend on some underlying stochastic factors that are the elements of a (n, n) matrix Wishart process Y_t. From Sections 3.1.1 and 3.2.1, the bond price $D(t, h)$ is

$$D(t,h) = E_t\big[M_{t,t+1}(Y_{t+1})\cdots M_{t+h-1,t+h}(Y_{t+h})\big]$$
$$= E_t \exp\left[\sum_{i=t+1}^{t+h} Tr(DY_i + d)\right] = \Psi_{t,h}(D,d,0).$$

As shown in Section 2.4.1, the conditional Laplace transform of the integrated Wishart process can be computed as

$$D(t,h) = \exp\big[Tr\big(B(h)Y_t\big) + b(h)\big], \tag{4.1}$$

where the symmetric matrix B and the scalar b satisfy the system of difference equations, for $h \geqslant 1$

$$B(h) = M'\big(B(h-1)+D\big)\big[Id - 2\Sigma\big(B(h-1)+D\big)\big]^{-1}M, \tag{4.2}$$
$$b(h) = b(h-1) + d - 0.5K\log\det\big[Id - 2\Sigma\big(B(h-1)+D\big)\big], \tag{4.3}$$

with initial conditions: $B(0) = 0$, $b(0) = 0$. The resulting term structure is called Wishart quadratic term structure (see Gourieroux and Sufana, 2003).

The reason for this terminology is the existence of a particular interpretation of the Wishart process, when the degree of freedom is integer. When K is integer, the Wishart process (Y_t) can be expressed as the sum of squares of K independent Gaussian vector autoregressions of order 1 (VAR(1)) with identical dynamics:

$$Y_t = \sum_{k=1}^{K} x_{kt}x'_{kt}, \tag{4.4}$$

where the n-dimensional vector x_{kt} follows:

$$x_{k,t+1} = Mx_{k,t} + \varepsilon_{k,t+1},\ \varepsilon_{k,t+1} \sim N(0,\Sigma), \tag{4.5}$$

for $k = 1,\ldots,K$. The sum of squares in Eq. (4.4) is a.s. positive definite, if $K \geqslant n$. With this interpretation of the Wishart process, the bond prices become

$$D(t,h) = \exp\left[Tr\left(B(h)\sum_{k=1}^{K} x_{kt}x'_{kt}\right) + b(h)\right]$$
$$= \exp\left[\sum_{k=1}^{K} Tr\big(B(h)x_{kt}x'_{kt}\big) + b(h)\right]$$
$$= \exp\left[\sum_{k=1}^{K} Tr\big(x'_{kt}B(h)x_{kt}\big) + b(h)\right]$$
$$= \exp\left[\sum_{k=1}^{K} x'_{kt}B(h)x_{kt} + b(h)\right],$$

since we can commute within the trace operator. Thus, the bond yields $r(t,h) = -(\log D(t,h))/h$ are linear affine with respect to Y_t, but quadratic in the vectors x_{kt}, $k = 1, \ldots, K$. In fact, the Wishart quadratic term structure is an extension of the standard quadratic term structure, which corresponds to $K = 1$ (see Ahn et al., 2002; Leippold and Wu, 2002, for the standard quadratic term structure model). The quadratic term structure models are very tractable since they are special cases of affine term structure models.

An important property of the Wishart quadratic term structure is related to the positivity of bond yields.

Proposition 3. *If the symmetric matrix D is negative semidefinite and $d \leqslant 0.5K \log\det(Id - 2\Sigma D)$, then, for $h \geqslant 1$,*

(i) *The domain for the yield $r(t,h)$ is $[-b(h)/h, \infty)$;*
(ii) *The lower bound $-b(h)/h$ is nonnegative and increases with h.*

As a consequence, under the parameter restrictions of Proposition 3, the bond yields are positive at all maturities. The fact that the lower bound of the domain for interest rates increases with the time-to-maturity is not a consequence of the Wishart assumption, but is a consequence of no-arbitrage opportunity, as shown in Gourieroux and Monfort (2005).

4.2 Extension of Heston's model

The Wishart process can be used to extend the stochastic volatility model introduced by Heston (1993) (see Gourieroux and Sufana, 2005). Let us consider n risky assets whose prices are the components of the n-dimensional vector S_t, and let Σ_t denote the volatility matrix of the infinitesimal geometric returns $d\log S_t$ of the risky assets. We assume that the joint dynamics of $\log S_t$ and Σ_t is given by the stochastic differential system

$$\mathrm{d}\log S_t = \left[\mu + \begin{pmatrix} Tr(D_1\Sigma_t) \\ \vdots \\ Tr(D_n\Sigma_t) \end{pmatrix}\right]\mathrm{d}t + \Sigma_t^{1/2}\,\mathrm{d}W_t^S, \tag{4.6}$$

$$\begin{aligned}\mathrm{d}\Sigma_t &= (KQQ' + A\Sigma_t + \Sigma_t A')\,\mathrm{d}t \\ &\quad + \Sigma_t^{1/2}\,\mathrm{d}W_t^{\sigma}(Q'Q)^{1/2} + (Q'Q)^{1/2}\big(\mathrm{d}W_t^{\sigma}\big)'\Sigma_t^{1/2},\end{aligned} \tag{4.7}$$

where W_t^S and W_t^{σ} are a n-dimensional vector and a (n,n) matrix, respectively, whose elements are independent unidimensional standard Brownian motions, μ is a constant n-dimensional vector, and D_i, $i = 1, \ldots, n$, A, Q are (n,n) matrices with Q invertible. The price equation includes a volatility-in-mean effect to account for a dynamic risk premium, and to capture the tendency for volatility and stock price to move together even without assuming an instantaneous correlation between the stock return and volatility innovations. However, the

model can be extended to include such a multivariate correlation (Fonseca et al., 2005).

If $n = 1$, the differential system reduces to

$$\begin{aligned}
&\mathrm{d}\log S_t = (\mu + D_1\Sigma_t)\,\mathrm{d}t + \sqrt{\Sigma_t}\,\mathrm{d}W_t^S,\\
&\mathrm{d}(\Sigma_t) = \left(KQ^2 + 2A\Sigma_t\right)\mathrm{d}t + 2Q\sqrt{\Sigma_t}\,\mathrm{d}W_t^\sigma,
\end{aligned}$$

with a Cox–Ingersoll–Ross specification for the stochastic volatility (Cox et al., 1985). Thus, the multivariate model in Equations (4.6)–(4.7) reduces to Heston's specification (see Heston, 1993; Ball and Roma, 1994).

The multivariate model is an affine process since it admits drift and volatility functions that are affine functions of $\log S_t$ and Σ_t (see Duffie and Kan, 1996; Duffie et al., 2003). Therefore, the theory of affine processes can be used to derive the conditional Laplace transform of the joint process $(\log S_t, \Sigma_t)$ and of its integrated values, which is defined by

$$\begin{aligned}
\Psi_{t,h}(\gamma, \gamma_0, \widetilde{\gamma}, C, c_0, \widetilde{C}) = E_t \exp\Bigg[&\int_t^{t+h} (\gamma' \log S_u + \gamma_0)\,\mathrm{d}u + \widetilde{\gamma}' \log S_{t+h}\\
&+ \int_t^{t+h} Tr(C\Sigma_u + c_0)\,\mathrm{d}u + Tr(\widetilde{C}\Sigma_{t+h})\Bigg].
\end{aligned}$$

This Laplace transform is the basis for pricing derivatives by transform analysis (Duffie et al., 2000).

Proposition 4. *The conditional Laplace transform of the joint process* $(\log S_t, \Sigma_t)$ *is*

$$\Psi_{t,h}(\gamma, \gamma_0, \widetilde{\gamma}, C, c_0, \widetilde{C}) = \exp\left[a(h)' \log S_t + Tr\left(B(h)\Sigma_t\right) + b(h)\right], \tag{4.8}$$

where a, b, and the symmetric matrix B satisfy the system of Riccati equations

$$\frac{\mathrm{d}a(h)}{\mathrm{d}h} = \gamma, \tag{4.9}$$

$$\begin{aligned}
\frac{\mathrm{d}B(h)}{\mathrm{d}h} = {} & B(h)A + A'B(h) + 2B(h)Q'QB(h)\\
&+ \frac{1}{2}a(h)a(h)' + \sum_{i=1}^{n} a_i(h)D_i + C,
\end{aligned} \tag{4.10}$$

$$\frac{\mathrm{d}b(h)}{\mathrm{d}h} = a(h)'\mu + KTr\left[B(h)QQ'\right] + \gamma_0 + c_0, \tag{4.11}$$

with initial conditions: $a(0) = \widetilde{\gamma}$, $B(0) = \widetilde{C}$, $b(0) = 0$.

The differential equation for a admits the explicit solution

$$a(h) = \gamma h + \widetilde{\gamma}.$$

The remaining equations admit a closed-form solution under some parameter restrictions.

Proposition 5. *For $\gamma = 0$, we get*

$$\begin{aligned} B(h) = B^* &+ \exp\bigl[(A + 2Q'QB^*)h\bigr]' \\ &\times \left\{ (\widetilde{C} - B^*)^{-1} - 2\int_0^h \exp\bigl[(A + 2Q'QB^*)u\bigr] Q'Q \right. \\ &\left. \times \exp\bigl[(A + 2Q'QB^*)u\bigr]' \,\mathrm{d}u \right\}^{-1} \exp\bigl[(A + 2Q'QB^*)h\bigr], \end{aligned}$$

where B^* is a symmetric matrix which satisfies

$$A'B^* + B^*A + 2B^*Q'QB^* + \frac{1}{2}\widetilde{\gamma}\widetilde{\gamma}' + \sum_{i=1}^{n} \widetilde{\gamma}_i D_i + C = 0.$$

The closed-form solution for $b(h)$ is deduced from the third differential equation

$$b(h) = (\widetilde{\gamma}'\mu + \gamma_0 + c_0)h + KTr\left[\int_0^h B(u)\,\mathrm{d}u QQ' \right].$$

4.3 Multifactor extension of Merton's model

An application of the multivariate stochastic volatility model in Section 4.2 is the extension of the credit risk model proposed by Merton (1974) to a framework with stochastic volatility, stochastic firm's debt and more than one firm (see Gourieroux and Sufana, 2005). In Merton's basic structural approach, a firm's asset value is assumed to follow a geometric Brownian motion under the risk-neutral probability, and the debt amount L and time-to-default are assumed predetermined. Under these assumptions, the firm's equity is a call option on the asset value with a strike equal to the debt level, and its price is computed from the Black–Scholes formula.

Let us now consider n firms indexed by $i = 1, \ldots, n$. The extended structural model assumes that the joint dynamics of firm i's asset value $A_{i,t}$ and liability value $L_{i,t}$ are represented by

$$\begin{pmatrix} \mathrm{d}\log A_{i,t} \\ \mathrm{d}\log L_{i,t} \end{pmatrix} = \begin{bmatrix} \mu_{A,i} + Tr(D_{A,i}\Sigma_{i,t}) \\ \mu_{L,i} + Tr(D_{L,i}\Sigma_{i,t}) \end{bmatrix} \mathrm{d}t + \Sigma_{i,t}^{1/2}\,\mathrm{d}W_{i,t}^{S},$$

where $\Sigma_{i,t}$, $i = 1, \ldots, n$, are Wishart processes. The n Wishart processes can be independent, or some firms can be driven by the same Wishart process.

The extended structural model for each firm is a bivariate specification of the multivariate model in Section 4.2. However, the terms $Tr(D_A\Sigma_t)$ and $Tr(D_L\Sigma_t)$ included in the drift do not admit an interpretation as risk premia, since the firm's asset value and liability value are not traded on a market. They capture the tendency of the firm's asset and liability values to move together in response to an increase in volatility, in a situation when the asset value has a mean higher than L, but a large volatility increase causes the asset value to approach the liability level. In this case, the medium-term rating of the firm allows it to increase its debt in order to stimulate its investments, and as a result increase the asset value. Prices of various credit derivatives can be computed as explained in Section 4.2.

4.4 Joint modeling of default intensity and recovery rates

As emphasized by the Basel Committee on Banking Supervision (2005), the credit risk models have to allow for "realized recovery rates to be lower than average during terms of high default rates." Thus, it is important to develop models sufficiently flexible to allow for negative (respectively positive) dependence between the default intensity and recovery rate (respectively loss-given-default). This section shows that the models with Wishart risk factors have this flexibility. The results are based on Gourieroux et al. (2006).

Let us consider the following specification for the default probability DP and loss-given-default LGD:

$$DP = \exp\big[-Tr(AY)\big] = \exp\big[-(a_{11}Y_{11} + 2a_{12}Y_{12} + a_{22}Y_{22})\big],$$
$$LGD = \exp\big[-Tr(BY)\big] = \exp\big[-(b_{11}Y_{11} + 2b_{12}Y_{12} + b_{22}Y_{22})\big],$$

where

$$Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{12} & Y_{22} \end{pmatrix}, \qquad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix},$$

are symmetric positive definite matrices, and Y follows the marginal (invariant) distribution of a Wishart process, with Laplace transform:

$$E\exp\big[-Tr(AY)\big] = \big[\det(Id + 2A)\big]^{-K/2}.$$

Since A and B are symmetric positive definite matrices, DP and LGD are between 0 and 1, and the elements of A and B are constrained by

$$a_{11} > 0, \qquad a_{11}a_{22} - a_{12}^2 > 0, \qquad b_{11} > 0, \qquad b_{11}b_{22} - b_{12}^2 > 0.$$

In particular, these restrictions are compatible with opposite signs of a_{12} and b_{12}. In such a case, a shock on the stochastic covolatility Y_{12} (or equivalently, on the stochastic correlation), for given volatilities, implies opposite effects on DP and LGD. Thus, we get a negative dependence, for given volatilities.

Let us now show that, unconditionally, the effects of shocks on the stochastic covolatility Y_{12} cannot dominate the effects due to shocks on Y_{11} and Y_{22}.

More precisely, the negative dependence, given the volatilities, is compatible with a positive unconditional dependence. For expository purposes, let us assume $K = 2$. We get

$$\begin{aligned}\operatorname{Cov}[DP, LGD] &= \operatorname{Cov}\big[\exp\big[-Tr(AY)\big], \exp\big[-Tr(BY)\big]\big] \\ &= E\exp\big[-Tr\big[(A+B)Y\big]\big] \\ &\quad - E\exp\big[-Tr(AY)\big]E\exp\big[-Tr(BY)\big] \\ &= \frac{1}{\det(Id+2(A+B))} - \frac{1}{\det(Id+2A)}\frac{1}{\det(Id+2B)}.\end{aligned}$$

The covariance is positive if and only if

$$\Delta = \det(Id+2A)\det(Id+2B) - \det\big(Id+2(A+B)\big) > 0.$$

Without loss of generality, we can assume $a_{12} = 0$. We get

$$\begin{aligned}\Delta &= 8\big(b_{11}b_{22} - b_{12}^2\big)(a_{11}+a_{22}+2a_{11}a_{22}) \\ &\quad + 4\big(a_{11}b_{11} + a_{22}b_{22} + 2a_{11}a_{22}(b_{11}+b_{22})\big) > 0,\end{aligned}$$

by the positive definiteness of matrices A and B.

5 Concluding remarks

The aim of this survey was to explain why Wishart risk factors are good candidates for representing the underlying multivariate risk, that is stochastic volatility–covolatility matrices. The affine (Car) properties of the Wishart processes explain the closed-form expression of the Laplace transform, and then of derivative prices at any horizon. The examples have shown the flexibility of Wishart factor models in studying various types of derivatives. In fact, an advantage of such models is the possibility of developing coherent approaches. For instance, Wishart factor models can be introduced in a coherent way for both historical and risk-neutral approaches. In credit risk analysis, this allows to price credit derivatives (risk-neutral approach) and jointly compute the Credit VaR or study the default occurrence (historical approach). The Wishart risk factor models are also appropriate for coherent analysis of asset prices across countries, for instance, to understand how the evolution of exchange rates influences the term structure patterns in two countries (Gourieroux et al., 2005, 2006).

References

Ahn, D., Dittmar, R., Gallant, A. (2002). Quadratic term structure models: Theory and evidence. *Review of Financial Studies* 15, 243–288.

Ball, C., Roma, A. (1994). Stochastic volatility option pricing. *Journal of Financial and Quantitative Analysis* 29, 589–607.

Basel Committee on Banking Supervision (2005). Guidance on Paragraph 468 of the framework document. Bank for International Settlements, July.
Bru, M. (1989). Diffusions of perturbed principal component analysis. *Journal of Multivariate Analysis* 29, 127–136.
Bru, M. (1991). Wishart processes. *Journal of Theoretical Probability* 4, 725–751.
Cox, J., Ingersoll, J., Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica* 53, 385–407.
Darolles, S., Gourieroux, C., Jasiak, J. (2006). Structural Laplace transform and compound autoregressive models. *Journal of Time Series Analysis* 27, 477–503.
Diebold, F., Nerlove, M. (1989). The dynamics of exchange rate volatility: A multivariate latent factor ARCH model. *Journal of Applied Econometrics* 4, 1–22.
Duffie, D., Kan, R. (1996). A yield factor model of interest rates. *Mathematical Finance* 6, 379–406.
Duffie, D., Pan, J., Singleton, K. (2000). Transform analysis and asset pricing for affine jump diffusions. *Econometrica* 68, 1343–1376.
Duffie, D., Filipovic, D., Schachermayer, W. (2003). Affine processes and applications in finance. *Annals of Applied Probability* 13, 984–1053.
Engle, R., Manganelli, S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles. *Journal of Business and Economic Statistics* 22, 367–381.
Engle, R., Ng, V., Rothschild, M. (1990). Asset pricing with a factor ARCH covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics* 45, 213–237.
Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. John Wiley, New York.
Fonseca, J., Grasselli, M., Tebaldi, C. (2005). Wishart multi-dimensional stochastic volatility. *Working paper*.
Gourieroux, C., Jasiak, J. (2001). Dynamic factor models. *Econometric Reviews* 20, 385–424.
Gourieroux, C., Jasiak, J. (2005). Dynamic quantile models. *Working paper*, CREST.
Gourieroux, C., Jasiak, J. (2006). Autoregressive gamma processes. *Journal of Forecasting* 25, 129–152.
Gourieroux, C., Monfort, A. (2005). Domain restrictions on interest rates implied by no arbitrage. *Working paper*, CREST.
Gourieroux, C., Sufana, R. (2003). Wishart quadratic term structure models. *Working paper*, University of Toronto.
Gourieroux, C., Sufana, R. (2005). Derivative pricing with multivariate stochastic volatility: Application to credit risk. *Working paper*, CREST.
Gourieroux, C., Monfort, A., Sufana, R. (2005). International money and stock market contingent claims. *Working paper*, CREST.
Gourieroux, C., Monfort, A., Polimenis, V. (2006). Affine models for credit risk analysis. *Journal of Financial Econometrics* 4, 494–530.
Gourieroux, C., Jasiak, J., Sufana, R. (2007). The Wishart autoregressive process of multivariate stochastic volatility. *Journal of Econometrics*, in press.
Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6, 327–343.
Hull, J., White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 281–300.
James, A. (1968). Calculation of zonal polynomial coefficients by use of the Laplace–Beltrami operator. *Annals of Mathematical Statistics* 39, 1711–1718.
Leippold, M., Wu, L. (2002). Asset pricing under the quadratic class. *Journal of Financial and Quantitative Analysis* 37, 271–295.
Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 13–37.
Merton, R. (1973). An intertemporal capital asset pricing model. *Econometrica* 41, 867–887.
Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449–470.
Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley, New York.
Roll, R. (1977). A critique of the asset pricing theory's tests. *Journal of Financial Economics* 4, 129–176.
Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425–442.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15005-2

Chapter 5
Volatility

Federico M. Bandi
Graduate School of Business, The University of Chicago
E-mail: federico.bandi@chicagogsb.edu

Jeffrey R. Russell
Graduate School of Business, The University of Chicago
E-mail: jeffrey.russell@chicagogsb.edu

Abstract

We provide a unified framework to understand current advances in two important fields in empirical finance: volatility estimation by virtue of microstructure noise-contaminated asset price data and transaction cost evaluation. In this framework, we review recently-proposed identification procedures relying on the unique possibilities furnished by asset price data sampled at high frequency. While discussing these procedures, we offer our perspective on the existing methods and findings, as well as on directions for future work.

Keywords: High-frequency data; Realized volatility; Market microstructure noise; Transaction cost; Volatility and asset pricing; Liquidity and asset pricing

1 Introduction

Recorded asset prices deviate from their equilibrium values due to the presence of market microstructure frictions. Hence, the volatility of the observed prices depends on two distinct volatility components, i.e., the volatility of the *unobserved* frictionless equilibrium prices (henceforth equilibrium prices) and the volatility of the equally *unobserved* market microstructure effects.

In keeping with this basic premise, this review starts from a model of price formation that allows for empirically relevant market microstructure effects to discuss current advances in the nonparametric estimation of both volatility notions using high-frequency asset price data.

Numerous insightful reviews have been written on volatility. The existing reviews concentrate on work that assumes observability of the equilibrium price

and study its volatility properties in the absence of measurement error (see, e.g., Andersen et al., 2002, and the references therein). Reviews have also been written on work that solely focuses on the measurement error and characterizes it in terms of frictions induced by the market's fine grain dynamics (see, e.g., Hasbrouck, 1996, and Stoll, 2000). Quantifying these frictions is of crucial importance to understand and measure the effective execution cost of trades. More recently, Barndorff-Nielsen and Shephard, 2007 have provided a discussion of current research on alternative nonparametric volatility estimators. While their review largely focuses on the frictionless case, it also offers interesting perspectives on the empirically more relevant case of equilibrium prices affected by market microstructure effects (see, also, McAleer and Medeiros, 2007).

The present review places emphasis on the volatilities of both unobserved components of a recorded price, i.e., equilibrium price and microstructure frictions. Specifically, our aim is to provide a unified framework to understand current advances in two important fields in empirical finance, namely equilibrium price volatility estimation and transaction cost evaluation. To this extent, we begin with a general price formation mechanism that expresses recorded (logarithmic) asset prices as the sum of (logarithmic) equilibrium prices and (logarithmic) market microstructure effects.

2 A model of price formation with microstructure effects

Write an observed logarithmic price as

$$p = p^* + \eta, \tag{1}$$

where p^* denotes the logarithmic equilibrium price, i.e., the price that would prevail in the absence of market microstructure frictions,[1] and η denotes a market microstructure contamination in the observed logarithmic price as induced by price discreteness and bid–ask bounce effects, for instance (see, e.g., Stoll, 2000). Fix a certain time period h (a trading day, for example) and assume availability of M equispaced high-frequency prices over h. Given Eq. (1) we can readily define continuously-compounded returns over any intra-period interval of length $\delta = \frac{h}{M}$ and write

$$\underbrace{p_{j\delta} - p_{(j-1)\delta}}_{r_{j\delta}} = \underbrace{p^*_{j\delta} - p^*_{(j-1)\delta}}_{r^*_{j\delta}} + \underbrace{\eta_{j\delta} - \eta_{(j-1)\delta}}_{\varepsilon_{j\delta}}. \tag{2}$$

The following assumptions are imposed on the equilibrium price process and market microstructure effects.

[1] We start by being deliberately unspecific about the nature of the equilibrium price. We will add more economic structure to the model when discussing transaction cost evaluation (Section 7).

Assumption 1 *(The equilibrium price process)*.

(1) The logarithmic equilibrium price process p_t^* is a continuous stochastic volatility semimartingale. Specifically,

$$p_t^* = \alpha_t + m_t, \tag{3}$$

where α_t (with $\alpha_0 = 0$) is a predictable drift process of finite variation and m_t is a continuous local martingale defined as $\int_0^t \sigma_s \, \mathrm{d}W_s$, with $\{W_t\colon\ t \geqslant 0\}$ denoting standard Brownian motion.

(2) The spot volatility process σ_t is càdlàg and bounded away from zero.

(3) The process $\int_0^t \sigma_s^4 \, \mathrm{d}s$ is bounded almost surely for all $t < \infty$.

Assumption 2 *(The microstructure frictions)*.

(1) The microstructure frictions in the price process η have mean zero and are covariance-stationary with joint density $f_M(.)$.

(2) The variance of $\varepsilon_{j\delta} = \eta_{j\delta} - \eta_{(j-1)\delta}$ is $O(1)$ for all δ.

(3) The η's are independent of the p^*'s ($\eta \perp\!\!\!\perp p^*$).

In agreement with classical asset-pricing theory in continuous time (see, e.g., Duffie, 1992), Assumption 1 implies that the equilibrium return process evolves in time as a stochastic volatility martingale difference plus an adapted process of finite variation. The stochastic spot volatility can display jumps, diurnal effects, high-persistence (possibly of the long-memory type), and non-stationarities. Leverage effects (i.e., dependence between σ and the Brownian motion W) are allowed.

Assumption 2 permits general dependence features for the microstructure friction components in the recorded prices. The correlation structure of the frictions can, for instance, capture first-order negative dependence in the recorded high-frequency returns as determined by bid–ask bounce effects (see, e.g., Roll, 1984) as well as higher order dependence as induced by clustering in order flow, for example. In general, the characteristics of the noise returns ε may be a function of the sampling frequency $\delta = \frac{h}{M}$. The joint density of the η's has a subscript M to make this dependence explicit. Similarly, the symbol $\mathbf{E}_M$ will be later used to denote expectations of the noise returns taken with respect to the measure $f_M(.)$.

While the equilibrium return process $r_{j\delta}^*$ is modeled as being $O_p(\sqrt{\delta})$ over any intra-period time horizon of size $\delta = \frac{h}{M}$, the contaminations $\varepsilon_{j\delta}$ in the observed return process are $O_p(1)$. This result, which is a consequence of Assumptions 1(1) and 2(2), implies that longer period returns are less contaminated by noise than shorter period returns. Differently put, the magnitude of the frictions does not decrease with the distance between subsequent time stamps. Provided sampling does not occur between high-frequency price updates, the rounding of recorded prices to a grid (price discreteness) and the existence of different prices for buyers and sellers *alone* make this feature of

the setup presented here empirically compelling. As we discuss in what follows, the different stochastic orders of r^* and ε are important aspects of some recent approaches to equilibrium price variance estimation as well as to transaction cost evaluation.

2.1 *The MA*(1) *case*

Sometimes the dependence structure of the microstructure friction process can be simplified. Specifically, one can modify Assumption 2 as follows:

Assumption 2b.

(1) The microstructure frictions in the price process η are i.i.d. mean zero.
(2) $\eta \perp\!\!\!\perp p^*$.

If the microstructure noise contaminations in the price process η are i.i.d., then the noise returns ε display an $MA(1)$ structure with a negative first-order autocorrelation. Importantly, the noise return moments do not depend on M, i.e., the number of observations over h. This is an important feature of the $MA(1)$ model which, as we discuss below, has been exploited in recent work on volatility estimation.

The $MA(1)$ model, as typically justified by bid–ask bounce effects, is bound to be an approximation. However, it is known to be a realistic approximation in decentralized markets where traders arrive in a random fashion with idiosyncratic price setting behavior, the foreign exchange market being a valid example (see, e.g., Bai et al., 2005). It can also be a plausible approximation, capturing first-order effects in the data, in the case of equities when considering transaction prices and/or quotes posted on multiple exchanges. Bandi and Russell (2006b) provide additional discussions. Awartani et al. (2004) propose a formal test of the $MA(1)$ market microstructure model.

3 The variance of the equilibrium price

The recent availability of quality high-frequency financial data has motivated a growing literature devoted to the model-free measurement of the equilibrium price variance. We refer the interested reader to the review paper by Andersen et al. (2002) and the references therein. The main idea is to aggregate intra-period squared continuously-compounded returns and compute $\widehat{V} = \sum_{j=1}^{M} r_{j\delta}^2$ over h. The quantity $\widehat{V}$, which has been termed "realized variance," is thought to approximate the increments of the quadratic variation of the semimartingale that drives the underlying logarithmic price process, i.e., $V = \int_0^h \sigma_s^2 \, \mathrm{d}s$. The consistency result justifying this procedure is the convergence in probability of $\widehat{V}$ to V as returns are computed over intervals that are increasingly small asymptotically, i.e., as $\delta \to 0$ or, equivalently, as $M \to \infty$

for a fixed h. This result is a cornerstone in semimartingale process theory (see, e.g., (Chung and Williams, 1990, Theorem 4.1, p. 76)).[2] More recently, the important work of Andersen et al. (2001, 2003) and Barndorff-Nielsen and Shephard (2002, 2004b) has championed empirical implementation of these ideas.

The theoretical validity of this identification procedure hinges on the observability of the equilibrium price process. However, it is widely accepted that the equilibrium price process and, as a consequence, the equilibrium return data are contaminated by market microstructure effects. Even though the early realized variance literature is aware of the potential importance of market microstructure effects, it has largely abstracted from them. The theoretical and empirical consequences of the presence of market microstructure frictions in the observed price process have been explored only recently.

3.1 Inconsistency of the realized variance estimator

Under the price formation mechanism in Section 2, the realized variance estimates are asymptotically dominated by noise as the number of squared return data increases over a fixed time period. Write

$$\widehat{V} = \sum_{j=1}^{M} r_{j\delta}^2 = \sum_{j=1}^{M} r_{j\delta}^{*2} + \sum_{j=1}^{M} \varepsilon_{j\delta}^2 + 2\sum_{j=1}^{M} r_{j\delta}\varepsilon_{j\delta}. \tag{4}$$

Since $r_{j\delta}^*$ is $O_p(\sqrt{\delta})$ and $\varepsilon_{j\delta}$ is $O_p(1)$, the term $\sum_{j=1}^{M} \varepsilon_{j\delta}^2$ is the dominating term in the sum. Specifically, this term diverges to infinity almost surely as $M \to \infty$. The theoretical consequence of this divergence is a realized variance estimator that fails to converge to the increment of the quadratic variation (integrated variance) of the underlying logarithmic price process but, instead, increases without bound almost surely over any fixed period of time, however small: $\widehat{V} \overset{a.s.}{\to} \infty$ as $M \to \infty$ (or $\delta = \frac{h}{M} \to 0$ given h). This point has been made in independent and concurrent work by Bandi and Russell (2003, 2006a) and Zhang et al. (2005).[3]

The divergence to infinity of the realized variance estimator over any fixed time period is an asymptotic approximation to a fairly pervasive empirical

[2] The corresponding weak convergence result is discussed in Jacod (1994), Jacod and Protter (1998), Barndorff-Nielsen and Shephard (2002), and Bandi and Russell (2003). Mykland and Zhang (2006) cover the case of irregularly-spaced data. Goncalves and Meddahi (2004)) discuss finite sample improvements through bootstrap methods (see, also, Goncalves and Meddahi, 2006).

[3] This theoretical result is general and relies on the different stochastic orders of the equilibrium returns and noise returns. The result does not hinge on an $MA(1)$ structure for the noise return component ε, as implied by Assumption 2b(1). Also, importantly, the result does not hinge on the independence between the price process and the noise, as implied by Assumption 2(3) and Assumption 2b(2). Bandi and Russell (2003) clarify both statements.

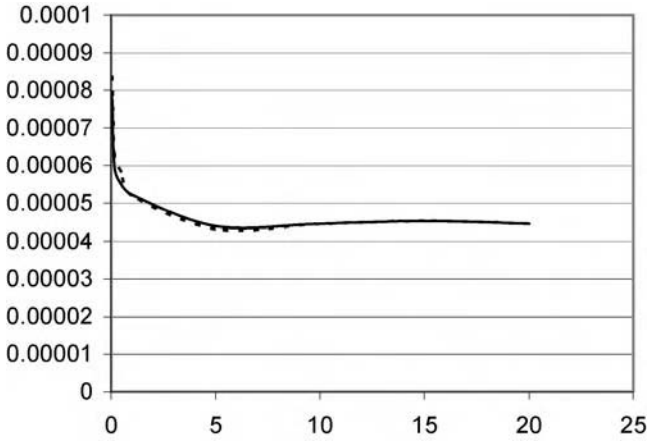


Fig. 1. "Volatility signature plots" for IBM using mid-quotes from (i) the NYSE only (solid line) and (ii) the NYSE and the Midwest exchange (dotted line). We plot realized variance as a function of the sampling frequency (in minutes). The data are collected for the month of February 2002 using the filter discussed in Bandi and Russell (2006a).

fact. When computing realized variance estimates for a variety of sampling frequencies δ, the resulting estimates tend to increase as one moves to high frequencies (as $\delta \to 0$). In the terminology of Andersen et al. (1999, 2000), "the volatility signature plots," namely the plots of realized variance estimates versus different sampling frequencies,[4] are often upward sloping at high frequencies. Figure 1 shows volatility signature plots constructed using IBM mid-quotes obtained from (i) NYSE quotes and (ii) NYSE and Midwest exchange quotes. Figure 2 presents volatility signature plots for IBM using (i) NYSE and NASDAQ quotes and (ii) all quotes from the consolidated market. Figure 3 presents volatility signature plots using mid-quotes obtained from two NASDAQ stocks (Cisco Systems and Microsoft). The data are collected for the month of February 2002. In all cases the realized variance estimates increase as the sampling intervals decrease (see, also, the discussion in Bandi and Russell, 2006b).

3.2 The mean-squared error of the realized variance estimator

The presence of market microstructure contaminations induces a bias–variance trade-off in integrated variance estimation through realized variance. When the equilibrium price process is observable, higher sampling frequencies over a fixed period of time result in more precise estimates of the integrated variance of the logarithmic equilibrium price (see, e.g., Andersen et al.,

[4] See, also, Fang (1996).

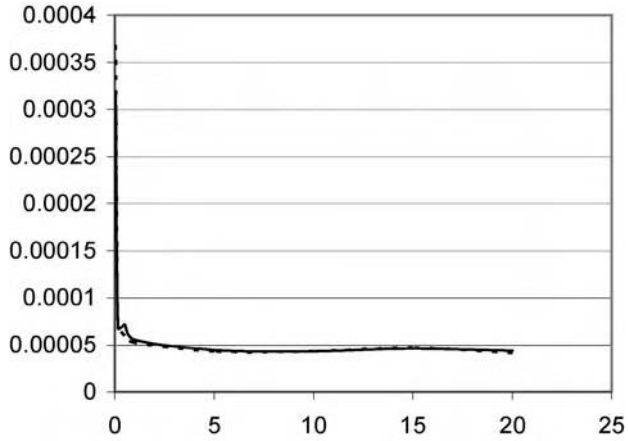


Fig. 2. "Volatility signature plots" for IBM using mid-quotes from (i) the NYSE and NASDAQ (solid line) and (ii) the consolidated market (dotted line). We plot realized variance as a function of the sampling frequency (in minutes). The data are collected for the month of February 2002 using the filter discussed in Bandi and Russell (2006a).

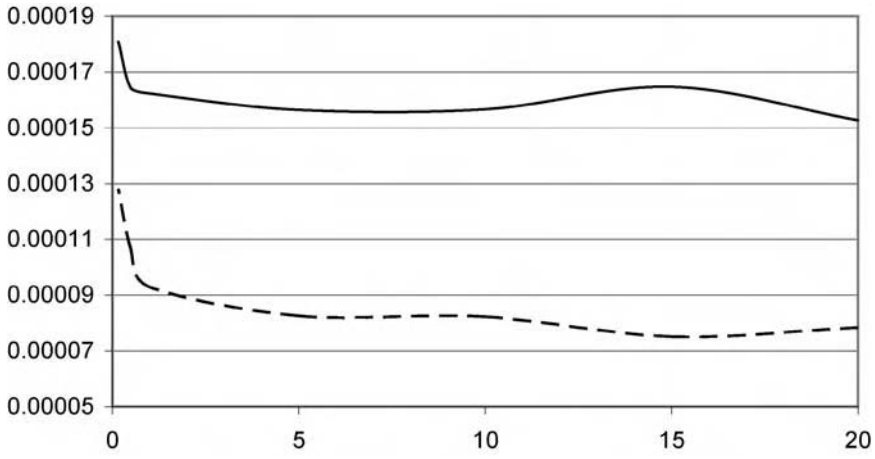


Fig. 3. "Volatility signature plots" for Cisco Systems (dotted line) and Microsoft (dashed line). We plot realized variance as a function of the sampling frequency (in minutes). The data are mid-quotes collected for the month of February 2002. We use the filter discussed in Bandi and Russell (2006a).

2003, and Barndorff-Nielsen and Shephard, 2002). When the equilibrium price process is not observable, as is the case in the presence of microstructure frictions, frequency increases provide information about the underlying integrated variance but, inevitably, entail accumulation of noise that affects both the bias and the variance of the estimator (Bandi and Russell, 2003, 2006a, and Zhang et al., 2005).

Under Assumptions 1 and 2, absence of leverage effects ($\sigma \perp\!\!\!\perp W$), and unpredictability of the equilibrium returns ($\alpha_t = 0$),[5] Bandi and Russell (2003) provide an expression for the conditional (on the underlying volatility path) mean-squared error (MSE) of the realized variance estimator as a function of the sampling frequency δ (or, equivalently, as a function of the number of observations M), i.e.,

$$\mathbf{E}_M\big(\widehat{V} - V\big)^2 = 2\frac{h}{M}\big(Q + o(1)\big) + \Pi_M, \tag{5}$$

where

$$\Pi_M = M\mathbf{E}_M\big(\varepsilon^4\big) + 2\sum_{j=1}^{M}(M-j)\mathbf{E}_M\big(\varepsilon^2\varepsilon_{-j}^2\big) + 4\mathbf{E}_M\big(\varepsilon^2\big)V, \tag{6}$$

and $Q = \int_0^h \sigma_s^4\,\mathrm{d}s$ is the so-called quarticity (see, e.g., Barndorff-Nielsen and Shephard, 2002). Notice that the bias of the estimator can be easily deduced by taking the expectation of $\widehat{V}$ in Eq. (4), i.e.,

$$\mathbf{E}_M\big(\widehat{V} - V\big) = M\mathbf{E}_M\big(\varepsilon^2\big). \tag{7}$$

As for the variance of $\widehat{V}$, we can write

$$\mathbf{E}_M\big(\widehat{V} - \mathbf{E}_M\big(\widehat{V}\big)\big)^2 = 2\frac{h}{M}\big(Q + o(1)\big) + \Pi_M - M^2\big(\mathbf{E}_M\big(\varepsilon^2\big)\big)^2. \tag{8}$$

As we discuss below, the conditional MSE of $\widehat{V}$ can serve as the basis for an optimal sampling theory designed to choose M in order to balance bias and variance.

[5] Both additional assumptions, namely absence of leverage effects and unpredictability of the equilibrium returns, can be justified. In the case of the latter, Bandi and Russell (2003) argue that the drift component α_t is rather negligible in practice at the sampling frequencies considered in the realized variance literature. They provide an example based on IBM. Assume a realistic annual constant drift of 0.08. The magnitude of the drift over a minute interval would be $0.08/(365 * 24 * 60) = 1.52 \times 10^{-7}$. Using IBM transaction price data from the TAQ data set for the month of February 2002, Bandi and Russell (2003) compute a standard deviation of IBM return data over the same horizon equal to 9.5×10^{-4}. Hence, at the one minute interval, the drift component is 1.6×10^{-4} or nearly 1/10, 000 the magnitude of the return standard deviation. Assuming absence of leverage effects is empirically reasonable in the case of exchange rate data. The same condition appears restrictive when examining high frequency stock returns. However, some recent work uses tractable parametric models to show that the effect of leverage on the unconditional MSE of the realized variance estimator in the absence of market microstructure noise is negligible (see Meddahi, 2002). This work provides some justification for the standard assumption of no-leverage in the realized variance literature. Andersen et al. (2002) discuss this issue.

4 Solutions to the inconsistency problem

4.1 The early approaches: sparse sampling and filtering

Thorough theoretical and empirical treatments of the consequences of market microstructure contaminations in realized variance estimation are recent phenomena. However, while abstracting from in-depth analysis of the implications of frictions for variance estimation, the early realized variance literature is concerned about the presence of microstructure effects in recorded asset prices (see, e.g., the discussion on this topic in Andersen et al., 2002).

In order to avoid substantial contaminations at high sampling frequencies, Andersen et al. (2001), for example, suggest sampling at frequencies that are lower than the highest frequencies at which the data arrives. The 5-minute interval was recommended as a sensible approximate choice. Relying on the leveling off of the volatility signature plots at frequencies around 15 minutes, Andersen et al. (1999, 2000) suggest using 15- to 20-minute intervals in practice. If the equilibrium returns are unpredictable ($\alpha_t = 0$), the correlation structure of the observed returns must be imputed to microstructure noise. Andersen et al. (2001, 2003), among others, filter the data using an $MA(1)$ filter. An $AR(1)$ filter is employed in Bollen and Inder (2002).

4.2 MSE-based optimal sampling

More recently, an MSE-based optimal sampling theory has been suggested by Bandi and Russell (2003, 2006a). Specifically, in the case of the model laid out above, the optimal frequency $\delta^* = \frac{h}{M^*}$ at which to sample continuously-compounded returns for the purpose of realized variance estimation can be chosen as the minimizer of the MSE expansion in Section 3.2.

Bandi and Russell's theoretical framework clarifies outstanding issues in the extant empirical literature having to do with sparse sampling and filtering. We start with the former. The volatility signature plots provide useful insights about the bias of the realized variance estimates. The bias typically manifests itself in an upward sloping pattern as the sampling intervals become short, i.e., the bias increases with M (see Eq. (7)).[6] However, it is theoretically fairly arbitrary to choose a single optimal frequency solely based on bias considerations. While it is empirically sensible to focus on low frequencies for the purpose of bias reduction, the bias is only one of the components of the estimator's estimation error. At sufficiently low frequencies the bias can be negligible. However,

[6] The possible dependence between the equilibrium price p^* and the market microstructure frictions η complicates matters. Negative dependence, for instance, might drastically reduce the upward trend of the volatility signature plots at high sampling frequencies. Equation (4) illustrates this point. The empirical relevance of negative dependence between equilibrium price and market microstructure frictions is discussed in Hansen and Lunde (2006). We focus on this issue in Section 5.2 below.

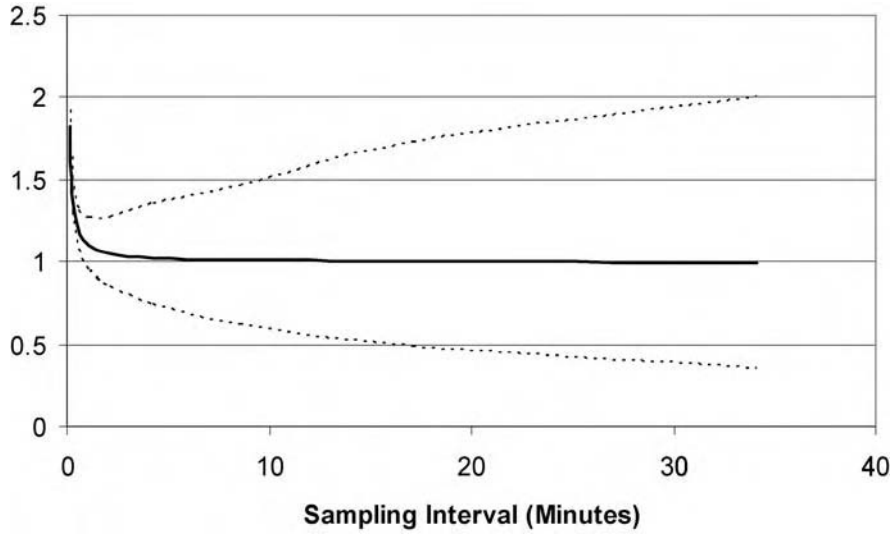


Fig. 4. Simulated "volatility signature plot" from a stochastic volatility diffusion model with parameter values consistent with IBM (see Bandi and Russell, 2003, 2006a, for details). The solid line is the average (across simulations) of the realized variance estimates for each sampling interval (in minutes). The dotted lines are 95% empirical intervals from the simulations. The true integrated variance is standardized to 1.

at the same frequencies, the variability of the estimates might be substantial (see Eq. (8)). Figure 4 is a picture from simulations for parameter values consistent with IBM.

The MSE-based sampling in Bandi and Russell (2003, 2006a) trades off bias and variance optimally. As for filtering, while the dependence that the noise induces in the return data can be reduced by it, residual contaminations are bound to remain in the data. These contaminations continue to give rise to inconsistent realized variance estimates. Bandi and Russell (2003) make this point while studying the theoretical properties of both filtering at the highest frequencies at which observations arrive and filtering at all frequencies.

Bandi and Russell (2003) discuss evaluation of the MSE under Assumptions 1 and 2 as well as in the $MA(1)$ case (i.e., under Assumptions 1 and 2b). In both cases, it is assumed that $\alpha_t = 0$ and $\sigma \perp\!\!\!\perp W$. When empirically justifiable, the $MA(1)$ case is very convenient in that the moments of the noise do not depend on the sampling frequency. Furthermore, the MSE simplifies substantially:

$$\mathbf{E}_M\big(\widehat{V} - V\big)^2 = 2\frac{h}{M}\big(Q + o(1)\big) + M\beta + M^2\alpha + \gamma, \tag{9}$$

where the parameters α, β, and γ are defined as

$$\alpha = \big(\mathbf{E}\big(\varepsilon^2\big)\big)^2, \tag{10}$$

$$\beta = 2\mathbf{E}\big(\varepsilon^4\big) - 3\big(\mathbf{E}\big(\varepsilon^2\big)\big)^2, \tag{11}$$

and

$$\gamma = 4\mathbf{E}\big(\varepsilon^2\big)V - \mathbf{E}\big(\varepsilon^4\big) + 2\big(\mathbf{E}\big(\varepsilon^2\big)\big)^2. \tag{12}$$

If M^* is large, the following approximation to the optimal number of observations applies:

$$M^* \approx \left(\frac{hQ}{(\mathbf{E}(\varepsilon^2))^2}\right)^{1/3}. \tag{13}$$

This approximation readily clarifies the nature of the microstructure-induced trade-off between the bias and variance of the realized variance estimator. If the signal coming from the underlying equilibrium price process (Q) is large relative to the noise determined by the frictions ($(\mathbf{E}(\varepsilon^2))^2$), then sampling can be conducted at relatively high frequencies. Hence, M^* is effectively a signal-to-noise ratio.

In the $MA(1)$ case, evaluation of the MSE does not need to be implemented on a grid of frequencies and simply relies on the consistent estimation of the frequency-independent moments of the noise ($\mathbf{E}(\varepsilon^2)$ and $\mathbf{E}(\varepsilon^4)$) as well as on the estimation of the quarticity term Q.[7] In this case, Bandi and Russell (2003, 2006a) show that sample moments of the *observable* contaminated return data can be employed to identify the moments of the *unobservable* noise process at *all* frequencies. Thus, while realized variance is inconsistent in the presence of microstructure noise, appropriately defined arithmetic averages of the observed returns consistently estimate the moments of the noise. Under $\mathbf{E}(\eta^8) < \infty$, the following result holds:

$$\frac{1}{M}\sum_{j=1}^{M} r_{j\delta}^q - \mathbf{E}\big(\varepsilon^q\big) \xrightarrow{p} 0 \quad 1 \leqslant q \leqslant 4, \tag{14}$$

[7] The quarticity term can be identified using the estimator proposed by Barndorff-Nielsen and Shephard (2002), namely

$$\widehat{Q} = \frac{M}{3h}\sum_{j=1}^{M} r_{j\delta}^4.$$

(See Barndorff-Nielsen and Shephard, 2004b, and Zhang et al., 2005, for alternative approaches.) However, $\widehat{Q}$ is not a consistent estimate of Q in the presence of noise. In practice, one could then sample the observed returns to be used in the definition of $\widehat{Q}$ at a lower frequency than the highest frequency at which observations arrive. Bandi and Russell (2006a) show by simulation that sampling returns in a reasonable (but possibly suboptimal) fashion for the purpose of quarticity estimation does not give rise to very imprecise sampling choices for realized variance. Using data, Bandi and Russell (2006a) find that sampling intervals for the quarticity between 10 and 20 minutes have a negligible effect on the resulting optimal frequency of the realized variance estimator. In light of the important role played by the quarticity in this and other identification procedures (see below), future research should study more efficient methods to estimate this term in the presence of realistic market microstructure noise effects.

as $M \to \infty$.[8] We provide intuition for this finding in the case $q = 2$. The sum of the squared contaminated returns can be written as in Eq. (4) above, namely as the sum of the squared equilibrium returns plus the sum of the squared noise returns and a cross-product term. The price formation mechanism in Section 2 is such that the orders of magnitude of the three terms in Eq. (4) above differ since $r^*_{j\delta} = O_p(\sqrt{\delta})$ and $\varepsilon_{j\delta} = O_p(1)$. Thus, the microstructure noise component dominates the equilibrium return process at very high frequencies, i.e., for small values of δ. This effect determines the diverging behavior of $\widehat{V}$, as discussed above. By the same logic, when we average the contaminated squared returns as in Eq. (14), the average of the squared noises constitutes the dominating term in the average. Naturally, then, while the remaining terms in the average vanish asymptotically due to the stochastic order of the equilibrium returns, i.e., $O_p(\sqrt{\delta})$, the average of the squared noise returns converges to the second moment of the noise returns as implied by Eq. (14).

Using a sample of mid-quotes for the S&P 100 stocks over the month of February 2002, Bandi and Russell (2006a) report (average) daily optimal sampling frequencies that are between 0.5 minutes and about 14 minutes with a median value of about 3.5 minutes. The MSE improvements that the MSE-based optimal frequencies guarantee over the 5- or 15-minute frequency can be substantial. Not only do the optimal frequencies vary cross-sectionally, they also change over time. Using mid-quotes dating back to 1993 for three stocks with various liquidity features, namely EXXON Mobile Corporation (XOM), SBC Communications (SBC), and Merrill Lynch (MER), Bandi and Russell (2006a) show that the daily optimal frequencies have substantially increased in recent times, generally due to decreases in the magnitude of the noise moments. This effect should in turn be attributed to an overall increase in liquidity.

In agreement with the analysis in Bandi and Russell (2003, 2006a), Oomen (2006) discusses an MSE-based approach to optimal sampling for the purpose of integrated variance estimation. However, some important novelties characterize Oomen's work. First, the underlying equilibrium price is not modeled as in Section 2 but as a compound Poisson process. Second, Oomen explores the relative benefits of transaction time sampling versus calendar time sampling.

Consider a Poisson process $N(t)$ with intensity $\lambda(t)$. In Oomen (2006) the observed logarithmic price process is expressed as

$$p_t = \underbrace{p_0 + \sum_{j=1}^{N(t)} \xi_j}_{p^*_t} + \sum_{j=1}^{N(t)} \eta_j, \tag{15}$$

[8] Importantly, this result is robust to the presence of a drift term ($\alpha_t \neq 0$), dependence between the frictions and the equilibrium price, and leverage effects (Bandi and Russell, 2003). Under assumptions, it is also robust to dependence in the frictions (Bandi and Russell, 2004). See Section 7 for additional discussions.

where $\xi_j \sim$ i.i.d. $N(\mu_\xi, \sigma_\xi^2)$ and $\eta_j = \Delta\nu_j + \rho_2\Delta\nu_{j-1} + \cdots + \rho_q\Delta\nu_{j-q+1}$ with $\Delta\nu_j = \nu_j - \nu_{j-1}$ and $\nu_j \sim$ i.i.d. $N(0, \sigma_\nu^2)$. The process $N(t)$ effectively counts the number of transactions up to time t. This process is assumed to be independent of both ξ and ν. Importantly, the equilibrium price p_t^* is equal to $p_0 + \sum_{j=1}^{N(t)} \xi_j$. Hence, p_t^* is a jump process of finite variation (in the tradition of Press, 1967) with integrated variance (i.e., the object of econometric interest) given now by $V = \sigma_\xi^2 \int_0^h \lambda(s)\,\mathrm{d}s = \sigma_\xi^2 \Lambda(h)$. The microstructure noise contaminations η have an $MA(q)$ structure. Setting q equal to one yields a negative first-order autocorrelation of the calendar time continuously-compounded returns since

$$p_t - p_{t-\tau} = \sum_{j=N(t-\tau)}^{N(t)} \xi_j + \nu_{N(t)} - \nu_{N(t-\tau)-1}, \tag{16}$$

for any calendar time interval τ.

Oomen (2006) provides closed-form expressions for the MSE of the realized variance estimator under both calendar time sampling, as earlier, and transaction time sampling. Given M (the total number of observations), transaction time sampling leads to a sequence of prices $\{p_{t_i}\}_{i=0}^{M}$ with sampling times implicitly defined as $N(t_i) = i\lfloor \frac{N(h)}{M} \rfloor$, where $\lfloor x \rfloor$ is the integer part of x.[9] Oomen (2006) also discusses optimal choice of M in an MSE sense. Using IBM transaction prices from the consolidated market over the period between 2000 and 2004, he finds that transaction time sampling generally outperforms calendar time sampling. In his sample the average decrease in MSE that transaction time sampling induces is about 5%. Gains up to 40% can be achieved. As intuition suggests, the largest gains are obtained for days with irregular trading patterns.

4.3 *Bias-correcting*

The microstructure-induced bias of the realized variance estimator represents a large component of the estimator's MSE. This point is emphasized by Hansen and Lunde (2006). Hansen and Lunde (2006) propose a bias-adjustment to the conventional realized variance estimator. The bias-corrected estimator they suggest is in the tradition of HAC estimators such as those of Newey and West (1987) and Andrews and Monahan (1992). Its general form

[9] Equivalently, given M, business time sampling can be obtained by sampling prices at times t_i so that $\Lambda(t_i) = i\frac{\Lambda(h)}{M}$. Because $\lambda(.)$ is latent and since, conditionally on $\lambda(.)$, $\mathbf{E}(N(t)) = \Lambda(t)$, transaction time sampling can be interpreted as a feasible version of business time sampling (Oomen, 2006).

is

$$\widetilde{V} = \sum_{j=1}^{M} r_{j\delta}^2 + 2\sum_{h=1}^{q_M} \frac{M}{M-h} \sum_{j=1}^{M-h} r_{j\delta} r_{(j+h)\delta}, \tag{17}$$

where q_M is a frequency-dependent number of covariance terms. If the correlation structure of the noise returns has a finite order and $\alpha_t = 0$, under appropriate conditions on q_M the estimator in Eq. (17) is unbiased for the underlying integrated variance over the period, i.e., $\mathbf{E}_M(\widetilde{V}) = V$.

The intuition readily derives from the $MA(1)$ noise case. In this case the estimator takes the simpler expression

$$\widetilde{V}^{MA(1)} = \sum_{j=1}^{M} r_{j\delta}^2 + 2\frac{M}{M-1} \sum_{j=1}^{M-1} r_{j\delta} r_{(j+1)\delta}. \tag{18}$$

Under Assumption 1, Assumption 2b, and $\alpha_t = 0$, the covariance between $r_{j\delta}$ and $r_{(j+1)\delta}$, i.e., $\mathbf{E}_M(r_{j\delta} r_{(j+1)\delta})$, is the same at all frequencies and equal to $-\mathbf{E}(\eta^2)$. Hence, $\mathbf{E}_M(2\frac{M}{M-1}\sum_{j=1}^{M-1} r_{j\delta} r_{(j+1)\delta}) = -2M\mathbf{E}(\eta^2)$. The bias of the estimator $\widehat{V}$ is equal to $M\mathbf{E}(\varepsilon^2) = 2M\mathbf{E}(\eta^2)$ (see Eq. (7)). Therefore, the second term in Eq. (18) provides the required bias correction. Interestingly, the finite sample unbiasedness of Hansen and Lunde's estimator is robust to the presence of some dependence between the underlying local martingale price process (under $\alpha_t = 0$) and market microstructure noise, i.e., Assumption 2(3) or Assumption 2b(2) can be relaxed. In the $MA(1)$ case, again, it is easy to see that if $\mathbf{E}_M(r^*_{j\delta}\eta_{(j-s)\delta}) = 0$ for all $s \geqslant 1$ (implying that microstructure noise does not predict equilibrium returns) and $\mathbf{E}_M(r^*_{j\delta}\eta_{(j+1)\delta}) = 0$, then $\mathbf{E}_M(\widetilde{V}^{MA(1)}) = V$ (see the discussion in Bandi and Russell, 2006b). In other words, the contemporaneous covariances $\mathbf{E}_M(r^*_{j\delta}\eta_{j\delta})$ are not required to be zero. This is an important property.

Under an assumed $MA(1)$ noise structure, Zhou (1996) is the first to use the bias-corrected estimator in Eq. (18) in the context of variance estimation through high-frequency data. His original setup assumes a constant return variance and Gaussian market microstructure noise. In this framework, Zhou (1996) characterizes the variance of the estimator and concludes that it can be minimized for a finite M. Under the more general assumptions in Section 2, but again in the presence of $MA(1)$ frictions, Hansen and Lunde (2006) have recently further studied the MSE properties of the estimator in Eq. (18). Using 5 years of DJIA price data from January 3, 2000, to December 31, 2004, they find that bias-correcting permits optimal sampling at higher frequencies than those obtained by Bandi and Russell (2006a) using the classical realized variance estimator. In addition, MSE improvements can be achieved. Consider Alcoa (AA), for example. They report an (average) daily optimal sampling frequency for their bias-corrected estimator equal to about 46 seconds. Their reported (average) optimal daily frequency for the realized variance estimator

is about 9 minutes. Bias-correcting yields a reduction in the root MSE of about 33%.

Alternative bias-corrections can be provided in both the correlated noise case and in the $MA(1)$ case. The correlated noise case is studied in Bandi and Russell (2003) and Zhang (2006a) (see, also, Aït-Sahalia et al., 2005b). The $MA(1)$ case is discussed in Bandi and Russell (2003) and Zhang et al. (2005). For conciseness, we focus only on the $MA(1)$ case. As we point out above, the bias of the realized variance estimator can be estimated consistently by computing an arithmetic average of the squared observed return data sampled at the highest frequencies (see Eq. (14)). The bias-corrected realized variance estimator is then equal to

$$\overleftrightarrow{V} = \widehat{V} - \overleftrightarrow{M}\frac{1}{M}\sum_{j=1}^{M} r_{j\delta}^2, \tag{19}$$

where M is the number of observations in the full sample and $\overleftrightarrow{M}$ is the number of observations used to compute $\widehat{V}$.[10] The approximate (MSE-based) optimal number of observations $\overleftrightarrow{M}^*$ of the estimator in Eq. (19) is now

$$\overleftrightarrow{M}^* \approx \left(\frac{hQ}{2\mathbf{E}(\varepsilon^4) - 3(\mathbf{E}(\varepsilon^2))^2}\right)^{1/2} \tag{20}$$

(Bandi and Russell, 2003).[11] In agreement with Hansen and Lunde's findings (Hansen and Lunde, 2006), this optimal frequency is generally higher then the optimal frequency of the realized variance estimator. Furthermore, it is associated with MSE improvements.

In the spirit of Zhou (1996) and Hansen and Lunde (2006), Oomen (2005) extends the framework in Oomen (2006) to the case of bias-corrected realized variance. Specifically, he studies the MSE properties of the estimator in

[10] When M is not large enough, the equilibrium return component in the estimated second moment $\frac{1}{M}\sum_{j=1}^{M} r_{j\delta}^2$ might be non-negligible. Specifically, conditional on the volatility path, the finite sample bias of $\frac{1}{M}\sum_{j=1}^{M} r_{j\delta}^2$, as an estimate of $\mathbf{E}(\varepsilon^2)$, is equal to $\frac{V}{M}$. Hence, this empirical moment can be purged of residual contaminations induced by the equilibrium price variance by subtracting from it a quantity defined as $\frac{1}{M}\sum_{j=1}^{\widetilde{P}} r_{j\delta}^2$, where $\widetilde{P}$ is an appropriate number of low frequency returns calculated using 15- or 20-minute intervals, for instance. The quantity $\frac{1}{M}\sum_{j=1}^{\widetilde{P}} r_{j\delta}^2$ is roughly unbiased for $\frac{V}{M}$. The resulting estimator, i.e., $\frac{1}{M}(\sum_{j=1}^{M} r_{j\delta}^2 - \sum_{j=1}^{\widetilde{P}} r_{j\delta}^2)$, has, of course, the same limiting properties as $\frac{1}{M}\sum_{j=1}^{M} r_{j\delta}^2$ for any fixed $\widetilde{P}$. A similar correction is discussed in Bandi and Russell (2003, 2004) and Hansen and Lunde (2006). The presence of dependence between the frictions and the equilibrium price process complicates matters. Bandi and Russell (2004) discuss a bias-correction in this case.

[11] The expression would be exact only if the estimator were defined as $\widehat{V} - \overleftrightarrow{M}\mathbf{E}(\varepsilon^2)$ which is, of course, infeasible in practice.

Eq. (18) for the case of an underlying jump process of finite variation (as in Eq. (15)) and transaction time sampling. Using IBM transaction data from the consolidated market for the period January 2, 2003–August 31, 2003, he confirms that (i) transaction time sampling can be beneficial in practice (as in Oomen, 2006) and (ii) bias-correcting can induce an increase in the optimal sampling frequency along with MSE gains. In the case of the bias-corrected estimator, he reports an (average) optimal daily frequency of about 12 seconds. The corresponding (average) optimal daily frequency of the classical realized variance estimator is around 2.5 minutes. While bias-correcting yields MSE gains of about 65% using his data, he reports that further gains (about 20%) can be obtained by employing transaction time sampling in place of calendar time sampling.

4.4 Sub-sampling

The bias-corrected estimators studied by Zhou (1996), Hansen and Lunde (2006), and Oomen (2005) are inconsistent. Biased in a finite sample, but consistent, is the estimator recently advocated by Zhang et al. (2005) in the presence of $MA(1)$ market microstructure noise. (See, also, Aït-Sahalia et al., 2005a, for a study of consistent maximum likelihood estimation of the constant variance of a scalar diffusion process in parametric models with microstructure effects.) This promising approach relies on subsampling.[12] Assume availability of n, generally non-equispaced, observations. Define q non-overlapping sub-grids $G^{(i)}$ of the full grid of n arrival times with $i = 1, \ldots, q$. The first sub-grid starts at t_0 and takes every qth arrival time, i.e., $G^{(1)} = (t_0, t_{0+q}, t_{0+2q}, \ldots)$, the second sub-grid starts at t_1 and also takes every qth arrival time, i.e., $G^{(2)} = (t_1, t_{1+q}, t_{1+2q}, \ldots)$, and so on. Given the generic ith sub-grid of arrival times, one can define the corresponding realized variance estimator as

$$\widehat{V}^{(i)} = \sum_{t_j, t_{j+} \in G^{(i)}} (p_{t_{j+}} - p_{t_j})^2, \tag{21}$$

where t_j and t_{j+} denote consecutive elements in $G^{(i)}$. Zhang et al.'s subsampling approach entails averaging the realized variance estimates over sub-grids as well as bias-correcting them. To this extent, define

$$\widehat{V}^{\text{sub}} = \frac{\sum_{i=1}^{q} \widehat{V}^{(i)}}{q} - \bar{n}\widehat{\mathbf{E}}(\varepsilon^2), \tag{22}$$

where $\bar{n} = \frac{n-q+1}{q}$, $\widehat{\mathbf{E}}(\varepsilon^2) = \frac{\sum_{j=1}^{n}(p_{t_{j+}} - p_{t_j})^2}{n}$ is a consistent estimate of the second moment of the noise return (as in Eq. (14)), and $\bar{n}\widehat{\mathbf{E}}(\varepsilon^2)$ is the required

[12] Müller (1993), Zhou (1996), and the review of Politis et al. (1999) contain early discussions of similar ideas.

bias-correction (as in Eq. (7)). Under Assumption 1 and Assumption 2b (i.e., the $MA(1)$ noise case), Zhang et al. (2005) show that, as $q, n \to \infty$ with $\frac{q}{n} \to 0$ and $\frac{q^2}{n} \to \infty$, $\widehat{V}^{\text{sub}}$ is a consistent estimator of the integrated variance V over h. Provided $q = cn^{2/3}$, the rate of convergence of $\widehat{V}^{\text{sub}}$ to V is $n^{1/6}$ and the asymptotic distribution is mixed-normal with an estimable asymptotic variance. Specifically,

$$n^{1/6}\big(\widehat{V}^{\text{sub}} - V\big) \Rightarrow \left(\sqrt{8c^{-2}\big(\mathbf{E}\big(\eta^2\big)\big)^2 + c\frac{4}{3}Q}\right)N(0,1), \tag{23}$$

where the symbol "$\Rightarrow$" denotes weak convergence. The proportionality factor c can be selected optimally in order to minimize the limiting variance in Eq. (23). This minimization leads to an asymptotically optimal number of subsamples given by

$$q^{\text{asy}} = c^{\text{asy}} n^{2/3} = \left(\frac{16(\mathbf{E}(\eta^2))^2}{h\frac{4}{3}Q}\right)^{1/3} n^{2/3} \tag{24}$$

(Zhang et al., 2005). Both components of the factor c^{asy}, namely $\mathbf{E}(\eta^2)$ and Q, can be readily evaluated from the data. Specifically, the second moment of the noise η can be estimated by using a (standardized) sample average of squared continuously-compounded returns sampled at the highest frequencies as discussed in Section 4.2.[13] The quarticity term Q can be identified by employing the Barndorff-Nielsen and Shephard's quarticity estimator, namely $\widehat{Q} = \frac{M}{3h}\sum_{j=1}^{M} r_{j\delta}^4$ (Barndorff-Nielsen and Shephard, 2002), with continuously-compounded returns sampled at relatively low frequencies, among other methods. The 15- or 20-minute frequency has been shown to work reasonably well in practice.

The estimator of Zhang et al. (2005) is effectively a "two-scale" estimator relying on very high-frequency return data to identify the bias component as well as on lower frequency return data to characterize the individual realized variances prior to averaging. In recent work, Zhang (2006a) has extended this approach to a "multi-scale" setup. This new estimator achieves the best attainable rate for this type of problems, $n^{1/4}$, and is robust to noise dependence in transaction time. See Aït-Sahalia et al. (2005b) for further discussions.

4.5 Kernels

The subsampling, or "two-scale," estimator is a kernel-based estimator. Specifically, Barndorff-Nielsen et al. (2005) have shown that it can be rewritten

[13] Recall that, under the $MA(1)$ market microstructure model, $\mathbf{E}(\varepsilon^2) = 2\mathbf{E}(\eta^2)$. Hence, $\frac{1}{2M}\sum_{j=1}^{M} r_{j\delta}^2 \underset{M\to\infty}{\overset{p}{\to}} \mathbf{E}(\eta^2)$.

as a "modified" Bartlett kernel estimator, i.e.,

$$\widehat{V}^{\text{sub}} = \left(1 - \frac{n-q+1}{nq}\right)\widehat{\gamma}_0 + 2\sum_{s=1}^{q}\left(\frac{q-s}{q}\right)\widehat{\gamma}_s - \frac{1}{q}\vartheta_q, \tag{25}$$

where $\widehat{\gamma}_s = \sum_{j=1}^{n-s} r_j r_{j+s}$, $\vartheta_1 = 0$ and $\vartheta_q = \vartheta_{q-1} + (r_1 + \cdots + r_{q-1})^2 + (r_{n-q+2} + \cdots + r_n)^2$ for $q \geqslant 2$.

The modification $\frac{1}{q}\vartheta_q$ (called "edge-effect" in Barndorff-Nielsen et al., 2005), which mechanically derives from the construction in the previous subsection, is crucial for the estimator's consistency. Consider a more traditional, "unmodified" Bartlett kernel estimator defined as

$$\widehat{V}^{\text{Bartlett}} = \left(\frac{n-1}{n}\frac{q-1}{q}\right)\widehat{\gamma}_0 + 2\sum_{s=1}^{q}\left(\frac{q-s}{q}\right)\widehat{\gamma}_s. \tag{26}$$

Under Assumption 1 and Assumption 2b (i.e., the $MA(1)$ noise case), Barndorff-Nielsen et al. (2005) find that $\widehat{V}^{\text{Bartlett}}$ is only "near-consistent" as $q, n \to \infty$ with $\frac{q}{n} \to 0$ and $\frac{q^2}{n} \to \infty$, namely under the same conditions yielding consistency of $\widehat{V}^{\text{sub}}$. The limiting variance of $\widehat{V}^{\text{Bartlett}}$ is given by $4(\mathbf{E}(\eta^2))^2$, which is small in practice when compared to V.[14]

Traditional kernel estimators can be rendered consistent. Barndorff-Nielsen et al. (2006b) have recently advocated unbiased symmetric kernels of the type

$$\widehat{V}^{\text{BNHLS}} = \widehat{\gamma}_0 + \sum_{s=1}^{q} w_s(\widehat{\gamma}_s + \widehat{\gamma}_{-s}), \tag{27}$$

where $\widehat{\gamma}_s = \sum_{j=1}^{n} r_j r_{j-s}$ with $s = -q, \ldots, q$, $w_s = k(\frac{s-1}{q})$ and k is a function on $[0, 1]$ satisfying $k(0) = 1$ and $k(1) = 0$ (see, also, Sun, 2006, for a class of unbiased, consistent estimators). If $q = cn^{2/3}$, this family of estimators is consistent (at rate $n^{1/6}$) and asymptotically mixed normal. Interestingly, when $k(x) = 1 - x$ (the Bartlett case), the limiting variance of $\widehat{V}^{\text{BNHLS}}$ is the same as that of the two-scale estimator. Hence, c can be chosen asymptotically as in the previous subsection. If, in addition, $k'(0) = 0$ and $k'(1) = 0$, then the number of autocovariances can be selected so that $q = cn^{1/2}$ and the estimator is consistent at rate $n^{1/4}$. When $k(x) = 1 - 3x^2 + 2x^3$, the limiting distribution of $\widehat{V}^{\text{BNHLS}}$ is the same as that of the multi-scale estimator of Zhang (2006a).

The limiting properties of the estimators in this subsection and in the previous subsection are derived under asymptotic conditions requiring the number

[14] The "unmodified" Bartlett kernel estimator and the "two-scale" estimator are quadratic estimators. A promising approach to consistent integrated variance estimation by virtue of unbiased quadratic estimators is contained in Sun (2006).

of autocovariances (or subsamples) q and the number of observations n to diverge to infinity as $\frac{q}{n} \to 0$ and $\frac{q^2}{n} \to \infty$ (or $\frac{q^2}{n} \to c^2$ when $k'(0) = 0$ and $k'(1) = 0$). Whether these classical conditions in HAC estimation lead to valid asymptotic approximations to the finite sample distributions of these estimators is a question addressed in Bandi and Russell (2005b). As in Barndorff-Nielsen et al. (2005, 2006b), Zhang et al. (2005), and Zhang (2006a), among others, Bandi and Russell (2005b) operate under Assumption 1 (with $\alpha_t = 0$ and $\sigma \perp\!\!\!\perp W$) and Assumption 2b. They recognize that, in practice, the number of autocovariances q is selected as a function of the number of observations n and set the ratio between q and n equal to a value ϕ such that $\phi \in (0, 1]$. Subsequently, they derive the finite sample MSE properties of the Bartlett kernel estimator in Eq. (26), of the subsampling estimator in Eq. (22), and of the class of symmetric kernels in Eq. (27) as a function of ϕ. Finally, they optimize these properties by choosing ϕ as the minimizer of each estimator's finite sample MSE. Their main results can be summarized as follows:

1. The finite sample MSE properties of the consistent, two-scale estimator and of the inconsistent, "unmodified" Bartlett kernel estimator are similar.
2. A large component of their mean-squared error is bias-induced.
3. Asymptotic bandwidth selection methods (as in Eq. (24) above) can be very suboptimal in their case and, more generally, in the case of biased kernel estimators. Because their finite sample bias washes out asymptotically, asymptotic methods do not take the finite sample bias into account and have a tendency to select an excessively small number of bandwidths. A small number of bandwidths can lead to a large bias component in a finite sample.
4. This bias component can be reduced by choosing q in order to minimize the estimator's finite sample MSE. In the case of the modified (two-scale) and "unmodified" Bartlett kernel estimator, Bandi and Russell (2005b) propose a simple (MSE-based) rule-of-thumb which is given by:

$$q^* \approx \left(\frac{3}{2} \frac{\frac{V^2}{n^2}}{Q} \right)^{1/3} n. \tag{28}$$

Since the finite sample bias of these estimators does not depend on the moments of the noise (it only depends on the underlying variance process and the number of observations), this ratio should not be surprising when compared to Eq. (13). As earlier, the ratio trades-off bias and variance. If the bias component (in the numerator) is large relative to the variance component (in the denominator), then the number of autocovariances should be large. Preliminary (roughly unbiased) V and Q estimates can be obtained by using the classical realized variance estimator and the quarticity estimator with returns sampled at low 15- or 20-minute frequencies, for instance.

5. While the optimal finite sample MSE values of the two-scale estimator and of the "unmodified" Bartlett kernel estimator are generally smaller than the optimal finite sample MSE value of the classical realized variance estimator, the gains that these useful estimators can provide over the classical realized variance estimator might be either lost, or severely reduced, by suboptimally choosing the number of autocovariances.
6. The class of estimators proposed by Barndorff-Nielsen et al. (2006b) is unbiased. Asymptotic bandwidth selection criteria are expected to be less detrimental in this case, i.e., suboptimal bandwidth choices will likely lead to smaller finite sample losses.
7. In general, even though all available consistent and "near-consistent" (in the terminology of Barndorff-Nielsen et al., 2005) estimators can yield accurate estimation of V when optimized, asymptotic approximations to their finite sample estimation error might be imprecise. A careful assessment of their accuracy requires paying close attention to their finite sample properties.

5 Equilibrium price variance estimation: directions for future work

5.1 Alternative integrated variance measures

The study of the implications of market microstructure noise for integrated variance estimation has largely focused on realized variance and its modifications. However, in the frictionless case promising alternative estimators have been studied in recent years. The (realized) range of Parkinson (1980) (Alizadeh et al., 2002; Brandt and Diebold, 2006; Christensen and Podolskij, 2005; Martens and Van Dijk, 2007; and Rogers and Satchell, 1991, inter alia), the Fourier approach of Malliavin and Mancino (2002) (see, also, Barucci and Renò, 2002a, 2002b; and Kanatani, 2004a, 2004b), the realized power variation (Jacod, 1994, and Barndorff-Nielsen and Shephard, 2003) and the bypower variation of Barndorff-Nielsen and Shephard (2004a)[15] – more on the last two measures in what follows – are notable examples. It is now of interest to fully understand the properties of these estimators (and other estimators recently proposed) in the presence of realistic market microstructure frictions. Nielsen and Frederiksen (2005) and Huang and Tauchen (2005) contain interesting simulation work on the subject. Much is left for future research.

[15] Corradi and Distaso (2006) use this statistic in the context of specification tests for parametric volatility models. Barndorff-Nielsen et al. (2006a) contain a broad discussion of this and other measures as well as additional references.

5.2 Relaxing the assumptions

Most of the current work on integrated variance estimation by virtue of noisy high-frequency data is conducted under an assumed diffusion process for the equilibrium price process and independent (of the equilibrium price process) $MA(1)$ market microstructure noise. While these assumptions provide a useful theoretical and empirical framework, important applications of interest might require a richer structure.

We start with the properties of the noise process and its relation with the underlying equilibrium price. Bandi and Russell (2003), Hansen and Lunde (2006), and Aït-Sahalia et al. (2005b) provide early, alternative approaches to noise persistence. A thorough discussion of the importance of allowing for dependent noise, mostly when sampling at very high frequencies, is contained in Hansen and Lunde (2006). Phillips and Yu (2005) emphasize that at high frequencies the noise process might even display dependence of the nonstationary type. The observations of Hansen and Lunde (2006) and Phillips and Yu (2005) can be understood in the context of the decomposition in Eq. (2):

$$r_{j\delta} = r^*_{j\delta} + \varepsilon_{j\delta}. \tag{29}$$

When gathering data at high frequencies, sampling between price updates occurring solely in correspondence with changes in the depth leads to observed returns that are equal to zero. In general, negligible observed returns $r_{j\delta}$ combined with unpredictable equilibrium returns $r^*_{j\delta}$ (as implied by our baseline model with $\alpha_t = 0$) induce highly persistent and potentially nonstationary microstructure noise components. Assume, for simplicity, $r_{j\delta} = 0$. Then,

$$0 = r^*_{j\delta} + \varepsilon_{j\delta} \Rightarrow \eta_{j\delta} = \eta_{(j-1)\delta} - r^*_{j\delta}. \tag{30}$$

A broader argument can be made: any factor inducing sluggishness in the adjustments to the observed prices *mechanically* determines persistence in the market microstructure noise components (Bandi and Russell, 2006b). Bandi and Russell (2006b) identify three main factors affecting the stickiness of the observed prices: the market structure (centralized versus decentralized markets), the type of price measurement (mid-quotes versus transaction prices), and the sampling method (calendar time sampling versus event time sampling). Mid-quotes that are posted on centralized markets and are sampled in calendar time are expected to have noise components that are substantially more dependent than transaction prices posted on decentralized markets and sampled in event time. In other words, the extent to which persistence is a first-order effect in the data depends on the economics of price formation as well as on the adopted sampling scheme.

In their articulate study of the properties of market microstructure noise, Hansen and Lunde (2006) also stress that attention should be paid to the dependence between the underlying equilibrium price process and market microstructure noise. Similarly to noise persistence, this dependence somewhat mechanically derives from the degree of stickiness in the observed prices

(Bandi and Russell, 2006b). Equation (29) shows that the more stable the observed prices are, the stronger is the negative dependence between equilibrium returns and noise returns. Hence, the factors inducing persistent noise components are also expected to be the factors inducing negative dependence between the noise returns and the equilibrium returns (Bandi and Russell, 2006b). Barndorff-Nielsen et al. (2006b) and Kalnina and Linton (2006) propose kernel-based approaches to the consistent estimation of integrated variance under some form of dependence between noise and equilibrium price. It is an important challenge for the literature on nonparametric variance estimation to study methods that provide satisfactory finite sample performance when noise persistence and dependence between noise and underlying equilibrium price are relevant effects in the data. In the case of kernel estimators, the issue of bandwidth selection is expected to be, as earlier, of first-order importance.

We now turn to models for the equilibrium price. The equilibrium price formation mechanism in Assumption 1 can be generalized to allow for a jump component in addition to the classical continuous semimartingale component. Barndorff-Nielsen and Shephard (2004a) have provided several stimulating theoretical results to show how to identify the integrated variance of the equilibrium price's continuous sample path component when finite activity jumps play a role (see, also, Mancini, 2003, 2004, for an alternative approach). Their main result is that realized power and realized bypower variation measures are, if properly constructed, "robust" to the presence of discontinuous components of this type. Assume the equilibrium price process is defined as in Assumption 1 and add a component to it expressed as $v_t = \sum_{j=1}^{N(t)} c_j$, where $N(t)$ is a finite activity, simple counting process and the c_j's are non-zero random variables.[16] Thus, $p_t^* = \alpha_t + m_t + v_t$. Now define the r, s-order bypower variation $BV_{(r,s)}$ as

$$BV_{(r,s)} = M^{-1+(r+s)/2} \sum_{j=1}^{M-1} \left|r_{j\delta}^*\right|^r \left|r_{(j+1)\delta}^*\right|^s. \tag{31}$$

In the absence of market microstructure frictions, Barndorff-Nielsen and Shephard (2004a) show that

$$BV_{(r,s)} \underset{M\to\infty}{\overset{p}{\to}} \mu_r \mu_s \int_0^h \sigma_s^{r+s}\, \mathrm{d}s, \tag{32}$$

[16] If $N(t)$ is an homogeneous Poisson process and the c_j's are i.i.d., then v_t is a compound Poisson process.

where $\mu_r = \mathbf{E}(|Z|^r) = 2^{r/2}\frac{\Gamma(\frac{1}{2}(r+1))}{\Gamma(\frac{1}{2})}$ with $Z \sim N(0,1)$, if $\max(r,s) < 2$.[17] This result readily implies that

$$\mu_r^{-1}\mu_{2-r}^{-1}BV_{(r,2-r)} \xrightarrow{p} V, \tag{33}$$

as $M \to \infty$. Since, when microstructure noise is assumed to be absent, realized variance converges to V plus the sum of the squared jumps over the period ($\sum_{j=1}^{N(h)} c_j^2$), subtracting $\mu_r^{-1}\mu_{2-r}^{-1}BV_{(r,2-r)}$ (with $r = 1$, for instance) from realized variance consistently estimates the sum of the squared jumps in the no noise case. This observation is employed by Andersen et al. (2007) and Huang and Tauchen (2005) in their analysis of the contribution of jumps to total price variance. Huang and Tauchen (2005) offer interesting simulation evidence about the robustness of this procedure to some form of market microstructure noise. More theoretical and empirical work ought to be done on the relative role played by jumps and continuous sample path price components in the presence of market frictions. Extensions to infinite activity jumps, and the impact of market frictions in this case, are also of interest. Barndorff-Nielsen et al. (2006c) and Woerner (2006) are recent work on the subject in the frictionless case.

As discussed above, Oomen (2005, 2006) models the underlying equilibrium price as a pure jump process. In Large (2006), it is the observed price process which is modeled as a pure jump process with *constant* jumps whereas, coherently with Assumption 1, the underlying equilibrium price process evolves as a stochastic volatility semimartingale. The difference between the observed price process and the underlying continuous semimartingale defines market microstructure noise. Write the observed price process as

$$p_t = p_0 + \int_0^t c_s \, \mathrm{d}N_s, \tag{34}$$

where N_s is a simple counting process and c is an adapted process taking values k and $-k$, with $k > 0$. The quadratic variation of the observed price process $[p]_h$ can then be expressed as $k^2N(h)$ since k represents the constant size of the jumps and $N(h)$ defines the number of jumps over the time interval h. Notice that the process $N(h)$ can be decomposed into the number of "continuations" $C(h)$, i.e., the number of jumps in the same direction as the previous jump, and the number of "alternations" $A(h)$, i.e., the number of jump reversals. Under assumptions, Large (2006) shows that the integrated variance

[17] Realized r-order power variation is defined as $PV_{(r)} = M^{-1+r/2}\sum_{j=1}^{M}|r_{j\delta}^*|^r$. The limiting properties of $PV_{(r)}$ are studied in Jacod (1994) and Barndorff-Nielsen and Shephard (2003, 2004a). See Barndorff-Nielsen et al. (2006a) for discussions.

of the underlying, unobservable semimartingale price process V can be consistently estimated using the quadratic variation of observed price process. Specifically, a consistent (in an asymptotic theory assuming increasingly frequent observations and small jumps) estimator can be defined by computing $[p]_h \frac{C(h)}{A(h)}$. While the quadratic variation of the observed price $[p]_h$ is generally a biased estimate of the quadratic variation of the underlying equilibrium price $[p^*]_h$, the bias can be corrected by using the factor $\frac{C(h)}{A(h)}$. The intuition goes as follows. The quadratic variation of the observed price process provides important information about the quadratic variation of the unobserved equilibrium price unless most of the jumps are alternations, for instance. In this case, $[p]_h$ will be an upward biased estimate of $[p^*]_h$. The correction factor $\frac{C(h)}{A(h)}$ will then act as a deflator.

In light of the local constancy of the observed price in the presence of an ever-evolving underlying equilibrium price, this approach captures the "mechanical effect" described in Bandi and Russell (2006b) yielding noise dependence and negative correlation between the noise and the underlying efficient price. The model's maintained assumption is that the observed prices change by fixed amounts or can be reduced, possibly by virtue of "rounding," to changes by fixed amounts. The practical applicability of this promising method will then depend on the nature of the data and hence on the price formation mechanism in specific markets. In general, an attentive analysis of the markets' fine grain dynamics has the potential to furnish important information about the process leading to market frictions. This information should be put to work to justify the use of different modeling and estimation approaches to integrated variance estimation.

5.3 Multivariate models

The provision of methods intended to identify integrated covariances and betas in the presence of market microstructure noise contaminations represents a necessary next step for the effective practice of portfolio choice and risk management through high-frequency asset price data. Barndorff-Nielsen and Shephard (2004b) study the asymptotic properties of realized covariance, i.e., the sum of the cross-products between two asset's calendar time returns over a period (a natural extension of the notion of "realized variance"), and realized beta in the frictionless case. To fix ideas, consider a second continuous stochastic volatility semimartingale price process $p^*_{(2)t}$ and re-define the original price process as $p^*_{(1)t}$. Assume, for simplicity, that the dynamics of the two price processes are driven by the same, scalar Brownian motion. The realized covariance (over h) between the original price 1 and price 2 is naturally defined as $\widehat{C}_{(1)(2)} = \sum_{j=1}^{M} r^*_{(1)j\delta} r^*_{(2)j\delta}$, where $r^*_{(u)j\delta} = p^*_{(u)j\delta} - p^*_{(u)(j-1)\delta}$ with $u = 1, 2$, and, as earlier, $\delta = h/M$. Similarly, the realized beta between asset 1 and asset 2 is defined as $\widehat{B}_{(1)(2)} = \widehat{C}_{(1)(2)}/(\sqrt{\widehat{V}_{(1)}\widehat{V}_{(2)}})$. In the absence of fric-

tions, Barndorff-Nielsen and Shephard (2004b) show that $\widehat{C}_{(1)(2)}$ is consistent for $\int_0^h \sigma_{(1)s}\sigma_{(2)s}\,\mathrm{d}s$, i.e., (the increment of) the covariation process between price 1 and price 2 over h, and asymptotically mixed-normally distributed with an estimable limiting variance as $M \to \infty$. The corresponding results in the $\widehat{B}_{(1)(2)}$ case follow from the consistency of the realized covariance and variance estimates, as well as from their joint mixed normality, in the no noise case. Barndorff-Nielsen et al. (2006a) contains a comprehensive discussion of these (and other) findings.

New issues arise in practice when computing high-frequency estimates of integrated covariances and betas. Information arrives at different frequencies for different assets, thereby leading to an additional microstructure effect having to do with nonsynchronicity in the underlying price formation processes. Even abstracting from the presence of a noise component as in previous sections, nonsynchronous trading leads to downward biased realized covariance estimates when sampling continuously-compounded returns in calendar time at high frequencies. This is the so-called Epps (1979) effect. The asset-pricing literature has long recognized the importance of this effect. Scholes and Williams (1977), Dimson (1979), and Cohen et al. (1983), among others, use leads and lags in nonparametric covariance measures to adjust for nonsynchronous trading. Martens (2005) reviews the early work on the subject. In the realized covariance case, the adjusted estimator with U lags and L leads can be simply defined as $\widehat{C}^{UL}_{(1)(2)} = \sum_{j=1}^{M}\sum_{s=-L}^{U} r^*_{(1)j\delta} r^*_{(2)(j-s)\delta}$. The logic behind this adjustment is well known (see, e.g., Cohen et al., 1983). Assume the equilibrium returns are martingale difference sequences ($\alpha_t = 0$). Then, $\widehat{C}^{UL}_{(1)(2)}$ is virtually unbiased for the true covariation over the period provided U and L are large enough. If U and L are small, then lack of price updates for either stock is bound to induce (downward) biases.

Initial work on realized covariance estimation in the presence of noisy high-frequency data is contained in Bandi and Russell (2005a) and Martens (2005). Bandi and Russell (2005a) study MSE-based optimal sampling for the purpose of realized covariance and beta estimation. Nonsynchronicity is accounted for by adding leads and lags to the optimized realized covariance estimator. Future research should study direct MSE-based optimization of the lead–lag estimator (for a certain number of leads and lags) as well as optimal choice of the number of leads and lags when noise is present. As is well known, the inclusion of a large number of leads and lags improves the bias properties of the estimator but increases its variability. Martens (2005) studies the MSE properties of a variety of covariance estimators (including realized covariance relying on equally-spaced returns and lead–lag estimators) through simulations based on Lo and MacKinlay's (1990) nonsynchronous trade model.

Recently, Hayashi and Yoshida (2005, 2006), Sheppard (2006), and Zhang (2006b), among others, have introduced promising, alternative approaches to high-frequency covariance estimation. The Hayashi and Yoshida's estimator, for instance, sums the products of all *overlapping* tick-by-tick returns rather

than the products of the calendar time returns, as is the case for realized covariance. Specifically, the estimator is defined as

$$\sum_{j=1}^{\#} \sum_{s \in S_j} r^*_{(1)j} r^*_{(2)s}, \tag{35}$$

where $r^*_{(u)j} = p^*_{(u)j} - p^*_{(u)(j-1)}$ with $u = 1, 2$; $S_j = \{s | (t^{(1)}_{j-1}, t^{(1)}_j) \cap (t^{(2)}_{s-1}, t^{(2)}_s) \neq 0\}$, the t_j's are transaction times and # denotes the number of transactions for asset 1 over the period. In the absence of classical microstructure noise contaminations, but in the presence of nonsynchronous trading, the Hayashi and Yoshida estimator is consistent and asymptotically normally distributed as the number of observations increases without bound over the trading day (Hayashi and Yoshida, 2005).

Voev and Lunde (2007) and Griffin and Oomen (2006) provide thorough finite sample studies of the MSE properties of several covariance estimators, including realized covariance, optimally-sampled realized covariance, and the Hayashi–Yoshida estimator, as well as recommendations for practical implementations.

Much remains to be done. While the first-order issues in high-frequency covariance estimation are likely to be fully understood, the methods and solutions are still in constant evolution. Arguably, the main goal of the literature is to provide reliable forecasts of *large* covariance matrices. We are far from this goal. On the one hand, the notion of reliability depends on the adopted metric (see below for discussions). On the other hand, the dimensionality of problems of practical interest continues to pose substantial issues when relying on high-frequency nonparametric estimates. Considerable effort is now being devoted to obtaining unbiased and efficient, in-sample, high-frequency covariance estimates. We welcome this effort and emphasize that out-of-sample performance will ultimately be the judge.

5.4 Forecasting and economic metrics

Understandably, the initial work on integrated variance estimation by virtue of high-frequency data was largely motivated by volatility prediction (see, e.g., Andersen et al., 2003, 2004, 2005, and the references therein). In the no noise case, high-frequency volatility forecasting using alternative reduced-form models, as well as alternative integrated variance estimators, has been successfully conducted by Ghysels et al. (2006) and Forsberg and Ghysels (2007), among many others (Andersen et al., 2006a, review this literature).

The noise case is now receiving substantial attention. Bandi and Russell (2006a) and Bandi et al. (2006) employ reduced-form models to show that optimally-sampled realized variances (covariances) outperform realized variances (covariances) constructed using ad hoc intervals in predicting variances (covariances) out-of-sample (see, also, Andersen et al., 2006b). Ghysels and

Sinko (2006) use the MIDAS approach of Ghysels et al. (2006) to evaluate the relative performance of realized variance based on fixed intervals, bias-corrected realized variance as in Eq. (17) above, and power variation. Confirming findings in Ghysels et al. (2006), their results point to the superior out-of-sample performance of power variation. Large (2006) employs the HAR-RV model of Corsi (2003), as in Andersen et al. (2007), to stress that his "alternation estimator" can have better forecasting properties than realized variance constructed using fixed, arbitrary intervals.

More work ought to be done. On the one hand, a comprehensive study using a variety of variance/covariance measures and reduced-form models appears to be needed. To this day, the literature appears to solely agree on the fact that realized variance constructed using ad hoc fixed intervals is generally dominated, in terms of forecasting performance, by alternative measures. A complete comparison between these alternative measures, including optimally-sampled realized variance, optimally-sampled bias-corrected realized variance, and consistent kernel estimators, appears to be an important topic for future empirical work on the subject. On the other hand, as forcefully emphasized by Bandi and Russell (2006b), assessing the out-of-sample performance of alternative variance estimates using relevant economic metrics is arguably the single most important test in the literature. Thus far, two economic metrics have been proposed. Bandi and Russell (2006a) consider a *portfolio allocation* problem and the long-run utility that a mean-variance representative investor derives from alternative variance forecasts as the relevant performance criterion. The same portfolio-based approach has been recently implemented by Bandi et al. (2006) and De Pooter et al. (2006) in a multivariate context (see Fleming et al. 2001, 2003, and West et al., 1993, in the no-noise case). Bandi and Russell (2005b, 2006c) study volatility forecasting for the purpose of *option pricing* in the context of a simulated derivative market (see Engle et al., 1990, in the no-noise case). In agreement with the forecasting results derived from reduced-form models, the use of economic metrics indicates that optimally-sampled realized variances (covariances) have the potential to substantially outperform realized variances (covariances) based on fixed intervals. In addition, optimally-sampled realized variances can yield more accurate forecasts than certain consistent kernel estimators (such as the two-scale estimator) when these estimators are implemented using asymptotic bandwidth selection methods. Consistent and "near-consistent" kernel estimators that are implemented at their full potential on the basis of finite sample criteria (as recommended by Bandi and Russell, 2005b) are likely to dominate optimally-sampled realized variance in the context of the above-mentioned metrics. Again, future work on the subject should provide a more comprehensive study focusing on a variety of suggested measures. In addition, alternative economic metrics should be investigated.

6 The variance of microstructure noise: a consistency result

Even though the classical realized variance estimator is not a consistent estimator of the integrated variance of the underlying equilibrium price, a re-scaled version of the standard realized variance estimator is, under assumptions, consistent for the variance of the noise return component (Bandi and Russell, 2003, and Zhang et al., 2005). More generally, sample moments of the observed return data can estimate moments of the underlying noise return process at high-frequencies (see Eq. (14) above). Bandi and Russell (2003) discuss this result and use it to characterize the MSE of the conventional realized variance estimator.

While the literature on integrated variance estimation focuses on the volatility features of the underlying equilibrium price, the empirical market microstructure literature places emphasis on the *other* component of the observed price process in Eq. (1), namely the price frictions η. When p is a transaction price, such frictions can be interpreted in terms of transaction costs in that they constitute the difference between the observed price p and the corresponding equilibrium price p^*.[18] Hasbrouck (1993) and Bandi and Russell (2004) provide related, but different, frameworks to use transaction price data in order to estimate the second moment of the transaction cost η (rather than moments of ε as generally needed in the integrated variance literature) under mild assumptions on the features of the price formation mechanism in Section 2. The implications of their results for measuring transaction costs are the subject of the next sections. We start with a discussion of traditional approaches to transaction cost evaluation.

7 The benefit of consistency: measuring market quality

7.1 Transaction cost estimates

Following Perold (1988), it is generally believed that an ideal measure of the execution cost of a trade should be based on the comparison between the trade price for an investor's order and the equilibrium price prevailing at the time of the trading decision. Although informed individual investors can plausibly construct this measure, researchers and regulators do not have enough information to do so (see Bessembinder, 2003, for discussions).

Most available estimates of transaction costs relying on high-frequency data hinge on the basic logic behind Perold's original intuition. Specifically, there are three high-frequency measures of execution costs that have drawn atten-

[18] Measuring the execution costs of stock market transactions and understanding their determinants is of importance to a variety of market participants, such as individual investors and portfolio managers, as well as regulators. In November 2000, the Security and Exchange Commission issued Rule 11 Ac. 1-5 requesting market venues to widely distribute (in electronic format) execution quality statistics regarding their trades.

tion in recent years, i.e., the so-called *bid–ask half spread*, the *effective half spread*, and the *realized half spread*. The bid–ask half spread is defined as half the difference between ask quote and bid quote. The effective half spread is the (signed[19]) difference between the price at which a trade is executed and the mid-point of reference bid–ask quotes. As for the realized half spread, this measure is defined as the (signed) difference between the transaction price and the mid-point of quotes in effect some time after the trade.[20] In all cases, an appropriately chosen bid–ask mid-point is used as an approximation for the relevant equilibrium price.

The limitations of these measures of the cost of trade have been pointed out in the literature (the interested reader is referred to the special issue of the Journal of Financial Markets on transaction cost evaluation, JFM 6, 2003, for recent discussions). The bid–ask half spread, for example, is known to overestimate the true cost of trade in that trades are often executed at prices within the posted quotes. As for the effective and realized spreads, not only do they require the trades to be signed as buyer or seller-initiated, but they also require the relevant quotes and transaction prices to be matched.

The first issue (assigning the trade direction) arises due to the fact that commonly used high-frequency data sets (the TAQ database, for instance) do not contain information about whether a trade is buyer or seller-initiated. Some data sets do provide this information (the TORQ database being an example) but the length of their time series is often insufficient. Naturally, then, a considerable amount of work has been devoted to the construction of algorithms intended to classify trades as being buyer of seller-initiated simply on the basis of transaction prices and quotes (see, e.g., Lee and Ready, 1991, and Ellis et al., 2000). The existing algorithms can of course missclassify trades (the Lee and Ready method, for example, is known to categorize incorrectly about 15% of the trades), thereby inducing biases in the final estimates. See Bessembinder (2003) and Peterson and Sirri (2002) for discussions.

The second issue (matching quotes and transaction prices) requires potentially arbitrary judgment calls. Since the trade reports are often delayed, when computing the effective spreads, for example, it seems sensible to compare the trade prices to mid-quotes occurring before the trade report time. The usual allowance is 5 seconds (see, e.g., Lee and Ready, 1991) but longer lags can of course be entertained.

This said, there is a well-known measure which can be computed using low frequency data and does not require either the signing of the trades or the matching of quotes and transaction prices, i.e., *Roll's effective spread estimator* (Roll, 1984). Roll's estimator does not even rely on the assumption that the mid-points of the bid and ask quotes are valid proxies for the unobserved equi-

[19] Positive for buy orders and negative for sell orders.

[20] The idea is that the traders possess private information about the security value and the trading costs should be assessed based on the trades' non-informational price impacts.

librium prices. The idea behind Roll's measure can be easily laid out using the model in Section 2. Write the model in transaction time. Assume

$$\eta_i = sI_i, \tag{36}$$

where I_i equals 1 for a buyer-initiated trade and -1 for a seller-initiated trade with $p(I_i = 1) = p(I_i = -1) = \frac{1}{2}$. If Assumption 1 (with $\alpha_t = 0$) and Assumption 2b are satisfied, then

$$\mathbf{E}(r, r_{-1}) = -s^2. \tag{37}$$

Equivalently,

$$s = \sqrt{-\mathbf{E}(r, r_{-1})}. \tag{38}$$

Thus, the constant width of the spread can be estimated consistently based on the (negative) first-order autocovariance of the observed (low-frequency) stock returns.

Roll's estimator hinges on potentially restrictive assumptions. The equilibrium returns r^* are assumed to be serially uncorrelated. More importantly, the microstructure frictions in the observed returns r follow an $MA(1)$ structure, as largely implied by bid–ask bounce effects, with a constant cost of trade s. Finally, the estimator relies on the microstructure noise components being uncorrelated with the equilibrium prices.

7.2 Hasbrouck's pricing errors

Hasbrouck (1993) assumes the price formation mechanism in Eq. (1). However, his setup is in discrete time and time is measured in terms of transaction arrival times. Specifically, the equilibrium price p^* is modeled as a random walk while the η's, which may or may not be correlated with p^*, are mean-zero covariance stationary processes. Hence, he considerably relaxes the assumptions that are needed to derive the classical Roll effective spread estimator. Hasbrouck (1993) interprets the difference η between the transaction price p and the equilibrium price p^* as a *pricing error* impounding microstructure effects. The standard deviation of the pricing error σ_η is the object of interest. Because stocks whose transaction prices track the equilibrium price can be regarded as being stocks that are less affected by barriers to trade, σ_η is thought to represent a natural measure of market quality.

Using methods in the tradition of Beveridge and Nelson (1981) and Watson (1986) to study nonstationary time series (the observed price p in this case) expressed as the sum of a nonstationary component (the equilibrium price p^*) and a residual stationary component (the pricing error η), Hasbrouck (1993) provides estimates (and lower bounds) for σ_η. His empirical work focuses on NYSE stocks and utilizes transaction data collected from the Institute for the Study of Securities Markets (ISSM) tape for the first quarter of 1989. His (average) estimated σ_η value is equal to about 33 basis points. Under an assumption of normality, the average value for the expected transaction cost $\mathbf{E}|\eta|$ is equal to about 26 basis points ($\frac{2}{\sqrt{\pi}}\sigma_\eta \approx 0.8\sigma_\eta$) in his data.

7.3 Full-information transaction costs

Bandi and Russell (2004) define an alternative notion of pricing error. Their approach imposes more economic structure on the model in Section 2. They begin by noting that in a rational expectation setup with asymmetric information two equilibrium prices can be defined in general: an "efficient price," i.e., the price that would prevail in equilibrium given public information, and a "full-information price," the price that would prevail in equilibrium given private information. Both the efficient price and the full-information price are unobservable. The econometrician only observes transaction prices.

In this setting, two sources of "market inefficiency" arise. First, transaction prices deviate from efficient prices due to classical market microstructure frictions (see Stoll's AFA presidential address, Stoll, 2000, for discussions). Second, the presence of asymmetric information induces deviations between efficient prices and full-information prices. Classical approaches to transaction cost evaluation (in Section 7.1) and Hasbrouck's important approach to pricing error estimation (in Section 7.2) refer to the efficient price as the relevant equilibrium price. Hence, these methods are meant to only account for the first source of inefficiency.[21]

A cornerstone of market microstructure theory is that uninformed agents learn about existing private information from observed order flow (see, e.g., the discussions in O'Hara, 1995). Since each trade carries information, meaningful revisions to the efficient price are made regardless of the time interval between trade arrivals. Hence, the efficient price is naturally thought of as a process changing discretely at transaction times. Contrary to the public-information set, the full-information set, by definition, contains all information used by the agents in their decisions to transact. Hence, the full-information price is unaffected by past order flow. Barring occasional news arrivals to the informed agents, the dynamic behavior of the full information price is expected to be relatively "smooth." As for the microstructure frictions, separate prices for buyers and sellers and discreteness of prices *alone* suggest that changes in the microstructure frictions from trade to trade are discrete in nature.

Bandi and Russell (2004) formalize these ideas by writing the model in Section 2 in transaction time. They add structure to the specification in Eq. (1) in order to account for the desirable properties of efficient price, full-information price, and microstructure frictions. Specifically, write

$$p_i = p^*_{t_i} + \eta_i \tag{39}$$

$$= p^*_{t_i} + \eta_i^{\text{asy}} + \eta_i^{\text{fri}}, \tag{40}$$

[21] There is of course a sense in which realized spreads can capture both components with a suitably chosen lagged midpoint. It is plausible that the notional efficient price used in Hasbrouck's approach could also be viewed as a full-information price. Both issues are worth further exploration.

where $p^*_{t_i}$ is now the (logarithmic) full-information price, $p^*_{t_i} + \eta_i^{\text{asy}}$ is the discretely-evolving (logarithmic) efficient price, and η_i^{fri} denotes conventional (discrete) microstructure frictions. The deviation η_i includes a classical friction component η_i^{fri} and a pure asymmetric information component η_i^{asy}. The former is affected by both liquidity and asymmetric information,[22] the latter should only be affected by asymmetric information. As in Section 2, it is convenient to rewrite the model in terms of observed continuously-compounded returns, i.e.,

$$r_i = r^*_{t_i} + \varepsilon_i, \tag{41}$$

where $r_i = p_i - p_{i-1}$, $r^*_{t_i} = p^*_{t_i} - p^*_{t_{i-1}}$, and $\varepsilon_i = \eta_i - \eta_{i-1}$. At very high frequencies, the observed return data (the r_i's) are dominated by return components that are induced by the microstructure effects (the ε_i's) since the full-information returns evolve smoothly in time. Technically, $r^*_{t_i} = O_p(\sqrt{\max|t_i - t_{i-1}|})$ and $\varepsilon_i = O_p(1)$. In this context, Bandi and Russell (2004) employ sample moments of the *observed* high-frequency return data to identify the moments of the *unobserved* trading cost η_i. They do so by using the informational content of observed return data whose full-information return component $r^*_{t_i}$ is largely swamped by the component ε_i when sampling is conducted at the high frequencies at which transactions occur in practice.

Assume the covariance structure of the η's is such that $\mathbf{E}(\eta\eta_{-j}) = \theta_j \neq 0$ for $j = 1, \ldots, k < \infty$ and $\mathbf{E}(\eta\eta_{-j}) = 0$ for $j > k$. This structure accommodates temporal dependence possibly induced by clustering in order flow. It is then easy to show that

$$\sigma_\eta = \sqrt{\left(\frac{1+k}{2}\right)\mathbf{E}\left(\varepsilon^2\right) + \sum_{s=0}^{k-1}(s+1)\mathbf{E}(\varepsilon\varepsilon_{-k+s})}. \tag{42}$$

For every sample period (a trading day, for instance), an estimate of σ_η can thus be obtained by replacing the moments of the *unobserved* contaminations ε with the corresponding sample moments of the *observed* returns. At very high frequencies (represented here by a large number of observations for each period), the full-information return component of each sample moment is expected to be negligible. Formally,

$$\widehat{\sigma}_\eta = \sqrt{\left(\frac{k+1}{2}\right)\left(\frac{\sum_{i=1}^{\widetilde{M}} r_i^2}{\widetilde{M}}\right) + \sum_{s=0}^{k-1}(s+1)\left(\frac{\sum_{i=k-s+1}^{\widetilde{M}} r_i r_{i-k+s}}{\widetilde{M}-k+s}\right)} \xrightarrow[\widetilde{M}\to\infty]{p} \sigma_\eta, \tag{43}$$

[22] Market microstructure theory imputes classical frictions to operating (order-processing and inventory-keeping) costs and adverse selection. See, for example, the discussion in Stoll (2000).

where $\widetilde{M}$ is now the total number of transactions over the period.[23] This consistency result only relies on the different stochastic orders of efficient price, full-information price, and classical frictions. These orders are simply meant to formalize the economics of price formation in markets with asymmetric information. The result is robust to predictability in the underlying full-information return process ($\alpha_t \neq 0$), presence of infrequent jumps in the full-information price, dependence between the full-information price and the frictions as well as, of course, dependence in the frictions.

When the number of observations for each time period is not large enough, the potential for (finite sample) contaminations in the estimates due to the presence of a non-negligible full-information price component is higher. Bandi and Russell (2004) suggest a finite sample adjustment.[24]

Bandi and Russell (2004) use the convention of calling the standard deviation $\widehat{\sigma}_\eta$, rather than the actual η, *full-information transaction cost*, or *FITC*. While the *FITCs* are standard deviations, one can either assume normality of the η's (as done in Hasbrouck, 1993, for similar purposes) or use the approach in Roll (1984) to derive expected costs. In the former case, a consistent estimate of $\mathbf{E}|\eta|$ can be obtained by computing $\frac{2}{\sqrt{\pi}}\widehat{\sigma}_\eta$. In the latter case, assume $\eta = sI$, where the random variable I, defined in Section 7.1, represents now the direction (higher or lower) of the transaction price with respect to the full-information price and s is a constant full-information transaction cost. Then, $\widehat{\sigma}_\eta$ consistently estimates s.

Using a sample of S&P 100 stocks over the month of February 2002, Bandi and Russell (2004) report an average value for $\widehat{\sigma}_\eta$ equal to 14 basis points. Under normality, their estimated average $\mathbf{E}|\eta|$ is then equal to about 11 basis points. This value is larger that the corresponding average effective spread (about 6 basis points). Consistent with the economic interpretation underlying the construction of the *FITC*s, Bandi and Russell (2004) find that the *FITC*s are cross-sectionally more correlated with private information proxies, such as the *PIN* measure of Easley and O'Hara (see, e.g., Easley et al., 1996), the turnover ratio (Stoll, 1989), and the number of analysts following the stock, than the average effective spreads and the average half bid–ask spreads. Furthermore, they find that the deviations of the efficient prices from the full-information prices, as determined by the existence of private information in the market

[23] Under uncorrelatedness of the full-information returns, k can be estimated based on the dependence properties of the observed returns.

[24] In the absence of correlation between the frictions η and the full-information price p^* the bias-adjustment is relatively straightforward and can be implemented by using nonparametric estimates of the full-information price variance as described in Footnote 10, Section 4.3. In the presence of correlation between η and p^*, as typically implied by models with learning, a complete bias-correction requires parametric assumptions on the underlying full-information price process. Bandi and Russell (2004) use the price formation mechanism proposed by Hasbrouck and Ho (1987) to quantify the estimates' finite sample bias.

place, can be as large as the departures of the transaction prices from the efficient prices.

Assume now σ_η is stochastic and latent. In keeping with the logic behind the vibrant and successful realized variance literature initiated by Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002), the high-frequency approach suggested in Bandi and Russell (2004) can be interpreted as providing a method to render the latent noise volatility observable (i.e., easily estimable without filtering) for each period of interest. While the realized variance literature has placed emphasis on the volatility of the underlying equilibrium price process, one can focus on the other volatility component of the observed returns, i.e., microstructure noise volatility. Treating the volatility of the noise component of the observed prices as being directly observable can allow one to address a broad array of fundamental issues. Some have a statistical flavor having to do with the distributional and dynamic properties of the noise variance and its relationship with the time-varying variance of the underlying equilibrium price process. Some have an economic importance having to do with the dynamic determinants of the cost of trade. Since the most salient feature of the quality of a market is how much agents have to pay in order of transact, much can be learned about the genuine market dynamics by exploiting the informational content of the estimated noise variances.

8 Volatility and asset pricing

Barring complications induced by the shorter observation span of asset price data sampled at high frequencies, the methods described in the previous sections can have important implications for the cross-sectional asset pricing literature.

A promising, recent strand of this literature has been devoted to assessing whether stock *market volatility* is priced in the cross-section of stock returns. Being innovations in volatility correlated with changes in investment opportunities, this is a relevant study to undertake. Ang et al. (2005), Adrian and Rosenberg (2006), and Moise (2004), among others, find that the price of market volatility risk is negative. Volatility is high during recessions. Stocks whose returns covary with innovations in market volatility are stocks which pay off during bad times. Investors are willing to pay a premium to hold them. The results in Ang et al. (2005), Adrian and Rosenberg (2006), and Moise (2004) are robust to the use of alternative, parametric and nonparametric, low-frequency volatility estimates. In virtue of the potential accuracy of the newly-developed high-frequency volatility measures, as described above, it is now of interest to re-evaluate the importance of market volatility as a systematic risk factor by studying the cross-sectional pricing implications of these measures. In this context, market microstructure issues ought to be accounted for.

Another strand of this literature has focused on the pricing implication of *market liquidity*. As is the case for market volatility, innovations in liquidity are

correlated with the business cycle. Stocks yielding high returns when illiquidity is high provide a hedge. Not surprisingly, the price of market illiquidity risk is found to be negative (see, e.g., Acharya and Pedersen, 2005, and Pástor and Stambaugh, 2003). Liquidity is hard to measure. The recent advances in high-frequency volatility estimation provide a rich set of tools to separate liquidity-induced components (named "market microstructure frictions" earlier) from the estimated moments of the observed high-frequency asset returns. When aggregated across stocks (for each period of interest), these components have the potential to provide important information about the time-series properties of the overall level of market liquidity. These properties can be put to work to better understand the pricing of (il-)liquidity risk from a novel standpoint.

The pricing of idiosyncratic risk is also of interest. Since individuals are likely to take into account the cost of acquiring and rebalancing their portfolios, expected stock returns should somehow embed idiosyncratic transaction costs in equilibrium. This observation has given rise to a convergence between classical market microstructure work on price determination and asset pricing in recent years (the interested reader is referred to the recent survey of Easley and O'Hara, 2002). The studies on the cross-sectional relationship between expected stock returns and cost of trade largely hinge on liquidity-based theories of execution cost determination (Amihud and Mendelson, 1986; Brennan and Subrahmanyam, 1996; Datar et al., 1998; and Hasbrouck, 2003; among others). Alternatively, some recent studies rely on information-based approaches to the same issue (see, e.g., Easley et al., 2002). Much remains to be done. Full-information transaction costs, among other tools discussed earlier, may provide a promising bridge between liquidity-based and information-based arguments.

Generally speaking, the convergence between market microstructure theory and methods and asset pricing is still in its infancy. We are convinced that the recent interest in microstructure issues in the context of volatility estimation is providing, and will continue to provide, an important boost to this inevitable process of convergence.

Acknowledgements

We survey new and fast-developing fields. We tried our best to keep this review updated until publication and apologize for unwanted omissions. We thank the William S. Fishman Faculty Research Fund at the Graduate School of Business of the University of Chicago (Bandi) and the Graduate School of Business at the University of Chicago (Russell) for financial support.

References

Acharya, V.V., Pedersen, L.H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics* 77, 375–410.

Adrian, T., Rosenberg, J. (2006). Stock returns and volatility: Pricing the short-run and long-run components of market risk. Working paper.
Aït-Sahalia, Y., Mykland, P., Zhang, L. (2005a). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* 18, 351–416.
Aït-Sahalia, Y., Mykland, P., Zhang, L., (2005b). Ultra high-frequency volatility estimation with dependent microstructure noise. Working paper.
Alizadeh, S., Brandt, M., Diebold, F. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance* 57, 1047–1091.
Amihud, Y., Mendelson, H. (1986). Asset pricing and the bid–ask spread. *Journal of Financial Economics* 17, 223–249.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (1999). (Understanding, optimizing, using, and forecasting) Realized volatility and correlation. Working paper.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2000). Great realizations. *Risk Magazine* 13, 105–108.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics* 61, 43–76.
Andersen, T.G., Bollerslev, T., Diebold, F.X. (2002). Parametric and nonparametric measurements of volatility. In: Aït-Sahalia, Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*, Elsevier, North-Holland. In press.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* 71, 579–625.
Andersen, T.G., Bollerslev, T., Meddahi, N. (2004). Analytic evaluation of volatility forecasts. *International Economic Review* 45, 1079–1110.
Andersen, T.G., Bollerslev, T., Meddahi, N. (2005). Correcting the errors: A note on volatility forecast evaluation based on high-frequency data and realized volatilities. *Econometrica* 73, 279–296.
Andersen, T.G., Bollerslev, T., Christoffersen, P.F., Diebold, F.X. (2006a). Volatility and correlation forecasting. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*. Elsevier, North-Holland, pp. 778–878.
Andersen, T.G., Bollerslev, T., Meddahi, N. (2006b). Market microstructure noise and realized volatility forecasting. Working paper.
Andersen, T.G., Bollerslev, T., Diebold, F.X. (2007). Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *Review of Economics and Statistics*, in press.
Andrews, D.J.C., Monahan, W.K. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60, 953–966.
Ang, A., Hodrick, R., Xing, Y., Zhang, X. (2005). The cross-section of volatility and expected returns. *Journal of Finance* 61, 259–299.
Awartani, B., Corradi, V., Distaso, W. (2004). Testing and modelling market microstructure effects with an application to the Dow Jones Industrial Average. Working paper.
Bai, X., Russell, J.R., Tiao, G. (2005). Effects of non-normality and dependence on the precision of variance estimates using high-frequency data. Working paper.
Bandi, F.M., Russell, J.R. (2003). Microstructure noise, realized volatility, and optimal sampling. Working paper.
Bandi, F.M., Russell, J.R. (2004). Full-information transaction costs. Working paper.
Bandi, F.M., Russell, J.R. (2005a). Realized covariation, realized beta, and microstructure noise. Working paper.
Bandi, F.M., Russell, J.R. (2005b) Market microstructure noise, integrated variance estimators, and the accuracy of asymptotic approximations. Working paper.
Bandi, F.M., Russell, J.R. (2006a). Separating microstructure noise from volatility. *Journal of Financial Economics* 79, 655–692.
Bandi, F.M., Russell, J.R. (2006b). Comment on Hansen and Lunde. *Journal of Business and Economic Statistics* 24, 167–173.
Bandi, F.M., Russell, J.R. (2006c). Microstructure noise, realized variance, and optimal sampling. Working paper.

Bandi, F.M., Russell, J.R., Zhu, Y. (2006). Using high-frequency data in dynamic portfolio choice. *Econometric Reviews*, in press.

Barndorff-Nielsen, O.E., Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 64, 253–280.

Barndorff-Nielsen, O.E., Shephard, N. (2003). Realized power variation and stochastic volatility. *Bernoulli* 9, 243–265.

Barndorff-Nielsen, O.E., Shephard, N. (2004a). Power and bypower variation with stochastic volatility and jumps (with discussions). *Journal of Financial Econometrics* 2, 1–48.

Barndorff-Nielsen, O.E., Shephard, N. (2004b). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* 72, 885–925.

Barndorff-Nielsen, O.E., Shephard, N. (2007). Variation, jumps, market frictions and high-frequency data in financial econometrics. In: Blundell, R., Torsten, P., Newey, W.K. (Eds.), *Advances in Economics and Econometrics. Theory and Applications*. In: *Econometric Society Monographs*. Cambridge Univ. Press. Ninth World Congress.

Barndorff-Nielsen, O.E., Hansen, P., Lunde, A., Shephard, N. (2005). Regular and modified kernel-based estimators of integrated variance: The case with independent noise. Working paper.

Barndorff-Nielsen, O.E., Graversen, S.E., Jacod, J., Shephard, N. (2006a). Limit theorems for realized bypower variation in econometrics. *Econometric Theory* 22, 677–719.

Barndorff-Nielsen, O.E., Hansen, P., Lunde, A., Shephard, N. (2006b). Designing realized kernels to measure ex post variation of equity prices in the presence of noise. Working paper.

Barndorff-Nielsen, O.E., Shephard, N., Winkel, M. (2006c). Limit theorems for multipower variation in the presence of jumps in financial econometrics. Stochastic Processes and their Applications, in press.

Barucci, E., Renò, R. (2002a). On measuring volatility and the GARCH forecasting performance. *Journal of International Financial Markets, Institutions and Money* 12, 182–200.

Barucci, E., Renò, R. (2002b). On measuring volatility of diffusion processes with high-frequency data. *Economic Letters* 74, 371–378.

Bessembinder, H. (2003). Issues in assessing trade execution costs. *Journal of Financial Markets* 6, 233–257.

Beveridge, S., Nelson, C. (1981). A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to the measurement of the "Business Cycle". *Journal of Monetary Economics* 7, 151–174.

Bollen, B., Inder, B. (2002). Estimating daily volatility in financial markets utilizing intraday data. *Journal of Empirical Finance* 9, 551–562.

Brandt, M.W., Diebold, F.X. (2006). A no-arbitrage approach to range-based estimation of return covariances and correlations. *Journal of Business* 79, 61–74.

Brennan, M.J., Subrahmanyam, A. (1996). Market microstructure and asset pricing: on the compensation for illiquidity in stock returns. *Journal of Financial Economics* 41, 441–464.

Christensen, K., Podolskij, M. (2005). Asymptotic theory for range-based estimation of integrated volatility of a continuous semimartingale. Working paper.

Chung, K.L., Williams, R.J. (1990). *Introduction to Stochastic Integration*, second ed. Birkhäuser.

Cohen, K.J., Hawanini, G.A., Maier, S.F., Schwartz, R.A., Whitcomb, D.K. (1983). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics* 12, 263–278.

Corradi, V., Distaso, W. (2006). Semiparametric comparison of stochastic volatility models via realized measures. *Review of Economic Studies*, in press.

Corsi (2003). A simple long-memory model of realized volatility. Working paper.

Datar, V.T., Naik, N.Y., Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets* 1, 203–219.

De Pooter, M., Martens, M., Van Dijk, D. (2006). Predicting the daily covariance matrix for S&P100 stocks using intraday data: But which frequency to use? *Econometric Reviews*, in press.

Dimson, E. (1979). Risk management when shares are subject to infrequent trading. *Journal of Financial Economics* 7, 197–226.

Duffie, D. (1992). *Dynamic Asset Pricing Theory*. Princeton Univ. Press, Princeton.

Easley, D., O'Hara, M. (2002). Microstructure and asset pricing. In: Constantinides, G., Harris, M., Stulz, R. (Eds.), *Handbook of Financial Economics*. Elsevier, North-Holland.
Easley, D., Kiefer, N., O'Hara, M., Paperman, J. (1996). Liquidity, information, and infrequently traded stocks. *Journal of Finance* 1, 1405–1436.
Easley, D., Hvidkjaer, S., O'Hara, M. (2002). Is information risk a determinant of asset returns? *Journal of Finance* 57, 2185–2221.
Ellis, K., Michaely, R., O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis* 35, 529–552.
Engle, R.F., Hong, C.H., Kane, A. (1990). Valuation of variance forecasts with simulated option markets. Working paper No. 3350. NBER.
Epps, T.W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association* 74, 291–298.
Fang, Y. (1996). Volatility modeling and estimation of high-frequency data with Gaussian noise. Unpublished PhD thesis, MIT.
Fleming, J., Kirby, C., Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance* 56, 329–352.
Fleming, J., Kirby, C., Ostdiek, B. (2003). The economic value of volatility timing using "realized volatility". *Journal of Financial Economics* 67, 473–509.
Forsberg, L., Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, in press.
Ghysels, E., Sinko, A. (2006). Comment on Hansen and Lunde. *Journal of Business and Economic Statistics* 24, 192–194.
Ghysels, E., Santa-Clara, P., Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131, 59–95.
Goncalves, S., Meddahi, N. (2004). Bootstrapping realized volatility. Working paper.
Goncalves, S., Meddahi, N. (2006). Box–Cox transforms for realized volatility. Working paper.
Griffin, J., Oomen, R. (2006). Covariance measurement in the presence of nonsynchronous trading and market microstructure noise. Working paper.
Hansen, P.R., Lunde, A. (2006). Realized variance and market microstructure noise (with discussions). *Journal of Business and Economic Statistics* 24, 127–161.
Hasbrouck, J. (1993). Assessing the quality of a security market: A new approach to transaction cost measurement. *Review of Financial Studies* 6, 191–212.
Hasbrouck, J. (1996). Modelling market microstructure time series. In: Maddala, G.S., Rao, C.R. (Eds.), *Handbook of Statistics*. Elsevier, North-Holland.
Hasbrouck, J. (2003). Trading costs and returns for US securities: The evidence from daily data. Working paper.
Hasbrouck, J., Ho, T. (1987). Order arrival, quote behavior, and the return generating process. *Journal of Finance* 4, 1035–1048.
Hayashi, T., Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* 11, 359–379.
Hayashi, T., Yoshida, N. (2006). Estimating correlations with nonsynchronous observations in continuous diffusion models. Working paper.
Huang, X., Tauchen, G. (2005). The relative contribution of jumps to total price variation. *Journal of Financial Econometrics* 3, 456–499.
Jacod, J. (1994). Limit of random measures associated with the increments of a Brownian semimartingale. Working paper.
Jacod, J., Protter, P. (1998). Asymptotic error distributions for the Euler method for stochastic differential equations. *Annals of Probability* 26, 267–307.
Kalnina, I., Linton, O. (2006). Estimating quadratic variation consistently in the presence of correlated measurement error. Working paper.
Kanatani, T. (2004a). High-frequency data and realized volatility. Unpublished PhD thesis, Kyoto University.
Kanatani, T. (2004b). Integrated volatility measuring from unevenly sampled observations. *Economics Bulletin* 3, 1–8.

Large (2006). Estimating quadratic variation when quoted prices change by constant increments. Working paper.
Lee, C., Ready, M. (1991). Inferring trade direction from intraday data. *Journal of Finance* 46, 733–746.
Lo, A., MacKinlay, A.C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics* 45, 181–212.
Malliavin, P., Mancino, M.E. (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics* 6, 49–61.
Mancini, C. (2003). Statistics of a Poisson–Gaussian process. Working paper.
Mancini, C. (2004). Estimation of the characteristics of jump of a general Poisson-diffusion process. *Scandinavian Actuarial Journal* 1, 42–52.
Martens, M. (2005). Estimating unbiased and precise realized covariances. Working paper.
Martens, M., Van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics* 138, 181–207.
McAleer, M., Medeiros, M. (2007). Realized volatility: A review. *Econometrics Reviews*, in press.
Meddahi, N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics* 17, 475–508.
Moise, C. (2004). Stochastic volatility risk and the size anomaly. Working paper.
Müller, U.A. (1993). Statistics of variables observed over overlapping intervals. Working paper.
Mykland, P.A., Zhang, L. (2006). ANOVA for diffusions and Ito processes. *Annals of Statistics* 34, 1931–1963.
Newey, W., West, K. (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.
Nielsen, M.O., Frederiksen, P.H. (2005). Finite sample accuracy of integrated volatility estimators. Working paper.
O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell Sci., Oxford.
Oomen, R.C.A. (2005). Properties of bias-corrected realized variance under alternative sampling schemes. *Journal of Financial Econometrics* 3, 555–577.
Oomen, R.C.A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business and Economic Statistics* 24, 219–237.
Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business* 53, 61–66.
Pástor, L., Stambaugh, R.F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy* 111, 642–685.
Perold, A. (1988). The implementation shortfall. *Journal of Portfolio Management* 14, 4–9.
Peterson, M., Sirri, E. (2002). Evaluation of the biases in execution cost estimates using trade and quote data. *Journal of Financial Markets* 6, 259–280.
Phillips, P.C.B., Yu, J. (2005). Comment on Hansen and Lunde. *Journal of Business and Economic Statistics* 24, 202–208.
Politis, D., Romano, J.P., Wolf, M. (1999). *Subsampling*. Springer, New York.
Press, J.S. (1967). A compound events model for security prices. *Journal of Business* 40, 317–335.
Rogers, L.C.G., Satchell, S.E. (1991). Estimating variance from high, low, open, and close prices. *Annals of Applied Probability* 1, 504–512.
Roll, R. (1984). A simple measure of the effective bid–ask spread in an efficient market. *Journal of Finance* 39, 1127–1139.
Scholes, M., Williams, J. (1977). Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5, 309–327.
Sheppard, K. (2006). Realized covariance and scrambling. Working paper.
Stoll, H.R. (1989). Inferring the components of the bid–ask spread: Theory and empirical evidence. *Journal of Finance* 44, 115–134.
Stoll, H.R. (2000). Friction. *Journal of Finance* 55, 1479–1514.
Sun, Y. (2006). Best quadratic unbiased estimators of integrated variance. Working paper.
Voev, V., Lunde, A. (2007). Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics* 5, 68–104.

Watson, M. (1986). Univariate detrending methods with stochastic trends. *Journal of Monetary Economics* 18, 49–75.

West, K.D., Edison, H.J., Cho, D. (1993). A utility-based comparison of some models of exchange rate volatility. *Journal of International Economics* 35, 23–45.

Woerner, J. (2006). Power and multipower variation: Inference for high-frequency data. In: Shiryaev, A.N., do Rosario Grossinho, M., Oliviera, P., Esquivel, M. (Eds.), *Proceedings of the International Conference on Stochastic Finance 2004*. Springer-Verlag.

Zhang, L. (2006a). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* 12, 1019–1043.

Zhang, L. (2006b). Estimating covariation: Epps effect, microstructure noise. Working paper.

Zhang, L., Mykland, P., Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100, 394–1411.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business and Economic Statistics* 14, 45–52.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15006-4

Chapter 6

Spectral Methods in Derivatives Pricing*

Vadim Linetsky

Department of Industrial Engineering and Management Sciences, McCormick School of Engineering and Applied Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA
E-mail: linetsky@iems.northwestern.edu
url: http://users.iems.northwestern.edu/~linetsky

Abstract

In this chapter we study the problem of valuing a (possibly defaultable) derivative asset contingent on the underlying economic state modeled as a Markov process. To gain analytical and computational tractability both in order to estimate the model from empirical data and to compute the prices of derivative assets, financial models in applications are often Markovian. In the Markovian framework, the key object is the *pricing operator* mapping (possibly defaultable) future payments (payoffs) into present values. The pricing operators indexed by time form a *pricing semigroup* $\{\mathcal{P}_t, t \geqslant 0\}$ in an appropriate payoff space, which can be interpreted as the transition semigroup of the underlying Markov process with the killing rate equal to the default-free interest rate plus default intensity. This framework encompasses a wide range of Markovian financial models. In applications it is important to have a tool kit of analytically tractable Markov processes with known transition semigroups that lead to closed-form expressions for value functions of derivative assets. In particular, an analytical simplification is possible when the process is a *symmetric Markov process* in the sense that there is a measure m on the state space D and the semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ is symmetric in the Hilbert space $L^2(D, m)$. In this case we apply the Spectral Representation Theorem to obtain spectral representations for the semigroup and value functions of derivative assets. In this Chapter we survey the spectral method in general, as well as those classes of symmetric Markov processes for which the spectral representation can be obtained in closed form, thus generating closed form solutions to Markovian derivative pricing problems.

* This research was supported by the US National Science Foundation under grants DMI-0200429 and DMI-0422937.

1 Introduction

In this Chapter we study the problem of valuing a (possibly defaultable) derivative asset contingent on the underlying state of the economy modeled as a Markov process. While general development of financial asset pricing theory does not require the Markovian assumption and is cast in the general setting of the semimartingale theory as surveyed in Chapter 1 by Jarrow and Protter, to gain analytical and computational tractability both in order to estimate the model from empirical data and in order to compute the prices of derivative assets, specific financial models in applications are often Markovian.

1.1 The Markovian derivatives pricing problem

In the Markovian setting, we assume that, in the risk-neutral economy, the underlying state variable describing economic uncertainty follows a continuous-time, time-homogeneous Markov process $\{X_t, t \geqslant 0\}$ taking values in some state space D. We assume that either $D = \mathbb{R}^d$ or $D \subset \mathbb{R}^d$, an open domain in $\mathbb{R}^d$. In the latter case, we assume that the process starts from some fixed state $x \in D$ and is instantaneously killed at the first exit time from D, $\tau_D = \inf\{t \geqslant 0\colon X_t \notin D\}$, and sent to an isolated state denoted by Δ (*cemetery state* in the terminology of Markov process theory or *default* or *bankruptcy state* in financial terminology), where it remains forever. We adjoin the cemetery state as an isolated point to the state space, so that the extended state space is $D_\Delta = D \cup \{\Delta\}$. If the process never exits D, then, by convention, $\tau_D = \infty$. We assume that X is a strong Markov process and has right-continuous with left limits sample paths in D_Δ.

A possibly defaultable *derivative asset* is described by its promised payment $f(X_T)$ at maturity date $T > 0$ that depends on the state of the underlying process X_T at T. Generally, the default event can occur in one of two ways. Either the underlying process X exits D and is killed at the first exit time τ_D, at which time the derivative asset defaults, or the default occurs at a random time τ_h with the hazard rate (intensity) $h_t = h(X_t)$ with $h(x) \geqslant 0$:

$$\tau_h = \inf\left\{ t \geqslant 0\colon \int_0^t h(X_u)\,\mathrm{d}u \geqslant e \right\},$$

where $e \sim \mathrm{Exp}(1)$ is an exponential random variable with unit mean and independent of X. Thus, the default time is ($x \wedge y := \min\{x, y\}$):

$$\zeta = \tau_h \wedge \tau_D.$$

When default occurs, the holder of the asset looses the promised payoff $f(X_T)$ and possibly receives some *recovery payment* instead (see the recent monographs Bielecki and Rutkowski, 2002; Duffie and Singleton, 2003; Lando, 2004 for surveys of credit risk modeling). If $h \equiv 0$ and the process never exits D, then the derivative asset is default-free.

Assuming that the instantaneous default-free interest rate (the *short rate*) follows the process $\{r_t = r(X_t), t \geqslant 0\}$ with $r(x) \geqslant 0$, the present value of the derivative asset at time zero is (for simplicity here we assume no recovery in default):

$$V(T,x) = \mathbb{E}\Big[e^{-\int_0^T r(X_t)\,\mathrm{d}t} f(X_T)\mathbf{1}_{\{\zeta>T\}}\big|X_0 = x\Big].$$

Recognizing that $\mathbf{1}_{\{\zeta>T\}} = \mathbf{1}_{\{\tau_h>T\}}\mathbf{1}_{\{\tau_D>T\}}$ and conditioning on the path of the process X, the asset value can be re-written in the following form using the intensity h:

$$\begin{aligned} V(T,x) &= \mathbb{E}\Big[e^{-\int_0^T r(X_t)\,\mathrm{d}t} f(X_T)\mathbf{1}_{\{\tau_D>T\}} \\ &\qquad \times \mathbb{E}\big[\mathbf{1}_{\{\tau_h>T\}}\big|\{X_t, t\in[0,T]\}\big]\big|X_0 = x\Big] \\ &= \mathbb{E}_x\Big[e^{-\int_0^T (r(X_t)+h(X_t))\,\mathrm{d}t} f(X_T)\mathbf{1}_{\{\tau_D>T\}}\Big], \end{aligned} \tag{1.1}$$

where $\mathbb{E}_x$ is with respect to the law of the process X started at $X_0 = x \in D$.

Thus, the Markovian valuation problem reduces to computing expectations of the form

$$V(t,x) = \mathcal{P}_t f(x) = \mathbb{E}_x\Big[e^{-\int_0^t k(X_u)\,\mathrm{d}u}\mathbf{1}_{\{\tau_D>t\}} f(X_t)\Big], \tag{1.2}$$

where the discount rate k is equal to the default-free short rate r plus the *instantaneous credit spread* equal to the default intensity h, $k(x) = r(x) + h(x)$.

The pricing operators $\{\mathcal{P}_t, t \geqslant 0\}$ indexed by time and considered as linear operators in the Banach space $B_b(D)$ of Borel measurable bounded payoff functions equipped with the uniform norm form an *operator semigroup* called the *pricing semigroup*. Introducing the *pricing kernel* $P_t(x, \mathrm{d}y)$ of the pricing semigroup, Eq. (1.2) can be rewritten as

$$V(t,x) = \mathcal{P}_t f(x) = \int_D f(y) P_t(x, \mathrm{d}y). \tag{1.3}$$

If the pricing kernel has a density with respect to the Lebesgue measure on D (the *state-price density*), then

$$V(t,x) = \mathcal{P}_t f(x) = \int_D f(y)\, p(t;x,y)\,\mathrm{d}y. \tag{1.4}$$

The concept of the *pricing semigroup* in financial economics goes back to Garman (1985). Applications of semigroups in financial economics have recently been studied by Ait-Sahalia et al. (2004) and Hansen and Scheinkman (2002) and by Linetsky (2004a). The primary focus of the recent survey by Ait-Sahalia et al. (2004) is on statistical estimation of Markovian models. Our primary focus in the present Chapter is on the valuation of derivative assets. In probability theory and in mathematical physics the semigroup with the dis-

count factor $e^{-\int_0^t k(X_u)\,\mathrm{d}u}$ is called the *Feynman–Kac semigroup* in honor of Richard Feynman and Mark Kac (Feynman, 1948; Kac, 1951, 1959; see also Chung and Zhao, 1995). Mathematical references on semigroups and Markov processes include Applebaum (2004) and Ethier and Kurtz (1986).

The pricing semigroup can also be interpreted as the transition semigroup of a Markov process $\{\hat{X}_t, t \geqslant 0\}$ defined as the process X killed at the random time $\hat{\zeta} = \tau_k \wedge \tau_D$, where τ_D is the first exit time of X from D and τ_k is

$$\tau_k = \inf\left\{ t \geqslant 0 \colon \int_0^t k(X_u)\,\mathrm{d}u \geqslant e \right\},$$
$$k(x) = r(x) + h(x),\ e \sim \mathrm{Exp}(1).$$

That is, we kill the process at the rate k equal to the short rate r plus default intensity h (we re-interpret the discount rate as the killing rate in this formulation). The resulting process $\hat{X}$ is sent to the cemetery state at its *lifetime* $\hat{\zeta}$:

$$\hat{X}_t = \begin{cases} X_t, & t < \hat{\zeta}, \\ \Delta, & t \geqslant \hat{\zeta}. \end{cases}$$

The pricing semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ is thus interpreted as a transition semigroup of the process $\hat{X}$ with lifetime $\hat{\zeta}$, so that the value function of the derivative asset is interpreted as:

$$V(t,x) = \mathcal{P}_t f(x) = \mathbb{E}_x\big[f\big(\hat{X}_t\big)\mathbf{1}_{\{t>\hat{\zeta}\}}\big], \tag{1.5}$$

and the pricing kernel and the state-price density (assuming it exists) are identified with the transition kernel and the transition density of the Markov process $\hat{X}$. The semigroup is *conservative*, i.e., $P_t(x, D) = 1$ for each $t \geqslant 0$ and $x \in D$, if and only if $r \equiv 0$, $h \equiv 0$, and the process never exits D, $\tau_D = \infty$. Otherwise, it is *non-conservative*.[1] We will take advantage of this interpretation of the pricing (Feynman–Kac) semigroup as the transition semigroup of the Markov process $\hat{X}$ with lifetime $\hat{\zeta}$ and will study the transition semigroup of $\hat{X}$. To distinguish the underlying process X and the process with killing at the rate $k(x) = r(x) + h(x)$, the latter will be denoted by $\hat{X}$. Having the equality of the pricing semigroup (1.2) with discounting and the transition semigroup (1.5) of the killed process in mind, we will often use the terminologies of discounting and killing interchangeably.

The Markovian valuation problem outlined above is quite general. A wide range of financial models (various one- and multi-dimensional diffusion models, such as geometric Brownian motion, CIR, CEV, stochastic volatility mod-

[1] The non-conservative case can be made into a conservative one as follows. Define an extended transition kernel on the extended state space $D_\Delta = D \cup \{\Delta\}$ as follows: $P_t(x, \{\Delta\}) = 1 - P_t(x, D)$, $P_t(\Delta, \{\Delta\}) = 1$, $P_t(\Delta, D) = 0$, so that $P_t(x, D_\Delta) = 1$ for any $x \in D_\Delta$.

els, Gaussian models, affine models, Lévy process-based pure jump and jump-diffusion models, general state-inhomogeneous pure jump and jump-diffusion models, etc.) across different markets (equity, credit, foreign exchange, commodity, energy, interest rate) and types of financial contracts (vanilla and exotic options, bonds, credit derivatives, mortgages) fit into this framework. However, in this general setting with the Banach space transition semigroup, the expectations (1.1)–(1.5) are generally intractable analytically (the pricing kernel and the state-price density are not known analytically) and can only be computed numerically either by Monte Carlo simulation (see Glasserman, 2003 for a survey) or numerical partial integro-differential equations (PIDE) methods by observing that the value function $V(t, x) = \mathcal{P}_t f(x)$ solves the initial value problem for the evolution equation:

$$V_t = \mathcal{G}V \quad \text{for } t \geqslant 0 \quad \text{and} \quad V(0, x) = f(x), \tag{1.6}$$

where $\mathcal{G}$ is the infinitesimal generator of the Markov process with killing $\hat{X}$ and its transition semigroup, generally an integro-differential operator (see Chapter 7 by Feng et al. in this volume for a brief survey).

However, in financial practice a premium is placed on models that are analytically tractable with analytical solutions for value functions of derivative assets that are fast and accurate to compute and, just as importantly, that allow easy computation of the hedge ratios (the *Greeks*) by direct differentiation with respect to parameters of interest. In this regard, it is important to note that in practice one is often interested in evaluating large portfolios. When evaluating large portfolios, availability of analytical solutions for individual securities becomes particularly important, as one needs to evaluate thousands of security prices and hedge ratios. In fact, arguably one of the reasons why the Black–Scholes–Merton option pricing model has been so widely successful and so rapidly adopted in the market place lies in the fact that the model is analytically tractable. Ever since Black, Scholes, and Merton, the hunt has been on for more general and more empirically realistic analytically tractable models that lead to closed form solutions for value functions of derivative assets.

1.2 Symmetric semigroups and symmetric Markov processes

An important analytical simplification is possible if the Hilbert space structure is available and the Markov process is such that its transition semigroup is symmetric in the Hilbert space metric. In particular, suppose there is a (finite or infinite) measure m on D with full support, such that the pricing operators $\mathcal{P}_t$ are symmetric operators in the real Hilbert space $\mathcal{H} := L^2(D, m)$ with the inner product

$$(f, g) = \int_D f(x)g(x)m(\mathrm{d}x), \tag{1.7}$$

i.e.,

$$(\mathcal{P}_t f, g) = (f, \mathcal{P}_t g) \quad \text{for any } t \geqslant 0 \text{ and } f, g \in \mathcal{H}. \tag{1.8}$$

Such a semigroup is said to be *m-symmetric* (or simply symmetric). Assuming the transition density exists and the symmetry measure has a density $m(x)$ with respect to the Lebesgue measure, the transition density of an m-symmetric Markov process is m-symmetric:

$$p(t; x, y)m(x) = p(t; y, x)m(y) =: p_m(t; x, y)m(x)m(y), \tag{1.9}$$

where

$$p_m(t; x, y) = p_m(t; y, x) = p(t; x, y)/m(y) \tag{1.10}$$

is the symmetric transition density with respect to the measure m (symmetry between x and y).

If the semigroup is symmetric, then its infinitesimal generator $\mathcal{G}$ is a (generally unbounded) non-positive self-adjoint operator in $L^2(D, m)$. In this special case one can invoke the Spectral Representation Theorem for self-adjoint operators in Hilbert space to obtain a unique *spectral representation* for the semigroup:

$$\mathcal{P}_t f = \int_{[0,\infty)} e^{-\lambda t} E(\mathrm{d}\lambda) f, \quad \forall f \in L^2(D, m), \tag{1.11}$$

where $E(\mathrm{d}\lambda)$ is the so-called *projection-valued spectral measure* corresponding to the negative of the infinitesimal generator $-\mathcal{G}$. When the spectrum of $-\mathcal{G}$ is purely discrete, the spectral representation simplifies to the *eigenfunction expansion*:

$$\mathcal{P}_t f = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} \varphi_n, \quad c_n = (f, \varphi_n), \ \forall f \in L^2(D, m), \tag{1.12}$$

where φ_n are a complete orthonormal system of eigenvectors of $-\mathcal{G}$ with eigenvalues λ_n, i.e., $-\mathcal{G}\varphi_n = \lambda_n \varphi_n$. The φ_n are also eigenvectors of $\mathcal{P}_t$ with eigenvalues $e^{-\lambda_n t}$, i.e., $\mathcal{P}_t \varphi_n = e^{-\lambda_n t}\varphi_n$. They form a complete orthonormal basis in $L^2(D, m)$, and c_n are the expansion coefficients of the payoff f in this basis. If for a particular model this spectral representation can be worked out in closed form, it provides explicit closed-form representation for the pricing operator, pricing kernel, state-price density, and value functions of derivative assets.

A Markov process whose transition semigroup is m-symmetric is said to be m-symmetric or simply symmetric. Fukushima et al. (1994) is the standard reference text on the general theory of symmetric Markov processes. When the process is a one-dimensional diffusion on some interval $I \subseteq \mathbb{R}$ (finite or infinite) with the (finite or infinite) speed measure m (see Eq. (3.3) for definition) and appropriate boundary conditions at the endpoints of the interval,

the pricing semigroup is symmetric in $L^2(I, m)$ without any further conditions, and the Spectral Representation Theorem yields a spectral representation for the state-price density and value functions of derivative assets. Applications of the spectral theory to one-dimensional diffusions go back to McKean (1956). Based on the work of Feller, McKean constructed a spectral representation for a general one-dimensional diffusion (see also Ito and McKean, 1974, Section 4.11). Wong (1964) investigated a class of diffusion processes possessing stationary densities in the Pearson class for which the spectral representation of the transition semigroup can be expressed in closed form in terms of certain special functions. Many of the diffusion processes important in finance (such as Ornstein–Uhlenbeck, square-root CIR, Jacobi, etc.) fall within this framework. In contrast to one-dimensional diffusions, for multi-dimensional diffusions the m-symmetry condition restricts the form of the drift vector. For jump processes, the m-symmetry restricts the form of the Lévy measure.

1.3 Chapter outline

The aim of this Chapter is to describe the spectral method in general, as well as to give a survey of specific classes of analytically tractable models for which the spectral representation can be obtained in closed form, leading to analytical solutions for the state-price density and value functions of derivative assets. Our plan is as follows. In Section 2 we survey some prerequisite notions and results from the spectral theory of self-adjoint operators in Hilbert spaces (Section 2.1) and symmetric semigroups of operators in Hilbert spaces (Section 2.2). The key results here are the Spectral Representation Theorem for self-adjoint operators in Hilbert space (Theorem 2.1) and the corresponding result for symmetric semigroups (Theorem 2.2).

Since essentially all one-dimensional diffusions are symmetric, in Sections 3 and 4 we undertake a survey of the spectral method for one-dimensional diffusions, both discussing the general spectral representation for a one-dimensional diffusion (Section 3), and surveying specific families of analytically tractable one-dimensional diffusions and associated analytically tractable financial models, including options pricing models, interest rate models, and credit risk models (Section 4). This survey of one-dimensional diffusion models is primarily based on the series of papers Davydov and Linetsky (2001, 2003), Gorovoi and Linetsky (2004, 2006), and Linetsky (2004a, 2004b, 2004c, 2004d, 2004e, 2005, 2006), where the spectral method has been profitably employed to find analytical solutions for a wide range of financial models, including models for vanilla and exotic options (such as Asian, barrier, and lookback options) under various assumptions for the underlying asset price process, as well as in models for interest rates, credit risk, and mortgage prepayments. Successful applications of the spectral method in these models convinced the author that the spectral method is a powerful tool to generate analytical solutions for Markovian asset pricing problems. The spectral method has also been applied

to derivatives pricing by Albanese and co-authors (Albanese et al., 2001; Albanese and Kuznetsov, 2004, 2005; Albanese and Lawi, 2005), Boyarchenko and Levendorskiy (2006), Lewis (1998, 2000) and Lipton (2001, 2002), Lipton and McGhee (2002), Larsen and Sorensen (2007).

In Section 5 we discuss symmetric multi-dimensional diffusions. In Section 6 we described a procedure to generate analytically tractable jump-diffusion and pure jump processes starting from an analytically tractable diffusion process via *Bochner's subordination* (stochastic time change, where the time change process is a Lévy subordinator), as well as processes with stochastic volatility via absolutely continuous time changes with the time change taken to be an integral of an independent Markov process. Section 7 concludes this survey, discusses some advantages of the spectral method in financial applications, and outlines some further research directions.

The spectral expansion method has also found interesting applications in econometrics for estimation of diffusion processes (Bibby et al., 2004; Hansen et al., 1998; Florens et al., 1998; Larsen and Sorensen, 2007). In this chapter we only survey derivatives pricing applications, and refer the reader to the recent survey Ait-Sahalia et al. (2004) for econometrics applications.

2 Self-adjoint semigroups in Hilbert spaces

2.1 Spectral theory of self-adjoint operators in Hilbert spaces

In this section we review some basic notions and results from the spectral theory of self-adjoint operators in Hilbert spaces. We primarily follow Reed and Simon (1980), Demuth and van Casteren (2000), and Dunford and Schwartz (1963).

Let $\mathcal{H}$ be a separable real Hilbert space. A *linear operator* $(\mathrm{Dom}(\mathcal{A}), \mathcal{A})$ is a pair, where $\mathrm{Dom}(\mathcal{A})$ is a linear subspace of $\mathcal{H}$ (called the *domain* of $\mathcal{A}$) and $\mathcal{A}$ is a linear map from $\mathrm{Dom}(\mathcal{A})$ into $\mathcal{H}$. The *range* of a linear operator $\mathcal{A}$ is the image of its domain under the map $\mathcal{A}$. Two operators $\mathcal{A}_1$ and $\mathcal{A}_2$ are equal if $\mathrm{Dom}(\mathcal{A}_1) = \mathrm{Dom}(\mathcal{A}_2)$ and $\mathcal{A}_1 f = \mathcal{A}_2 f$ for all $f \in \mathrm{Dom}(\mathcal{A}_1)$. A linear operator $\mathcal{A}_2$ is an *extension* of $\mathcal{A}_1$ if $\mathrm{Dom}(\mathcal{A}_1) \subseteq \mathrm{Dom}(\mathcal{A}_2)$ and $\mathcal{A}_1 f = \mathcal{A}_2 f$ for all $f \in \mathrm{Dom}(\mathcal{A}_1)$. In this situation $\mathcal{A}_1$ is called a *restriction* of $\mathcal{A}_2$. The sum of two operators $\mathcal{A}_1 + \mathcal{A}_2$ has the domain $\mathrm{Dom}(\mathcal{A}_1 + \mathcal{A}_2) = \mathrm{Dom}(\mathcal{A}_1) \cap \mathrm{Dom}(\mathcal{A}_2)$ and $(\mathcal{A}_1 + \mathcal{A}_2) f := \mathcal{A}_1 f + \mathcal{A}_2 f$ for $f \in \mathrm{Dom}(\mathcal{A}_1 + \mathcal{A}_2)$.

A linear operator $\mathcal{A}$ is said to be *bounded* if there is a constant $M > 0$ such that

$$\|\mathcal{A} f\| \leqslant M \|f\| \quad \text{for all } f \in \mathrm{Dom}(\mathcal{A})$$

(here $\|f\| \equiv (f, f)$ is the norm in $\mathcal{H}$). Any bounded operator $\mathcal{A}$ has a unique extension $\overline{\mathcal{A}}$ with $\mathrm{Dom}(\overline{\mathcal{A}}) = \overline{\mathrm{Dom}(\mathcal{A})}$, the closure of $\mathrm{Dom}(\mathcal{A})$. In particular, if $\mathcal{A}$ is *densely defined* (i.e., $\mathrm{Dom}(\mathcal{A})$ is a dense subset of $\mathcal{H}$), then the domain of $\overline{\mathcal{A}}$ coincides with the Hilbert space $\mathcal{H}$. We will not distinguish between a

bounded operator $\mathcal{A}$ and its bounded extension $\overline{\mathcal{A}}$. If we consider bounded operators we always mean operators with $\mathrm{Dom}(\mathcal{A}) = \mathcal{H}$.

For a densely defined operator $\mathcal{A}$, its *adjoint* $\mathcal{A}^*$ is defined as follows: $g \in \mathrm{Dom}(\mathcal{A}^*)$ if there exists $h \in \mathcal{H}$ such that $(\mathcal{A}f, g) = (f, h)$, for all $f \in \mathrm{Dom}(\mathcal{A})$. Then $\mathcal{A}^*$ is a linear mapping from $\mathrm{Dom}(\mathcal{A}^*)$ to $\mathcal{H}$ given by $\mathcal{A}^* g = h$, $g \in \mathrm{Dom}(\mathcal{A}^*)$. This means

$$(\mathcal{A}f, g) = (f, \mathcal{A}^* g), \quad \forall f \in \mathrm{Dom}(\mathcal{A}),\ g \in \mathrm{Dom}(\mathcal{A}^*).$$

A densely defined linear operator is called *symmetric* if

$$(\mathcal{A}f, g) = (f, \mathcal{A}g), \quad \forall f, g \in \mathrm{Dom}(\mathcal{A}) \subseteq \mathrm{Dom}(\mathcal{A}^*).$$

If $\mathcal{A}$ is symmetric and furthermore $\mathrm{Dom}(\mathcal{A}) = \mathrm{Dom}(\mathcal{A}^*)$, then the operator is called *self-adjoint*. If $\mathrm{Dom}(\mathcal{A}) = \mathcal{H}$ and $\mathcal{A}$ is symmetric, then $\mathcal{A}$ is bounded and self-adjoint.

To formulate the spectral theorem for self-adjoint operators we need to define *projection-valued measures* or *spectral measures*.

Definition 2.1 *(Spectral Measure)*. Let $\mathcal{H}$ be a separable real Hilbert space and let $\mathcal{B}(\mathbb{R})$ be a Borel σ-algebra on $\mathbb{R}$. A family of bounded linear operators $\{E(B), B \in \mathcal{B}(\mathbb{R})\}$ in $\mathcal{H}$ such that:

- (i) Each $E(B)$ is an orthogonal projection (i.e., $E^2(B) = E(B)$ and $E^*(B) = E(B)$);
- (ii) $E(\emptyset) = 0$, $E(\mathbb{R}) = I$ (I is the identity operator in $\mathcal{H}$);
- (iii) If $B = \bigcup_{n=1}^{\infty} B_n$ with $B_n \cap B_m = \emptyset$ if $n \neq m$, then $E(B) = \sum_{n=1}^{\infty} E(B_n)$ (where the limit involved in the infinite series is taken in the strong operator topology);
- (iv) $E(B_1)E(B_2) = E(B_1 \cap B_2)$;

is called a *projection-valued measure*, *spectral measure*, or the *resolution of the identity*.

As in the case of scalar measures, the *support of a spectral measure* ($\mathrm{Supp}(E)$) can be defined as the smallest closed subset in $\mathbb{R}$ such that $E(\mathrm{Supp}(E)) = I$. For $f \in \mathcal{H}$, $\mu_f(B) := (f, E(B)f)$ is a well-defined Borel measure on $\mathbb{R}$ normalized so that $\mu_f(\mathbb{R}) = \|f\|^2$.

Theorem 2.1 *(Spectral Representation Theorem for Self-Adjoint Operators). There is a one-to-one correspondence between self-adjoint operators $\mathcal{A}$ and projection-valued measures $\{E(B), B \in \mathcal{B}(\mathbb{R})\}$ in $\mathcal{H}$, the correspondence being given by*:

$$\mathrm{Dom}(\mathcal{A}) = \left\{ f \in \mathcal{H}: \int_{\mathbb{R}} \lambda^2 \mu_f(\mathrm{d}\lambda) < \infty \right\}$$

and

$$\mathcal{A}f = \int_{\mathbb{R}} \lambda E(\mathrm{d}\lambda) f, \quad f \in \mathrm{Dom}(\mathcal{A}). \tag{2.1}$$

The integral in the spectral representation (2.1) of $\mathcal{A}$ can be understood in the weak sense, that is,

$$(f, \mathcal{A}g) = \int_{\mathbb{R}} \lambda\big(f, E(\mathrm{d}\lambda)g\big), \quad f \in \mathcal{H},\ g \in \mathrm{Dom}(\mathcal{A}).$$

In fact, this integral also converges in the strong sense in $\mathcal{H}$. The spectral representation of a self-adjoint operator $\mathcal{A}$ is abbreviated as

$$\mathcal{A} = \int_{\mathbb{R}} \lambda E(\mathrm{d}\lambda).$$

The *spectrum* of $\mathcal{A}$ coincides with the support of its spectral measure E.

The spectral representation theorem gives rise to the following *functional calculus* for self-adjoint operators. Let E be the spectral measure corresponding to the self-adjoint operator $\mathcal{A}$ and let ϕ be a real-valued Borel measurable function on $\mathbb{R}$. Then one can define a new operator (a function $\phi(\mathcal{A})$ of the operator $\mathcal{A}$)

$$\phi(\mathcal{A}) := \int_{\mathbb{R}} g(\lambda) E(\mathrm{d}\lambda), \tag{2.2}$$

which is a self-adjoint operator in $\mathcal{H}$ with domain

$$\mathrm{Dom}\big(\phi(\mathcal{A})\big) = \left\{ f \in \mathcal{H}\colon \int_{\mathbb{R}} \phi^2(\lambda)\mu_f(\mathrm{d}\lambda) < \infty \right\}.$$

It is bounded if and only if ϕ is bounded.

The *resolvent set* $\rho(\mathcal{A})$ of a linear operator $\mathcal{A}$ consists of all $\alpha \in \mathbb{C}$ such that there exists a bounded operator $\mathcal{R}_\alpha := (\alpha I - \mathcal{A})^{-1}$ called the *resolvent*. From the Spectral Representation Theorem, we have the spectral representation for the resolvent of a self-adjoint operator:

$$\mathcal{R}_\alpha f = (\alpha I - \mathcal{A})^{-1} f = \int_{\mathbb{R}} (\alpha - \lambda)^{-1} E(\mathrm{d}\lambda) f, \quad \alpha \in \rho(\mathcal{A}),\ f \in \mathcal{H}.$$

The complement $\sigma(\mathcal{A}) := \mathbb{C}\backslash\rho(\mathcal{A})$ of the resolvent set is called the *spectrum* of $\mathcal{A}$. The support of the spectral measure E of $\mathcal{A}$ coincides with the spectrum $\sigma(\mathcal{A})$. The resolvent set $\rho(\mathcal{A})$ is open and the spectrum $\sigma(\mathcal{A})$ is closed. If $\mathcal{A}$ is self-adjoint, then the spectrum of $\mathcal{A}$ is non-empty and lies on the real axis.

We say that a self-adjoint operator $\mathcal{A}$ is *non-negative* (*non-positive*) if

$$(f, \mathcal{A}f) \geqslant 0 \quad \big((f, \mathcal{A}f) \leqslant 0\big), \ \forall f \in \mathrm{Dom}(\mathcal{A}).$$

If $\mathcal{A}$ is non-positive (non-negative), then its spectrum $\sigma(\mathcal{A})$ lies on the non-positive (non-negative) half-axis. If $\mathcal{A}$ is a non-negative self-adjoint operator, its spectral representation takes the form

$$\mathcal{A} = \int\limits_{[0,\infty)} \lambda E(\mathrm{d}\lambda).$$

If $\mathcal{A}$ is a bounded operator, then its spectrum is non-empty and compact.

The *pure point spectrum* of $\mathcal{A}$ is defined as the set of all *eigenvalues* of $\mathcal{A}$:

$$\sigma_{pp}(\mathcal{A}) := \big\{\lambda \in \mathbb{R}\colon \exists f \in \mathrm{Dom}(\mathcal{A}) \text{ such that } \mathcal{A}f = \lambda f\big\}.$$

One can distinguish the following subspaces in $\mathcal{H}$: $\mathcal{H}_{pp}$ – the closure of the linear hull of all *eigenspaces* of $\mathcal{A}$, $\mathcal{H}_{ac}$ – the set of all $f \in \mathcal{H}$ such that the measure $\mu_f(B) = (f, E(B)f)$ is absolutely continuous with respect to the Lebesgue measure, and $\mathcal{H}_{cs}$ — the orthogonal complement of $\mathcal{H}_{pp} \oplus \mathcal{H}_{ac}$ in $\mathcal{H}$. If $f \in \mathcal{H}_{cs}$, then $\mu_f(B) := (f, E(B)f)$ is a continuous measure singular with respect to the Lebesgue measure. The subspaces $\mathcal{H}_{pp}$, $\mathcal{H}_{ac}$, and $\mathcal{H}_{cs}$ are orthogonal to each other and invariant with respect to $\mathcal{A}$. The restrictions $\mathcal{A}_{pp}$, $\mathcal{A}_{ac}$, and $\mathcal{A}_{cs}$ of $\mathcal{A}$ to $\mathrm{Dom}(\mathcal{A}) \cap \mathcal{H}_{pp}$, $\mathrm{Dom}(\mathcal{A}) \cap \mathcal{H}_{ac}$, and $\mathrm{Dom}(\mathcal{A}) \cap \mathcal{H}_{cs}$ are self-adjoint as operators in $\mathcal{H}_{pp}$, $\mathcal{H}_{ac}$, and $\mathcal{H}_{cs}$, respectively. They are called the *pure point*, *absolutely continuous*, and *continuous singular* components of $\mathcal{A}$. One can also define the *absolutely continuous*, *continuous singular*, and *pure point* components of the spectrum of $\mathcal{A}$: $\sigma_{ac}(\mathcal{A}) := \sigma(\mathcal{A}_{ac})$, $\sigma_{cs}(\mathcal{A}) := \sigma(\mathcal{A}_{cs})$ and $\sigma_{pp}(\mathcal{A}) := \sigma(\mathcal{A}_{pp})$. The pure point spectrum so defined is the *closure* of the pure point spectrum defined previously as the actual set of all eigenvalues. Both definitions are used in the literature. With the definition of the pure point spectrum as the closure of the set of all eigenvalues, one has a decomposition:

$$\sigma(\mathcal{A}) = \sigma_{ac}(\mathcal{A}) \cup \sigma_{cs}(\mathcal{A}) \cup \sigma_{pp}(\mathcal{A})$$

(however, the three components do not have to be disjoint). The *continuous spectrum* is

$$\sigma_c(\mathcal{A}) = \sigma_{ac}(\mathcal{A}) \cup \sigma_{cs}(\mathcal{A}).$$

We say that $\lambda \in \sigma(\mathcal{A})$ is in the *essential spectrum* of $\mathcal{A}$, $\sigma_e(\mathcal{A})$, if and only if the range of $E((\lambda - \epsilon, \lambda + \epsilon))$ is infinite-dimensional for all $\epsilon > 0$. If $\lambda \in \sigma(\mathcal{A})$, but the range of $E((\lambda - \epsilon, \lambda + \epsilon))$ is finite-dimensional for some $\epsilon > 0$, we say that $\lambda \in \sigma_d(\mathcal{A})$, the *discrete spectrum* of $\mathcal{A}$. This gives a decomposition of the spectrum into two *disjoint* components,

$$\sigma(\mathcal{A}) = \sigma_d(\mathcal{A}) \cup \sigma_e(\mathcal{A}).$$

The set $\sigma_d(\mathcal{A})$ is not necessarily closed, but $\sigma_e(\mathcal{A})$ is always closed. $\lambda \in \sigma_d(\mathcal{A})$ if and only if λ is an isolated point of $\sigma(\mathcal{A})$ *and* λ is an eigenvalue of finite multiplicity. Obviously, $\sigma_d(\mathcal{A}) \subset \sigma_{pp}(\mathcal{A})$. $\lambda \in \sigma_e(\mathcal{A})$ if and only if one or more of the following holds: (a) $\lambda \in \sigma_c(\mathcal{A}) \equiv \sigma_{ac}(\mathcal{A}) \cap \sigma_{cs}(\mathcal{A})$; (b) λ is a limit point of the pure point spectrum $\sigma_{pp}(\mathcal{A})$; (c) λ is an eigenvalue of infinite multiplicity. Some authors (e.g., Dunford and Schwartz, 1963, p. 1393) define the essential spectrum as follows: $\sigma_e(\mathcal{A}) := \{\lambda \in \mathbb{R}\text{: Range of } (\lambda I - \mathcal{A}) \text{ is not closed}\}$. In this definition, the essential spectrum of a self-adjoint operator $\mathcal{A}$ is the set of non-isolated points of $\sigma(\mathcal{A})$ (Dunford and Schwartz, 1963, p. 1395), and $\lambda \in \sigma_e(\mathcal{A})$ if and only if one or more of the following holds: (a) $\lambda \in \sigma_c(\mathcal{A}) \equiv \sigma_{ac}(\mathcal{A}) \cap \sigma_{cs}(\mathcal{A})$; (b) λ is a limit point of the pure point spectrum $\sigma_{pp}(\mathcal{A})$. Eigenvalues of ordinary differential operators always have finite multiplicity, so for such operators the two definitions of the essential spectrum coincide.

2.2 Self-adjoint semigroups in Hilbert spaces

We start by recalling some notions from the theory of operator semigroups in Banach spaces (see Applebaum, 2004, Chapter 3; Davies, 1980; Demuth and van Casteren, 2000; Ethier and Kurtz, 1986; Fukushima et al., 1994; Hille and Phillips, 1957).

Definition 2.2. A family of bounded linear operators $\{\mathcal{P}_t, t \geqslant 0\}$ is called a *strongly continuous semigroup* in a real Banach space $(\mathbf{B}, \|\cdot\|)$ if it possesses the following properties:

- (i) (semigroup property) $\mathcal{P}_s\mathcal{P}_t = \mathcal{P}_{s+t}$ for all $s, t \geqslant 0$;
- (ii) (identity) $\mathcal{P}_0 = I$;
- (iii) (strong continuity) for every $f \in \mathbf{B}$ the mapping $t \mapsto \mathcal{P}_t f$ is continuous from $[0, \infty)$ to $(\mathbf{B}, \|\cdot\|)$ (i.e., $\lim_{s\to t} \|\mathcal{P}_t f - \mathcal{P}_s f\| = 0$ for all $t \geqslant 0$ and $f \in \mathbf{B}$).

For every strongly continuous semigroup there exist constants $M \geqslant 0$ and $b \in \mathbb{R}$ such that

$$\|\mathcal{P}_t\| \leqslant M e^{bt}, \quad t \geqslant 0$$

(where $\|\mathcal{A}\| = \sup_{f\in\mathbf{B}} \|\mathcal{A}f\|/\|f\|$ is the operator norm). The semigroup is called a *contraction semigroup* if the above inequality is valid with $M = 1$ and $b = 0$, i.e.,

$$\|\mathcal{P}_t\| \leqslant 1, \quad t \geqslant 0.$$

To every semigroup one can associate an operator which is called an *infinitesimal generator* of the semigroup:

$$\mathcal{G}f := \lim_{t\downarrow 0} \frac{1}{t}(\mathcal{P}_t f - f), \quad f \in \mathrm{Dom}(\mathcal{G}),$$

$$\mathrm{Dom}(\mathcal{G}) = \left\{ f \in \mathbf{B}\colon \lim_{t \downarrow 0} \frac{1}{t}(\mathcal{P}_t f - f) \text{ exists in } \mathbf{B} \right\}.$$

We now assume that $\mathbf{B} = \mathcal{H}$ is a separable real Hilbert space and $\mathcal{P}_t$ is a strongly continuous contraction semigroup in $\mathcal{H}$. The semigroup is called *self-adjoint* in $\mathcal{H}$ if each $\mathcal{P}_t$ is a self-adjoint operator, i.e.,

$$(\mathcal{P}_t f, g) = (f, \mathcal{P}_t g), \quad f, g \in \mathcal{H} \tag{2.3}$$

(for a bounded operator the domain coincides with the whole Hilbert space $\mathcal{H}$ and, hence, a bounded symmetric operator is self-adjoint; for such operators there is no distinction between symmetric and self-adjoint).

We have the following theorem (Davies, 1980, p. 99; Hille and Phillips, 1957, Theorem 22.3.1).

Theorem 2.2. *The operator $\mathcal{G}$ is the infinitesimal generator of a strongly continuous self-adjoint contraction semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ in $\mathcal{H}$ if and only if $\mathcal{G}$ is a non-positive self-adjoint operator in $\mathcal{H}$. If*

$$-\mathcal{G} = \int_{[0,\infty)} \lambda E(\mathrm{d}\lambda) \tag{2.4}$$

is the spectral representation of $-\mathcal{G}$, then for every $t \geqslant 0$

$$\mathcal{P}_t = e^{t\mathcal{G}} = \int_{[0,\infty)} e^{-\lambda t} E(\mathrm{d}\lambda). \tag{2.5}$$

(*Note*: *we express the spectral expansion relative to the negative of the infinitesimal generator to have a non-negative spectral parameter $\lambda \geqslant 0$.*)

The spectral representation (2.5) comes from applying the functional calculus representation (2.2) to the exponential function. We thus have a one-to-one correspondence between strongly continuous self-adjoint contraction semigroups in $\mathcal{H}$ and non-positive self-adjoint operators in $\mathcal{H}$, their generators.

The *resolvent* of a strongly continuous self-adjoint contraction semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ is defined as a family $\{\mathcal{R}_\alpha, \alpha > 0\}$ of the resolvent operators $\mathcal{R}_\alpha$ for the infinitesimal generator $\mathcal{G}$ with $\alpha > 0$ (note that since the spectrum $\sigma(\mathcal{G}) \subseteq (-\infty, 0]$, $(0, \infty) \subset \rho(\mathcal{G})$):

$$\mathcal{R}_\alpha = (\alpha I - \mathcal{G})^{-1} = \int_{[0,\infty)} (\alpha + \lambda)^{-1} E(\mathrm{d}\lambda). \tag{2.6}$$

The resolvent can also be written as the Laplace transform of the semigroup:

$$\mathcal{R}_\alpha = \int_0^\infty e^{-\alpha t} \mathcal{P}_t \, \mathrm{d}t. \tag{2.7}$$

Things simplify considerably when the infinitesimal generator has a purely discrete spectrum (i.e., the essential spectrum is empty). Let $-\mathcal{G}$ be a non-negative self-adjoint operator with purely discrete spectrum $\sigma_d(-\mathcal{G}) \subseteq [0, \infty)$. Then the spectral measure can be defined by

$$E(B) = \sum_{\lambda \in \sigma_d(-\mathcal{G}) \cap B} P(\lambda),$$

where $P(\lambda)$ is the orthogonal projection onto the *eigenspace* corresponding to the eigenvalue $\lambda \in \sigma_d(-\mathcal{G})$, i.e., the subspace $\{f \in \mathcal{H}\colon -\mathcal{G}f = \lambda f\}$, and the sum is over all eigenvalues that fall into the set B. Then the spectral representation of the semigroup generated by $\mathcal{G}$ takes the simpler form:

$$\mathcal{P}_t f = e^{t\mathcal{G}} f = \sum_{\lambda \in \sigma_d(-\mathcal{G})} e^{-\lambda t} P(\lambda) f, \quad t \geqslant 0,\ f \in \mathcal{H},$$

$$-\mathcal{G}f = \sum_{\lambda \in \sigma_d(-\mathcal{G})} \lambda P(\lambda) f, \quad f \in \mathrm{Dom}(\mathcal{G}).$$

For $t = 0$ we have the spectral expansion for any $f \in \mathcal{H}$:

$$f = \sum_{\lambda \in \sigma_d(-\mathcal{G})} P(\lambda) f, \qquad \|f\|^2 = \sum_{\lambda \in \sigma_d(-\mathcal{G})} \|P(\lambda) f\|^2.$$

If $\{\varphi_n\}_{n=1}^{\infty}$ is a complete orthonormal system of eigenvectors of $\mathcal{G}$ with eigenvalues[2] $\{-\lambda_n\}_{n=1}^{\infty}$, $\lambda_n \geqslant 0$ (each eigenvalue is counted as many times as its multiplicity, i.e., the dimension of the eigenspace; here all eigenvalues are assumed to have finite multiplicity),

$$-\mathcal{G}\varphi_n = \lambda_n \varphi_n, \tag{2.8}$$

then the eigenvector expansion is valid for any $f \in \mathcal{H}$:

$$f = \sum_{n=1}^{\infty} c_n \varphi_n, \quad c_n = (f, \varphi_n), \tag{2.9}$$

and the Parseval equality holds, $\|f\|^2 = \sum_{n=1}^{\infty} c_n^2$. For each $t > 0$, the φ_n are also eigenvectors of the operator $\mathcal{P}_t$ with eigenvalues $e^{-\lambda_n t}$,

$$\mathcal{P}_t \varphi_n = e^{-\lambda_n t} \varphi_n, \tag{2.10}$$

and the spectral representations of the semigroup and the resolvent take the form:

$$\mathcal{P}_t f = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} \varphi_n, \quad t \geqslant 0,\ f \in \mathcal{H}, \tag{2.11}$$

[2] In our notation, λ_n are eigenvalues of the negative of the generator, $-\mathcal{G}$, so that $-\lambda_n$ are eigenvalues of $\mathcal{G}$.

$$\mathcal{R}_\alpha f = \sum_{n=1}^{\infty} \frac{c_n \varphi_n}{\alpha + \lambda_n}, \quad \alpha > 0,\ f \in \mathcal{H}. \tag{2.12}$$

So far we have worked with an abstract Hilbert space $\mathcal{H}$. We now come back to the framework of Section 1.2 with the state space D, $D = \mathbb{R}^d$ or $D \subset \mathbb{R}^d$, an open domain in $\mathbb{R}^d$, and m a (finite or infinite) measure on D and specialize to $\mathcal{H} = L^2(D, m)$, the Hilbert space of square-integrable functions on D endowed with the inner product (1.7). We further assume that the underlying Markov process is symmetric, i.e., its transition semigroup is symmetric in the sense of Eq. (1.8). All of the results in this section are thus applicable and yield a spectral expansion of the transition semigroup of the symmetric Markov process. Our interest is in the specific processes for which the spectral expansion can be obtained in closed form, as it yields closed-form expressions for value functions of derivative assets.

3 One-dimensional diffusions: general results

3.1 Preliminaries

The symmetry (1.8) is always satisfied for one-dimensional diffusions with the speed measure taken to be the symmetry measure. Indeed, assume that $\{\hat{X}_t, t \geqslant 0\}$ is a one-dimensional, time-homogeneous regular (i.e., it reaches every point in (e_1, e_2) with positive probability) diffusion whose state space is some interval $I \subseteq \mathbb{R}$ with endpoints e_1 and e_2, $-\infty \leqslant e_1 < e_2 \leqslant \infty$, and with the infinitesimal generator

$$\mathcal{G}f(x) = \frac{1}{2}a^2(x)f''(x) + b(x)f'(x) - k(x)f(x), \quad x \in (e_1, e_2). \tag{3.1}$$

We assume that the diffusion (volatility) coefficient $a(x)$ is twice continuously differentiable and strictly positive on the open interval (e_1, e_2), the drift $b(x)$ is continuously differentiable on (e_1, e_2), and the killing rate $k(x)$ is continuous and non-negative on (e_1, e_2) (these regularity assumptions are not necessary, but will simplify further development; in what follows we always assume that these assumptions are in force). The infinitesimal generator of $\hat{X}$ can be rewritten in the formally self-adjoint form[3]:

$$\mathcal{G}f(x) = \frac{1}{m(x)}\left(\frac{f'(x)}{s(x)}\right)' - k(x)f(x), \quad x \in (e_1, e_2), \tag{3.2}$$

[3] We say *formally self-adjoint* because we have not said anything yet about the domain of $\mathcal{G}$. In order to specify $\mathcal{G}$ as a self-adjoint operator in the Hilbert space $L^2(I, m)$ with the speed measure m, which, in our case, has a density $m(x)$ given by Eq. (3.3), we need to describe its domain $\mathrm{Dom}(\mathcal{G}) \subset L^2(I, m)$, which involves careful consideration of boundary conditions at the endpoints e_1 and e_2, which we will do shortly.

where $s(x)$ and $m(x)$ are the *scale* and *speed densities* (see Borodin and Salminen, 2002, Chapter II for details on one-dimensional diffusions; we include 2 in the definition of the speed density to conform to the usual convention):

$$s(x) := \exp\left(-\int_{x_0}^{x} \frac{2b(y)}{a^2(y)}\,\mathrm{d}y\right), \qquad m(x) := \frac{2}{a^2(x)s(x)}, \tag{3.3}$$

where $x_0 \in (e_1, e_2)$ is an arbitrary point in the state space. Under our assumptions, both $s(x)$ and $m(x)$ are twice continuously differentiable and strictly positive on (e_1, e_2).

The endpoints e_i, $i = 1, 2$, are either natural, entrance, exit, or regular boundaries for the diffusion $\hat{X}$. Feller's boundary classification for diffusions with killing is made as follows (e.g., Borodin and Salminen, 2002, Chapter II). For any $x, y \in (e_1, e_2)$ define the *scale function* (here $x_0 \in (e_1, e_2)$ is an arbitrary point in the state space):

$$\mathcal{S}(x) := \int_{x_0}^{x} s(z)\,\mathrm{d}z, \qquad \mathcal{S}[x, y] := \mathcal{S}(y) - \mathcal{S}(x) = \int_{x}^{y} s(z)\,\mathrm{d}z,$$

and the limits

$$\mathcal{S}(e_1, y] := \lim_{x \downarrow e_1} \mathcal{S}[x, y], \qquad \mathcal{S}[x, e_2) := \lim_{y \uparrow e_2} \mathcal{S}[x, y]$$

(the limits may be infinite). Furthermore, fix some $\epsilon \in (e_1, e_2)$ and define (we did some obvious interchanges of integrations in formulas on pp. 14–15 of Borodin and Salminen (2002) to present these formulas in this convenient for us form)

$$I_1 := \int_{e_1}^{\epsilon} \mathcal{S}(e_1, x]\big(1 + k(x)\big)m(x)\,\mathrm{d}x,$$

$$I_2 := \int_{\epsilon}^{e_2} \mathcal{S}[x, e_2)\big(1 + k(x)\big)m(x)\,\mathrm{d}x,$$

$$J_1 := \int_{e_1}^{\epsilon} \mathcal{S}[x, \epsilon]\big(1 + k(x)\big)m(x)\,\mathrm{d}x,$$

$$J_2 := \int_{\epsilon}^{e_2} \mathcal{S}[\epsilon, x]\big(1 + k(x)\big)m(x)\,\mathrm{d}x.$$

The boundary $e_i \in \{e_1, e_2\}$ is said to be

- *regular* if $I_i < \infty$ and $J_i < \infty$,
- *exit* if $I_i < \infty$ and $J_i = \infty$,
- *entrance* if $I_i = \infty$ and $J_i < \infty$,
- or *natural* if $I_i = \infty$ and $J_i = \infty$.

The diffusion process is instantaneously killed at the first hitting time of an exit boundary (and is sent to the cemetery state Δ) and cannot be started from an exit boundary. The process never reaches an entrance boundary if started in the interior of the state space, but can be started from an entrance boundary, in which case it immediately enters the state space and never returns to the entrance boundary. The process never reaches a natural boundary if started in the interior of the state space, and cannot be started from a natural boundary. Exit, entrance, and natural boundaries are not included in the state space I (the interval I is open at an endpoint which is classified as exit, entrance, or natural). At the regular boundary we may impose a boundary condition. Here we consider either killing or instantaneously reflecting boundary conditions at regular boundaries. In the former case the process is sent to the cemetery state Δ at the first hitting time of the regular boundary, and the boundary point is not included in the state space. In the latter case, the process is instantaneously reflected from the boundary (in this case the boundary point is included in the state space; for models with instantaneously reflecting boundaries see Linetsky, 2005). Thus, the state space I is taken to be an open interval, $I = (e_1, e_2)$, unless any of the endpoints is a regular instantaneously reflecting boundary, in which case the interval is closed at that end.

Note that adding the killing rate $k(x) \geqslant 0$ may change the nature of the boundaries, i.e., the processes X with $k = 0$ and $\hat{X}$ with $k(x) \geqslant 0$ will, in general, have different boundary classifications. In particular, an *accessible* boundary (regular or exit) may become *inaccessible* if the killing rate increases fast enough towards the boundary so that the process is killed almost surely prior to reaching the boundary.

Before we can proceed with discussing spectral expansions, we need some preliminary material on one-dimensional diffusions. We follow Borodin and Salminen (2002, Chapter II). Let $T_z := \inf\{t \geqslant 0\colon X_t = z\}$ be the first hitting time of $z \in I$. Then for $\alpha > 0$, the non-negative random variable T_z has the Laplace transform:

$$\mathbb{E}_x\left[e^{-\alpha T_z}\right] = \begin{cases} \frac{\psi_\alpha(x)}{\psi_\alpha(z)}, & x \leqslant z, \\ \frac{\phi_\alpha(x)}{\phi_\alpha(z)}, & x \geqslant z, \end{cases} \tag{3.4}$$

where $\psi_\alpha(x)$ and $\phi_\alpha(x)$ are continuous solutions of the second-order ordinary differential equation (here $\mathcal{G}$ is the infinitesimal generator (3.1); Eq. (3.5) is the so-called *Sturm–Liouville (SL) equation*):

$$\mathcal{G}u(x) = \frac{1}{2}a^2(x)u''(x) + b(x)u'(x) - k(x)u(x) = \alpha u(x). \tag{3.5}$$

The functions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ can be characterized as the unique (up to a multiplicative constant dependent on α but independent of x) solutions of (3.5) by firstly demanding that $\psi_\alpha(x)$ is increasing in x and $\phi_\alpha(x)$ is decreasing, and secondly posing boundary conditions at regular boundary points. For $\psi_\alpha(x)$ the boundary condition is only needed at e_1 if e_1 is a regular boundary. If e_1 is a regular boundary specified as a killing boundary, we have a Dirichlet boundary condition:

$$\psi_\alpha(e_1+) = 0.$$

If e_1 is a regular boundary specified as an instantaneously reflecting boundary, we have a Neumann boundary condition:

$$\lim_{x\downarrow e_1} \frac{\psi'_\alpha(x)}{s(x)} = 0.$$

Similarly, if e_2 is a regular boundary specified as killing (instantaneously reflecting), we have a Dirichlet (Neumann) boundary condition for $\phi_\alpha(x)$ at e_2. At non-regular boundaries, the functions ψ_α and ϕ_α have the following properties for all $\alpha > 0$. If e_1 is entrance:

$$\psi_\alpha(e_1+) > 0, \quad \lim_{x\downarrow e_1} \frac{\psi'_\alpha(x)}{s(x)} = 0,$$
$$\phi_\alpha(e_1+) = +\infty, \quad \lim_{x\downarrow e_1} \frac{\phi'_\alpha(x)}{s(x)} > -\infty.$$

If e_1 is exit:

$$\psi_\alpha(e_1+) = 0, \quad \lim_{x\downarrow e_1} \frac{\psi'_\alpha(x)}{s(x)} > 0,$$
$$\phi_\alpha(e_1+) < +\infty, \quad \lim_{x\downarrow e_1} \frac{\phi'_\alpha(x)}{s(x)} = -\infty.$$

If e_1 is natural:

$$\psi_\alpha(e_1+) = 0, \quad \lim_{x\downarrow e_1} \frac{\psi'_\alpha(x)}{s(x)} = 0,$$
$$\phi_\alpha(e_1+) = +\infty, \quad \lim_{x\downarrow e_1} \frac{\phi'_\alpha(x)}{s(x)} = -\infty.$$

Analogous properties hold at e_2 with ψ and ϕ interchanged.

The functions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ are called fundamental solutions of the Sturm–Liouville equation (3.5). They are linearly independent for all $\alpha > 0$ and all solutions can be expressed as their linear combinations. Moreover, the *Wronskian* defined by (where $s(x)$ is the scale density defined in Eq. (3.3))

$$w_\alpha := \phi_\alpha(x)\frac{\psi'_\alpha(x)}{s(x)} - \psi_\alpha(x)\frac{\phi'_\alpha(x)}{s(x)} \tag{3.6}$$

is independent of x.

In the standard Markovian set-up, one considers a Banach space $C_b(I)$ of real-valued, bounded continuous functions on I. Then the transition operators $\mathcal{P}_t$ form a semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ in $C_b(I)$. The domain $\mathrm{Dom}(\mathcal{G})$ of the infinitesimal generator $\mathcal{G}$ of $\{\mathcal{P}_t, t \geqslant 0\}$ in $C_b(I)$ is:

$$\mathrm{Dom}(\mathcal{G}) = \big\{f \in C_b(I)\colon\ \mathcal{G}f \in C_b(I),\\ \text{boundary conditions at } e_1 \text{ and } e_2\big\}.$$

The boundary conditions are as follows (McKean, 1956, p. 522; Borodin and Salminen, 2002, pp. 16–17). If $e \in \{e_1, e_2\}$ is an exit boundary or a regular boundary specified as a killing boundary for the process $\hat{X}$, then the appropriate boundary condition at e is the vanishing (Dirichlet) boundary condition:

$$\lim_{x\to e} f(x) = 0. \tag{3.7}$$

If $e \in \{e_1, e_2\}$ is an entrance boundary or a regular boundary specified as an instantaneously reflecting boundary for the process X, then the appropriate boundary condition at e is the Neumann boundary condition:

$$\lim_{x\to e} \frac{f'(x)}{s(x)} = 0. \tag{3.8}$$

No boundary conditions are needed at natural boundaries in addition to the boundedness requirement $f, \mathcal{G}f \in C_b(I)$.

3.2 *The Laplace transform of the transition density*

Let $p_m(t; x, y)$ be the symmetric transition density with respect to the speed measure $m(\mathrm{d}x) = m(x)\,\mathrm{d}x$ and introduce the *Green's function*, the Laplace transform of the transition density taken with respect to time (where $\alpha > 0$):

$$G_\alpha(x, y) = \int_0^\infty e^{-\alpha t} p_m(t; x, y)\,\mathrm{d}t. \tag{3.9}$$

Then it is classical that (Borodin and Salminen, 2002, p. 19)

$$G_\alpha(x, y) = \begin{cases} w_\alpha^{-1}\psi_\alpha(x)\phi_\alpha(y), & x \leqslant y, \\ w_\alpha^{-1}\psi_\alpha(y)\phi_\alpha(x), & y \leqslant x. \end{cases} \tag{3.10}$$

Therefore, the transition density of a one-dimensional diffusion can be found by, firstly, determining the increasing and decreasing fundamental solutions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ of the Sturm–Liouville equation (3.5) and, secondly, inverting the Laplace transform (3.9):

$$p_m(t; x, y) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{\alpha t} G_\alpha(x, y)\,\mathrm{d}\alpha. \tag{3.11}$$

In this Bromwich Laplace transform inversion formula the integration is along the contour in the right half-plane parallel to the imaginary axes ($c > 0$) and $G_\alpha(x, y)$ is the analytic continuation of the Green's function (3.10) into the complex plane $\alpha \in \mathbb{C}$. If the functions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ are known in closed form, one can study the analytic continuation of the Green's function as a function of the complex variable $\alpha \in \mathbb{C}$ and apply the Cauchy Residue Theorem to calculate the Bromwich Laplace inversion integral. Since the Green's function is the integral kernel (*resolvent kernel*) of the resolvent operator $\mathcal{R}_\alpha$ (2.6), the singularities of G_α will lie on the negative real half-line $\alpha \leqslant 0$ (the spectrum of the non-positive self-adjoint operator $\mathcal{G}$ lies on the negative half-line), and the result of the application of the Cauchy Residue Theorem will produce the spectral expansion of the transition density. In Section 3.7 we will give more precise results on the analytic continuation of the fundamental solutions and Green's function. Recovering the spectral expansion via the Cauchy Residue Theorem constitutes the *Weyl–Titchmarsh complex variable approach to the Sturm–Liouville problem* (see Titchmarsh, 1962). Alternatively, the spectral expansion can also be constructed by purely real-variable techniques (for the real variable approach see Linetsky, 2004a, 2004b and references therein).

3.3 *The general form of the spectral representation for one-dimensional diffusions*

The semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ in the Banach space $C_b(I)$ restricted to $C_b(I) \cap L^2(I, m)$ extends uniquely to a strongly continuous semigroup of self-adjoint contractions in $L^2(I, m)$ with the infinitesimal generator $\mathcal{G}$, an unbounded self-adjoint, non-positive operator in $L^2(I, m)$ (McKean, 1956; see also Langer and Schenk, 1990 for a more recent reference). The domain of $\mathcal{G}$ in $L^2(I, m)$ is (McKean, 1956, p. 526 and Langer and Schenk, 1990, p. 15):

$$\mathrm{Dom}(\mathcal{G}) = \big\{ f \in L^2(I, m) \colon\ f, f' \in AC_{\mathrm{loc}}(I), \mathcal{G}f \in L^2(I, m), \\ \text{boundary conditions at } e_1 \text{ and } e_2 \big\},$$

where $AC_{\mathrm{loc}}(I)$ is the space of functions absolutely continuous over each compact subinterval of I. If $e \in \{e_1, e_2\}$ is either an exit or a regular boundary specified as a killing boundary for the process $\hat{X}$, then we have a vanishing (Dirichlet) boundary condition at e, Eq. (3.7). If $e \in \{e_1, e_2\}$ is an entrance boundary or a regular boundary specified as an instantaneously reflecting boundary for the process $\hat{X}$, then the appropriate boundary condition at e is the Neumann boundary condition equation (3.8). The self-adjointness of $\mathcal{G}$ is proved in McKean (1956) (see also Langer and Schenk, 1990, p. 15, Theorem 3.2).

The Spectral Theorem for self-adjoint semigroups in Hilbert space can now be applied to produce a spectral representation of the form (2.5) for the transition semigroup $\{\mathcal{P}_t, t \geqslant 0\}$. In this case the semigroup has a density. General

results on the spectral representation for the density of the transition semigroup of a one-dimensional diffusion were obtained by McKean (1956) (see also Ito and McKean, 1974, Section 4.11). Specializing McKean's general results to our case where the scale function, speed measure, and killing measure are absolutely continuous with respect to the Lebesgue measure, the symmetric transition density has the spectral representation:

$$p_m(t; x, y) = \int_{[0,\infty)} e^{-\lambda t} \sum_{i,j=1}^{2} u_i(x, \lambda) u_j(y, \lambda) \rho_{ij}(\mathrm{d}\lambda),$$
$$t > 0, \ x, y \in I, \tag{3.12}$$

where $u_i(x, \lambda)$, $i = 1, 2$, are solutions of the following two initial value problems for the Sturm–Liouville equation (3.5) ($x_0 \in (e_1, e_2)$ can be selected arbitrarily):

$$-\mathcal{G}u_i(x, \lambda) = \lambda u_i(x, \lambda), \quad x \in (e_1, e_2), \ i = 1, 2, \tag{3.13}$$

$$u_1(x_0, \lambda) = 1, \quad \frac{u_1'(x_0, \lambda)}{s(x_0)} = 0, \tag{3.14}$$

$$u_2(x_0, \lambda) = 0, \quad \frac{u_2'(x_0, \lambda)}{s(x_0)} = 1, \tag{3.15}$$

and $\rho(\mathrm{d}\lambda) = (\rho_{ij}(\mathrm{d}\lambda))_{i,j=1}^{2}$ is a Borel measure from $[0, \infty)$ to 2×2 symmetric non-negative definite matrices (*spectral matrix*):

$$\rho_{11}(\mathrm{d}\lambda) \geqslant 0, \quad \rho_{22}(\mathrm{d}\lambda) \geqslant 0, \quad \rho_{12}(\mathrm{d}\lambda) = \rho_{21}(\mathrm{d}\lambda),$$
$$\big(\rho_{12}(\mathrm{d}\lambda)\big)^2 \leqslant \rho_{11}(\mathrm{d}\lambda)\rho_{22}(\mathrm{d}\lambda). \tag{3.16}$$

The integral in Eq. (3.12) converges uniformly on compact squares in $I \times I$. McKean (1956) proved a number of smoothness properties for the symmetric transition density $p_m(t; x, y)$.

The spectral representation for the value function of a derivative asset with payoff $f \in L^2(I, m)$ can now be written in the form:

$$V(t, x) = \mathcal{P}_t f(x) = \int_I f(y)\, p_m(t; x, y) m(y)\, \mathrm{d}y \tag{3.17}$$

$$= \int_{[0,\infty)} e^{-\lambda t} \sum_{i,j=1}^{2} u_i(x, \lambda) F_j(\lambda) \rho_{ij}(\mathrm{d}\lambda), \quad x \in I, \ t \geqslant 0, \tag{3.18}$$

with the expansion coefficients

$$F_i(\lambda) = \int_I f(y) u_i(y, \lambda) m(y)\, \mathrm{d}y, \tag{3.19}$$

satisfying the Parseval equality

$$\|f\|^2 = \int\limits_{[0,\infty)} \sum_{i,j=1}^{2} F_i(\lambda) F_j(\lambda) \rho_{ij}(\mathrm{d}\lambda).$$

When there are no natural boundaries (both boundaries are exit, entrance, or regular), the spectrum is simple and purely discrete (McKean, 1956, Theorem 3.1). Let $\{-\lambda_n\}_{n=1}^{\infty}$, $0 \leqslant \lambda_1 < \lambda_2 < \cdots$, $\lim_{n\uparrow\infty} \lambda_n = \infty$, be the eigenvalues of $\mathcal{G}$ and let $\{\varphi_n\}_{n=1}^{\infty}$ be the corresponding eigenfunctions normalized so that $\|\varphi_n\|^2 = 1$. Then the spectral representation (3.12) for the symmetric density and the spectral representation (3.18) for the value function simplify to (for $t > 0$ the eigenfunction expansion converges uniformly on compact squares in $I \times I$):

$$p_m(t; x, y) = \sum_{n=1}^{\infty} e^{-\lambda_n t} \varphi_n(x) \varphi_n(y), \quad x, y \in I, \ t > 0, \tag{3.20}$$

$$V(t, x) = \mathcal{P}_t f(x) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} \varphi_n(x), \quad c_n = (f, \varphi_n). \tag{3.21}$$

When one or both boundaries are natural, the spectrum of $\mathcal{G}$ may have some non-empty (and possibly non-simple) essential spectrum, and the spectral representation in general takes the form (3.12), (3.18). However, under some additional regularity assumptions, the form of the spectral expansion can be significantly simplified can be simplified.

Remark on non-L^2 payoffs. The spectral expansion (3.18) (or (3.21) when the spectrum is purely discrete) is valid for payoffs in $L^2(I, m)$. If the payoff is not in $L^2(I, m)$ but the integral in Eq. (3.17) exists, one can write the value function in the form (3.17) with the transition probability density given by Eq. (3.12) (or (3.20) when the spectrum is discrete). However, one cannot interchange the integration in y and the integration with respect to the spectral measure to obtain the spectral expansion (3.18) (or (3.22)) for the value function. This situation is frequent in finance applications. Among applications discussed in Section 4, examples of non-L^2 payoffs include call and put options in Merton's cash dividend model, Asian-style call options, vanilla call options and down-and-out call options in the CEV model, and defaultable bonds in the jump-to-default extended Black–Scholes model in Linetsky (2004b). When X is an asset price process, in many of these applications constants and/or linear functions are not in $L^2(I, m)$. In some cases (e.g., Asian call, CEV call), one is able to decompose the payoff into a linear combination of an $L^2(I, m)$-payoff (e.g., put) plus an affine position $a + bX$ (e.g., payoff of a forward contract). The value function for the affine position is found directly, and the value function for the $L^2(I, m)$-payoff is given by the spectral expansion (a typical

application of the call-put parity relationship). However, in some applications either the payoff cannot be represented as a linear combination of an affine position plus an $L^2(I, m)$-payoff, or the value function for the affine payoff cannot be easily found directly. An example of the first type arises in Merton's cash dividend model (Lewis, 1998). An example of the second type arises in pricing down-and-out call options in the CEV model (Davydov and Linetsky, 2003). In the latter problem, we can decompose the down-and-out call payoff into a linear combination of the down-and-out put payoff (which is L^2) and a down-and-out forward position (not L^2). However, we need to determine the value function for the down-and-out forward, which is not straightforward. In such cases, one can start with the representation (3.17) for the value function, but instead of the spectral representation (3.12) for the transition density one uses the alternative representation (3.11) as the inverse Laplace transform of the Green's function. The next steps are interchanging the integration in y with the Laplace inversion, integrating the Green's function against the payoff, and finally inverting the Laplace transform. The Laplace inversion leads to a spectral-type representation. However, it includes additional terms in addition to the ones in Eq. (3.18) that come from additional singularities in the Laplace transform when analytically continued to the complex plane in problems with non-L^2 payoffs. See Lewis (1998), Davydov and Linetsky (2003), Linetsky (2004b, 2006) for more details.

3.4 *Spectral classification of one-dimensional diffusions*

To investigate the qualitative nature of the spectrum and describe simplifications in the general form of the spectral expansion when natural boundaries are present we need some background from the Sturm–Liouville theory. The volume by Amrein et al. (2005) is an excellent recent reference (see also Titchmarsh, 1962; Dunford and Schwartz, 1963; Glazman, 1965; Levitan and Sargsjan, 1975; Weidmann, 1987; Fulton et al., 1996). Consider the SL equation (3.5) rewritten in the self-adjoint form and with $\alpha = -\lambda \in \mathbb{C}$:

$$-\frac{1}{2}a^2(x)u''(x) - b(x)u'(x) + k(x)u(x)$$
$$\equiv -\frac{1}{m(x)}\left(\frac{u'(x)}{s(x)}\right)' + k(x)u(x) = \lambda u(x), \quad x \in (e_1, e_2). \tag{3.22}$$

The oscillatory/non-oscillatory classification based on Sturm's theory of oscillations of solutions is of fundamental importance in determining the qualitative nature of the spectrum of the SL operator. For a given real λ, Eq. (3.22) is *oscillatory* at an endpoint e if and only if every solution has infinitely many zeros clustering at e. Otherwise it is called *non-oscillatory* at e. This classification is mutually exclusive for a fixed λ, but can vary with λ. For Eq. (3.22), there are two distinct possibilities at each endpoint.

Theorem 3.1 *(Oscillatory/Non-oscillatory Classification of Boundaries). Let* $e \in \{e_1, e_2\}$ *be an endpoint of Eq.* (3.22)*. Then* e *belongs to one and only one of the following two cases*:

(i) *Equation* (3.22) *is non-oscillatory at* e *for all real* λ*. Correspondingly, the endpoint* e *is said to be non-oscillatory.*

(ii) *There exists a real number* $\Lambda \geqslant 0$ *such that Eq.* (3.22) *is oscillatory at* e *for all* $\lambda > \Lambda$ *and non-oscillatory at* e *for all* $\lambda < \Lambda$*. Correspondingly,* e *is said to be oscillatory with cutoff* Λ*. Equation* (3.22) *can be either oscillatory or non-oscillatory at* e *for* $\lambda = \Lambda > 0$*. It is always non-oscillatory for* $\lambda = 0$*.*

Based on the oscillatory/non-oscillatory classification of boundaries, the spectrum of the non-negative operator $-\mathcal{G}$ is classified as follows.

Theorem 3.2 *(Spectral Classification).*

(i) **Spectral Category I.** *If both endpoints are non-oscillatory, then the spectrum is simple, non-negative and purely discrete.*

(ii) **Spectral Category II.** *If one of the endpoints is non-oscillatory and the other endpoint is oscillatory with cutoff* $\Lambda \geqslant 0$*, then the spectrum is simple and non-negative, the essential spectrum is nonempty,* $\sigma_e(-\mathcal{G}) \subset [\Lambda, \infty)$*, and* Λ *is the lowest point of the essential spectrum. If the SL equation is non-oscillatory at the oscillatory endpoint for* $\lambda = \Lambda \geqslant 0$*, then there is a finite set of simple eigenvalues in* $[0, \Lambda]$ (*it may be empty*)*. If the SL equation is oscillatory at the oscillatory endpoint for* $\lambda = \Lambda > 0$*, then there is an infinite sequence of simple eigenvalues in* $[0, \Lambda)$ *clustering at* Λ*.*

(iii) **Spectral Category III.** *If* e_1 *is oscillatory with cutoff* $\Lambda_1 \geqslant 0$ *and* e_2 *is oscillatory with cutoff* $\Lambda_2 \geqslant 0$*, then the essential spectrum is nonempty,* $\sigma_e(-\mathcal{G}) \subset [\underline{\Lambda}, \infty)$*,* $\underline{\Lambda} := \min\{\Lambda_1, \Lambda_2\}$*, and* $\underline{\Lambda}$ *is the lowest point of the essential spectrum. The spectrum is simple* (*has multiplicity one*) *below* $\overline{\Lambda} := \max\{\Lambda_1, \Lambda_2\}$ *and is not simple* (*has multiplicity two*) *above* $\overline{\Lambda}$*. If the SL equation is non-oscillatory for* $\lambda = \underline{\Lambda} \geqslant 0$*, then there is a finite set of simple eigenvalues in* $[0, \underline{\Lambda}]$ (*it may be empty*)*. If the SL equation is oscillatory for* $\lambda = \underline{\Lambda} > 0$*, then there is an infinite sequence of simple eigenvalues in* $[0, \underline{\Lambda})$ *clustering at* $\underline{\Lambda}$*.*

Regular, exit, and entrance boundaries in Feller's classification are always non-oscillatory for the associated SL equation, and, if there are no natural boundaries, the spectrum of the infinitesimal generator is purely discrete. Natural boundaries can be either non-oscillatory or oscillatory with cutoff $\Lambda \geqslant 0$.

3.5 The Liouville transformation

To determine when a natural boundary is non-oscillatory or oscillatory with cutoff Λ, it is convenient to transform the SL equation (3.22) to the so-called

Liouville normal form (e.g., Everitt, 2005, p. 280). Before introducing the *Liouville transformation* that transforms the SL equation to the Liouville normal form, we first consider a general class of transformations of the SL equation. Let $g(x)$ be a strictly increasing twice continuously differentiable function on (e_1, e_2) and $H(x)$ a twice continuously differentiable function on (e_1, e_2). Introduce new independent and dependent variables in the SL equation (3.22) according to:

$$y := g(x), \qquad v(y) := \left\{e^{H(x)}u(x)\right\}\big|_{x=g^{-1}(y)}, \tag{3.23}$$

where $x = g^{-1}(y)$ is the inverse of $y = g(x)$. Then it is easy to show by direct calculation that the function $v(y)$ satisfies the transformed SL equation of the form (3.22) on the transformed interval $(g(e_1), g(e_2))$ and with the transformed coefficients:

$$\tilde{a}(y) = \left\{a(x)g'(x)\right\}\big|_{x=g^{-1}(y)}, \tag{3.24}$$

$$\tilde{b}(y) = \left\{b(x)g'(x) + \frac{1}{2}a^2(x)\left[g''(x) - 2H'(x)g'(x)\right]\right\}\Big|_{x=g^{-1}(y)}, \tag{3.25}$$

$$\tilde{k}(y) = \left\{k(x) + b(x)H'(x) + \frac{1}{2}a^2(x)\left[H''(x) - \left(H'(x)\right)^2\right]\right\}\Big|_{x=g^{-1}(y)}. \tag{3.26}$$

In particular, fix some $x_0 \in (e_1, e_2)$ and consider a mapping $g : (e_1, e_2) \to (g(e_1), g(e_2))$:

$$g(x) := \int_{x_0}^{x} \frac{\mathrm{d}z}{a(z)}. \tag{3.27}$$

Since $a(x) > 0$ on (e_1, e_2), $g(x)$ is strictly increasing on (e_1, e_2). Let g^{-1} denote its inverse. Now we transform the independent and dependent variables in the SL equation as follows (in this case $e^{H(x)} = (a(x)s(x))^{-1/2}$, or $H(x) = \int_{x_0}^{x} \frac{b(z)}{a^2(z)}\,\mathrm{d}z - \frac{1}{2}\ln a(x)$):

$$y = g(x) = \int_{x_0}^{x} \frac{\mathrm{d}z}{a(z)}, \qquad v(y) = \left\{\frac{u(x)}{\sqrt{a(x)s(x)}}\right\}\Big|_{x=g^{-1}(y)}. \tag{3.28}$$

Then the function $v(y)$ satisfies the SL equation in the *Liouville normal form* with $\tilde{a}(y) = 1$, $\tilde{b}(y) = 0$, and $\tilde{c}(y) = Q(y)$:

$$-\frac{1}{2}v''(y) + Q(y)v(y) = \lambda v(y), \quad y \in \left(g(e_1), g(e_2)\right), \tag{3.29}$$

where the *potential function* $Q(y)$ is given by

$$Q(y) = U\left(g^{-1}(y)\right), \tag{3.30}$$

where

$$U(x) := \frac{1}{8}\big(a'(x)\big)^2 - \frac{1}{4}a(x)a''(x) + \frac{b^2(x)}{2a^2(x)} + \frac{1}{2}b'(x) - \frac{b(x)a'(x)}{a(x)} + k(x). \tag{3.31}$$

This transformation of the dependent and independent variable[4] is called the *Liouville transformation* in the Sturm–Liouville theory. It reduces the SL equation (3.22) to the Liouville normal form. The SL equation in the Liouville normal form has the form of the celebrated (stationary) *one-dimensional Schrödinger equation*: the coefficient in front of the second derivative term is equal to (negative) one-half[5], there is no first derivative term, and all the information about the dynamics is encoded in the potential function $Q(y)$ (as well as in the boundary conditions). In Section 4.1 we will give a probabilistic interpretation to the Liouville transform.

3.6 Further results on spectral classification

The oscillatory/non-oscillatory classification of boundaries of the SL equation remains invariant under the Liouville transformation, i.e., the SL equation (3.22) is non-oscillatory at $e \in \{e_1, e_2\}$ for a particular λ if and only if the Schrödinger equation (3.29) is non-oscillatory at $g(e)$ for that λ. The oscillatory/non-oscillatory classification of the Schrödinger equation depends on the behavior of the potential function Q near the endpoints. We have the following classification result.

Theorem 3.3 *(Oscillatory/Non-Oscillatory Classification of Natural Boundaries). Suppose $e \in \{e_1, e_2\}$ is a natural boundary, $U(x)$ is defined in Eq. (3.31), and the limit $\lim_{x\to e} U(x)$ exists (it is allowed to be infinite).*

(i) *If e is transformed into a finite endpoint by the Liouville transformation, i.e., $g(e) = \int_{x_0}^{e} \frac{\mathrm{d}z}{a(z)}$ is finite, then e is non-oscillatory.*

(ii) *Suppose e is transformed into $-\infty$ or $+\infty$ by the Liouville transformation. If $\lim_{x\to e} U(x) = +\infty$, then e is non-oscillatory. If $\lim_{x\to e} U(x) = \Lambda$ for some finite Λ, then e is oscillatory with cutoff Λ. Since the operator $-\mathcal{G}$ is non-negative, it follows that $\Lambda \geqslant 0$. If $\Lambda > 0$ and $\lim_{x\to e} g^2(x)(U(x) - \Lambda) > -1/4$, then e is non-oscillatory for $\lambda = \Lambda > 0$. If $\Lambda > 0$ and*

[4] Note that since the choice of x_0 is arbitrary, g is defined up to a constant. Different choices of the constant correspond to translations of the interval $(e_1, e_2) \to (e_1 + c, e_2 + c)$.

[5] The standard form of the Schrödinger operator is $-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + Q(x)$. We retain the "probabilistic" factor $1/2$ in $-\frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + Q(x)$ to interpret it as the infinitesimal generator of standard Brownian motion killed at the rate Q. Alternatively, Brownian motion can be taken to run twice as fast.

$\lim_{x\to e} g^2(x)(U(x) - \Lambda) < -1/4$, *then* e *is oscillatory for* $\lambda = \Lambda > 0$. *If* $\Lambda = 0$, e *is always non-oscillatory for* $\lambda = \Lambda = 0$.

Theorem 3.2 tells us that oscillatory natural boundaries generate some non-empty essential spectrum above the cutoff, but it does not identify the essential spectrum. It is well known that when the potential function oscillates towards an infinite boundary, the essential spectrum above the cutoff may have a complicated structure. In particular, it may consist of an infinite sequence of disjoint intervals separated by gaps. Furthermore, eigenvalues may be present in the gaps or embedded in the continuous spectrum. The assumption in Theorem 3.3 on the existence of the limit $\lim_{x\to e} U(x)$ combined with the assumption that $U(x)$ has bounded variation in a neighborhood of a natural boundary yields a drastic simplification of the essential spectrum.

Theorem 3.4 *(Essential Spectrum Generated by oscillatory Natural Boundaries). Suppose the limit* $\lim_{x\to e} U(x)$ *exists* (*it is allowed to be infinite*) *and* $U(x)$ *has bounded variation in a neighborhood of each oscillatory natural boundary.*[6]

(i) **Spectral Category II.** *If one of the boundaries is non-oscillatory and the other boundary is oscillatory with cutoff* $\Lambda \geqslant 0$, *then the essential spectrum of* $\mathcal{A}$ *is* $\sigma_e(\mathcal{A}) = [\Lambda, \infty)$. *Moreover,* $\mathcal{A}$ *has purely absolutely continuous spectrum in* (Λ, ∞).

(ii) **Spectral Category III.** *If* e_1 *is oscillatory with cutoff* $\Lambda_1 \geqslant 0$ *and* e_2 *is oscillatory with cutoff* $\Lambda_2 \geqslant 0$, *then the essential spectrum of* $\mathcal{A}$ *is* $\sigma_e(\mathcal{A}) = [\underline{\Lambda}, \infty)$, $\underline{\Lambda} := \min\{\Lambda_1, \Lambda_2\}$. *Moreover,* $\mathcal{A}$ *has purely absolutely continuous spectrum in* $(\underline{\Lambda}, \infty)$. *The part of the spectrum below* $\overline{\Lambda} = \max\{\Lambda_1, \Lambda_2\}$ *is simple* (*has multiplicity one*). *The part of the spectrum above* $\overline{\Lambda}$ *is not simple.*

Theorem 3.4 tells us that, under our assumptions, the spectrum above the cutoff is purely absolutely continuous. The bounded variation assumption rules out coefficients that oscillate towards oscillatory natural boundaries. They are not particularly restrictive for financial applications (essentially all diffusion models in finance satisfy these assumptions), but yield a drastic simplification of the structure of the essential spectrum. Under this assumption, one can read off the qualitative structure of the spectrum directly from the behavior of the coefficient functions a, b, and c near the boundaries, by studying the behavior of $U(x)$ near the boundaries.

[6] We assume that if e_1 (e_2) is a oscillatory natural boundary, then there is such $c \in (e_1, e_2)$ that $U(x)$ has bounded variation on $(e_1, c]$ ($[c, e_2)$).

3.7 The simplified form of the spectral representation for one-dimensional diffusions

3.7.1 Spectral Category I

When there are no oscillatory natural boundaries, the spectrum is simple, non-negative and purely discrete and the spectral representation for the transition density and the spectral expansion for the value function take the form of eigenfunction expansions (3.20) and (3.21), respectively (Linetsky, 2004a). We now described a general procedure to explicitly determine the eigenvalues and eigenfunctions in this case. Recall that, for $\alpha > 0$, $\psi_\alpha(x)$ and $\phi_\alpha(x)$ were defined as the fundamental solutions of the SL equation (3.5). Assume that both boundaries are non-oscillatory. It turns out that in this case $\psi_\alpha(x)$ and $\phi_\alpha(x)$ can be normalized[7] so that they can be analytically continued to the whole complex plane $\alpha \in \mathbb{C}$ and are entire functions of the complex variable α for each fixed x (see Linetsky, 2004a, Lemma 1, p. 351). Hence, the Wronskian w_α can also be analytically continued to the whole complex plane $\alpha \in \mathbb{C}$ and is entire in α.

An eigenfunction $\varphi_n(x)$ satisfies the SL equation (3.5) with $\alpha = -\lambda_n$, is square-integrable with m in a neighborhood of e_1, and satisfies the appropriate boundary condition at e_1. Hence it must be equal to $\psi_{-\lambda_n}(x)$ up to a non-zero constant multiple. But $\varphi_n(x)$ is also square-integrable with m in a neighborhood of e_2 and satisfies the appropriate boundary condition at e_2. Hence it must also be equal to $\phi_{-\lambda_n}(x)$ up to a non-zero constant multiple. Thus, for $\alpha = -\lambda_n$, $\psi_{-\lambda_n}(x)$ and $\phi_{-\lambda_n}(x)$ must be linearly dependent: $\phi_{-\lambda_n}(x) = A_n \psi_{-\lambda_n}(x)$ for some non-zero constant A_n and, hence, their Wronskian must vanish for $\alpha = -\lambda_n$, $w_{-\lambda_n} = 0$.

Conversely, let $\alpha = -\lambda_n$ be a zero of the Wronskian. Then $\psi_{-\lambda_n}(x)$ and $\phi_{-\lambda_n}(x)$ are linearly dependent with some non-zero constant A_n and, hence, $\psi_{-\lambda_n}(x)$ is a solution that is square-integrable with m on (e_1, e_2) and satisfies the appropriate boundary conditions at *both* endpoints e_1 and e_2, i.e., $\psi_{-\lambda_n}(x)$ is a *non-normalized* eigenfunction corresponding to the eigenvalue λ_n. Therefore, $\{-\lambda_n\}_{n=1}^{\infty}$ are zeros of w_α, and $\psi_{-\lambda_n}(x)$ are the corresponding non-normalized eigenfunctions. Since all eigenvalues of $-\mathcal{G}$ with $k(x) \geqslant 0$ are simple and non-negative, all zeros of the Wronskian w_α are simple and non-positive. Thus, practically, to determine the eigenvalues, find the fundamental solutions ψ and ϕ normalized so that their analytic continuations to complex α are entire functions of the complex variable α for each fixed x, compute their Wronskian, and find its zeros. The negatives of the zeros are the sought after non-negative eigenvalues of $-\mathcal{G}$. Finally, the *normalized* eigenfunctions can be

[7] Recall that $\psi_\alpha(x)$ and $\phi_\alpha(x)$ have been defined up to overall factors independent of x (but the normalization factors can depend on α).

taken in the form:

$$\varphi_n(x) = \pm\sqrt{\frac{A_n}{w'_{-\lambda_n}}}\psi_{-\lambda_n}(x) = \pm\frac{\phi_{-\lambda_n}(x)}{\sqrt{A_n w'_{-\lambda_n}}},$$

$$\text{where } w'_{-\lambda_n} \equiv -\frac{\mathrm{d}w_\alpha}{\mathrm{d}\alpha}\bigg|_{\alpha=-\lambda_n}. \tag{3.32}$$

This normalization can be obtained by applying the Cauchy Residue Theorem to invert the Laplace transform (3.11) (e.g., Davydov and Linetsky, 2003, p. 188) or by direct calculation of the norms (see Linetsky, 2004a, pp. 352–353).

3.7.2 Spectral Category II

Suppose e_1 is non-oscillatory and e_2 is an oscillatory natural boundary with cutoff $\Lambda \geqslant 0$ (the case of oscillatory e_1 and non-oscillatory e_2 is treated similarly). Under our assumptions, the essential spectrum of $-\mathcal{G}$ is $\sigma_e(-\mathcal{G}) = [\Lambda, \infty)$ and is simple. Moreover, $-\mathcal{G}$ has purely absolutely continuous spectrum in (Λ, ∞). If e_2 is non-oscillatory for $\lambda = \Lambda \geqslant 0$, then there is a finite set of simple eigenvalues in $[0, \Lambda]$ (it may be empty). If e_2 is oscillatory for $\lambda = \Lambda > 0$, then there is an infinite sequence of simple eigenvalues in $[0, \Lambda)$ clustering towards Λ. Accordingly, the spectral representation for the symmetric transition density (3.12) and the spectral expansion for the value function (3.18) simplify to:

$$p_m(t; x, y) = \sum_n e^{-\lambda_n t}\varphi_n(x)\varphi_n(y) + \int_\Lambda^\infty e^{-\lambda t}\psi_{-\lambda}(x)\psi_{-\lambda}(y)\,\mathrm{d}\rho_{ac}(\lambda),$$

$$t > 0, \tag{3.33}$$

$$V(t, x) = \sum_n c_n e^{-\lambda_n t}\varphi_n(x) + \int_\Lambda^\infty e^{-\lambda t}F(\lambda)\psi_{-\lambda}(x)\,\mathrm{d}\rho_{ac}(\lambda),$$

$$x \in I,\ t \geqslant 0, \tag{3.34}$$

$$c_n = (f, \varphi_n), \quad F(\lambda) = \int_I f(y)\psi_{-\lambda}(y)m(y)\,\mathrm{d}y,$$

$$f \in L^2(I, m). \tag{3.35}$$

Here $\varphi_n(x)$ are eigenfunctions corresponding to the eigenvalues λ_n (if any), $\psi_{-\lambda}(x)$ is the fundamental solution of the SL equation normalized so that it is an entire function of λ (recall that e_1 is non-oscillatory and, hence, such a normalization is possible), and $\rho_{ac}(\lambda)$ is the spectral function absolutely continuous on (Λ, ∞) and normalized relative to $\psi_{-\lambda}(x)$.

The eigenvalues and eigenfunctions below Λ are determined in the same way as for the Spectral Category I with one modification. Consider the endpoint e_1 and suppose it is oscillatory with cutoff Λ_1. Then the solution $\psi_{-\lambda}(x)$

can be normalized so that its analytic continuation to complex λ is analytic in the half-plane S_{Λ_1} for each fixed x, where $S_{\Lambda_1} := \{\lambda \in \mathbb{C}\colon \mathrm{Re}(\lambda) < \Lambda_1\}$. Similarly, if e_2 is oscillatory with cutoff Λ_2, then the solution $\phi_{-\lambda}(x)$ is analytic in the half-plane $S_{\Lambda_2} := \{\lambda \in \mathbb{C}\colon \mathrm{Re}(\lambda) < \Lambda_2\}$ (recall that for non-oscillatory endpoints these solutions are analytic in the whole complex plane). If e_1 is non-oscillatory and e_2 is oscillatory with cutoff Λ, then the eigenvalues λ_n can be found as zeros (if any) of the Wronskian of the two solutions $\psi_{-\lambda}(x)$ and $\phi_{-\lambda}(x)$ analytic in the half-plane S_Λ (in this case $\psi_{-\lambda}(x)$ is analytic in the whole complex plane and $\phi_{-\lambda}(x)$ is analytic in the half-plane S_Λ; hence, their Wronskian is analytic in the half-plane S_Λ). The case with non-oscillatory e_2 and oscillatory is treated similarly.

Since the absolutely continuous spectrum is simple, we are able to write down the continuous part of the spectral expansion in terms of just one spectral function in contrast to the general case that requires a 2×2 spectral matrix. There are two approaches to obtain the spectral function, the Weyl–Titchmarsh complex variable approach (Titchmarsh, 1962) and the real variable approach of Levitan (1950) and Levinson (1951) (see also Coddington and Levinson, 1955; McKean, 1956, and Levitan and Sargsjan, 1975). The complex variable approach consists in inverting the Laplace transform via the Cauchy Residue Theorem. This calculation is illustrated in detail in Linetsky (2004d, proof of Proposition 1), and Linetsky (2004b, proof of Proposition 3.3 in Appendix D). The alternative real variable approach proceeds as follows. Suppose e_2 is the oscillatory natural boundary. Then consider the problem on (e_1, b) with some $b < e_2$, impose the killing boundary condition at b (Dirichlet boundary condition for the associated SL problem), obtain the spectral expansion for this problem, and then take the limit $b \uparrow e_2$. This calculation is illustrated in detail in Linetsky (2004b; 2004a, p. 355).

3.7.3 Spectral Category III

Suppose e_1 is an oscillatory natural boundary with cutoff $\Lambda_1 \geqslant 0$ and e_2 is an oscillatory natural boundary with cutoff $\Lambda_2 \geqslant 0$. To be specific, suppose $\Lambda_1 < \Lambda_2$. Under our assumptions, the essential spectrum of $-\mathcal{G}$ is $\sigma_e(-\mathcal{G}) = [\Lambda_1, \infty)$. Moreover, $-\mathcal{G}$ has purely absolutely continuous spectrum in (Λ_1, ∞). The part of the spectrum below Λ_2 is simple. The part of the spectrum above Λ_2 is not simple. If the SL equation (3.22) is non-oscillatory for $\lambda = \Lambda_1$, there is a finite set of simple eigenvalues in $[0, \Lambda_1]$ (it may be empty). If the SL equation (3.22) is oscillatory for $\lambda = \Lambda_1 > 0$, then there is an infinite sequence of simple eigenvalues in $[0, \Lambda_1)$ clustering towards Λ_1. The general form of the spectral expansion (3.22) reduces to the form containing three parts: the sum over the eigenvalues (if any), the integral over the simple absolutely continuous spectrum (Λ_1, Λ_2) similar to the Spectral Category II with the single spectral function, and the integral over the non-simple portion of the absolutely continuous spectrum above Λ_2 that has the general form (3.12) with the 2×2 spectral matrix.

The explicit form of the spectral expansion can be obtained either via the Weyl–Titchmarsh complex variable approach by inverting the Laplace transform of the Green's function (an example calculation is given in Section 4.2 below), or via the real variable approach. In the latter approach, pick some a and b, $e_1 < a < b < e_2$, and kill the process at the first exit time from (a, b) (impose Dirichlet boundary conditions for the SL equation at both a and b). Consider the spectral representation for this process with two regular killing boundaries, and then take the limit $a \downarrow e_1$ and $b \uparrow e_2$. McKean's (1956) original derivation followed this approach. An example of this calculation is given in Linetsky (2004a, pp. 356–357).

4 One-dimensional diffusions: a catalog of analytically tractable models

4.1 Transformations of one-dimensional diffusions: transforming the state space and changing the probability measure

In order to determine the spectral representation explicitly, one needs explicit solutions of the Sturm–Liouville equation. The Liouville transformation reduces the SL equation (3.22) on the interval (e_1, e_2) and with coefficients $(a(x), b(x), k(x))$ to the Schrödinger equation (3.29) with potential function (3.30)–(3.31). If analytical solutions are available in terms of known special functions for the Schrödinger equation, inverting the Liouville transformation yields analytical solutions for the original Sturm–Liouville equation, and the spectral representation can be constructed explicitly. The celebrated Schrödinger equation is the fundamental equation of quantum mechanics and has been intensively studied in mathematical physics. Various specifications of potentials such that the Schrödinger equation admits analytical solutions in terms of certain classes of special functions (in particular, hypergeometric and confluent hypergeometric functions) have been intensively studied (Grosche and Steiner, 1998 provide extensive bibliography).

Conversely, suppose that $Q(x)$ defined on some interval (e_1, e_2) is an analytically tractable potential function. Then for each strictly increasing and twice differentiable function $g(x)$ and twice differentiable function $H(x)$, we can construct a diffusion process on $(g(e_1), g(e_2))$ with the infinitesimal parameters $(\tilde{a}(y), \tilde{b}(y), \tilde{k}(y))$ given by Eqs. (3.24)–(3.26) with $a(x) = 1$, $b(x) = 0$, and $k(x) = Q(x)$. We can thus generate a family of diffusion processes associated with a given potential function and parameterized by two functions $g(x)$ and $H(x)$. Since we have the analytical solution of the Schrödinger equation in hand, we immediately obtain analytical solutions to the SL equations for this family of diffusions by inverting the Liouville transformation for each of the processes.

We now give a probabilistic interpretation to the Liouville transformation as a composition of the state space transformation and the change of probability measure. Suppose we are given an open interval (e_1, e_2) and two functions

$a(x)$ and $b(x)$ and consider an SDE (our regularity assumptions on a and b are in force, i.e., $a \in C^2(e_1, e_2)$, $a(x) > 0$ on (e_1, e_2), and $b \in C^1(e_1, e_2)$):

$$\mathrm{d}X_t = a(X_t)\,\mathrm{d}B_t + b(X_t)\,\mathrm{d}t, \quad X_0 = x \in (e_1, e_2), \tag{4.1}$$

where B is a standard Brownian motion. Under our assumptions this SDE can be solved uniquely up to the first exit time from (e_1, e_2), $\tau_{e_1,e_2} = T_{e_1} \wedge T_{e_2}$. If according to Feller's boundary classification a and b are such that both boundaries are inaccessible (natural or entrance) for the diffusion process with volatility a and drift b, then $\tau_{e_1,e_2} = \infty$ a.s., there is no explosion, and the process lives in the open interval (e_1, e_2) forever. Otherwise, at the first hitting time of an accessible boundary (exit or regular) the solution process X is killed, i.e., sent to the cemetery state Δ (in particular, in this section we assume that regular boundaries are always specified as killing boundaries; we do not deal with reflection here).

Suppose further that $k(x) \in C(e_1, e_2)$ is also given and is such that $k(x) \geqslant 0$ on (e_1, e_2). Our aim is to calculate the expectation:

$$V(t, x) = \mathbb{E}_x^X\Big[e^{-\int_0^t k(X_u)\,\mathrm{d}u}\mathbf{1}_{\{\tau_{e_1,e_2}>t\}}f(X_t)\Big], \tag{4.2}$$

for some f such that the expectation exists ($\mathbb{E}_x^X$ is with respect to the law of X starting at x; if both boundaries are inaccessible for X, we can drop the indicator $\mathbf{1}_{\{\tau_{e_1,e_2}>t\}}$).

Introduce a process Y: $\{Y_t = g(X_t),\ t < \tau_{e_1,e_2},\ Y_t = \Delta,\ t \geqslant \tau_{e_1,e_2}\}$, where g is defined as in (3.27). The process Y is a one-dimensional diffusion on the interval $(g(e_1), g(e_2))$ with unit diffusion and drift:

$$a^Y(y) = 1, \qquad b^Y(y) = \mu(y) := \left\{\frac{b(x)}{a(x)} - \frac{1}{2}a'(x)\right\}\Big|_{x=g^{-1}(y)}. \tag{4.3}$$

In terms of the process Y the expectation to be computed takes the form:

$$V(t, x) = \mathbb{E}_{g(x)}^Y\Big[e^{-\int_0^t k(g^{-1}(Y_u))\,\mathrm{d}u}\mathbf{1}_{\{\tau_{g(e_1),g(e_2)}>t\}}f\big(g^{-1}(Y_t)\big)\Big],$$

where $\tau_{g(e_1),g(e_2)}$ is the first exit time of the process Y from $(g(e_1), g(e_2))$ ($\mathbb{E}_y^Y$ is with respect to the law of Y starting at y).

We now observe that, by Girsanov's theorem and Ito's formula, up to the first exit time from $(g(e_1), g(e_2))$ we have:

$$\begin{aligned} V(t, x) &= \mathbb{E}_{g(x)}^W\Big[e^{M(W_t)-M(g(x))-\frac{1}{2}\int_0^t(\mu^2(W_u)+\mu'(W_u))\,\mathrm{d}u} \\ &\qquad \times e^{-\int_0^t k(g^{-1}(W_u))\,\mathrm{d}u}\mathbf{1}_{\{\tau_{g(e_1),g(e_2)}>t\}}f\big(g^{-1}(W_t)\big)\Big] \\ &= \sqrt{s(x)a(x)}\,\mathbb{E}_{g(x)}^W\Big[e^{-\int_0^t Q(W_u)\,\mathrm{d}u}\mathbf{1}_{\{\tau_{g(e_1),g(e_2)}>t\}} \\ &\qquad \times \frac{f(g^{-1}(W_t))}{\sqrt{s(g^{-1}(W_t))a(g^{-1}(W_t))}}\Big], \end{aligned} \tag{4.4}$$

where $\mathbb{E}_y^W$ is with respect to the law of standard Brownian motion W started at y, $M(y) = \int^y \mu(y)\,\mathrm{d}y$ is the indefinite integral of the drift $\mu(y)$ (by Ito's formula, $M(W_t) = M(y) + \int_0^t \mu(W_u)\,\mathrm{d}W_u + \frac{1}{2}\int_0^t \mu'(W_u)\,\mathrm{d}u$), and $\tau_{g(e_1),g(e_2)}$ is the first exit time of standard Brownian motion W started at y from $(g(e_1), g(e_2))$. Direct calculation shows that

$$e^{M(z)-M(y)} = \sqrt{\frac{s(g^{-1}(y))a(g^{-1}(y))}{s(g^{-1}(z))a(g^{-1}(z))}}$$

and

$$k\big(g^{-1}(y)\big) + \frac{1}{2}\big(\mu^2(y) + \mu'(y)\big) = Q(y),$$

where $Q(y)$ is given by Eqs. (3.30)–(3.31).

Thus, the transition semigroup of the process $\hat{X}$ on the interval (e_1, e_2) with the infinitesimal parameters $(a(x), b(x), k(x))$ and with regular boundaries (if any) specified as killing can be expressed in terms of the Feynman–Kac semigroup of Brownian motion on the interval $(g(e_1), g(e_2))$ discounted at the rate $Q(x)$ and killed at the first exit time from $(g(e_1), g(e_2))$ if the boundaries are accessible. The Schrödinger equation (3.29) with potential $Q(x)$ (3.30)–(3.31) is then the Sturm–Liouville equation associated with the Brownian motion killed at the rate $Q(x)$. If $p(t; x, y)$ is the transition density (with respect to the Lebesgue measure) of the process $\hat{X}$ solving the backward Kolmogorov equation

$$\frac{1}{2}a^2(x)p_{xx} + b(x)p_x - k(x)p = p_t \tag{4.5}$$

with the Dirac delta initial condition $p(0; x, y) = \delta(x - y)$ and appropriate boundary conditions at the endpoints, and

$$p^Q(t; x, y) = \frac{\partial}{\partial y}\mathbb{E}_x^W\left[e^{-\int_0^t Q(W_u)\,\mathrm{d}u}\mathbf{1}_{\{\tau_{g(e_1),g(e_2)}>t\}}\mathbf{1}_{\{W_t\leqslant y\}}\right] \tag{4.6}$$

is the density of the Feynman–Kac semigroup of Brownian motion discounted at the rate $Q(x)$ (also called the *heat kernel of the Schrödinger operator with potential* $Q(x)$ because it solves the (time dependent) Schrödinger equation

$$\frac{1}{2}p^Q_{xx} - Q(x)p^Q = p^Q_t \tag{4.7}$$

with the Dirac delta initial condition $p^Q(0; x, y) = \delta(x - y)$ and appropriate boundary conditions at the endpoints[8]), then we have the following relationship between the two densities

[8] See Jeanblanc et al. (1997) for closely related work on integral functionals of Brownian motion, the Feynman–Kac formula, and related references.

$$p(t;x,y)=\frac{1}{a(y)}\sqrt{\frac{s(x)a(x)}{s(y)a(y)}}\,p^{Q}\big(t;g(x),g(y)\big), \tag{4.8}$$

and for the value function we obtain:

$$\begin{aligned}V(t,x)&=\int_{e_1}^{e_2}p(t;x,y)f(y)\,\mathrm{d}y\\&=\int_{e_1}^{e_2}\sqrt{\frac{s(x)a(x)}{s(y)a(y)}}\,p^{Q}\big(t;g(x),g(y)\big)f(y)\frac{\mathrm{d}y}{a(y)}.\end{aligned} \tag{4.9}$$

Table 1 lists some of the most important analytically tractable cases whose spectral representations can be expressed in terms of classical special functions.[9] The table contains three columns. The first column presents the Schrödinger potential[10] $Q(x)$ on the interval $x\in(e_1,e_2)$[11], the second column lists some of the practically important diffusion processes for which the SL equation reduces to the Schrödinger equation with this potential, and the third column lists financial models that feature these processes. In this section we survey these analytically tractable models. This catalog of analytically tractable models is incomplete (some further specifications of analytically tractable Schrödinger equations can be found in Grosche and Steiner, 1998 and references therein), but it does include some of the most important families of Schrödinger potentials and associated diffusion processes. For each of these Schrödinger potentials $Q(x)$, explicit expressions for the associated resolvent kernels (Green's function) $G_{\alpha}^{Q}(x,y)$ and transition densities $p^{Q}(t;x,y)$ are available. The reader can then immediately obtain the transition density for any diffusion process with the associated SL equation reducible to this normal form (hence, the sate-price density of any asset pricing model depicted by this diffusion process). To price derivative assets with $L^2((e_1,e_2),m)$ payoffs, one then needs to calculate expansion coefficients. If the payoff of interest is not in $L^2((e_1,e_2),m)$, one needs to proceed as described in the Remark in Section 3.3, by either pricing an L^2 payoff related to the one of interest by a parity relationship, such as put-call parity, or by directly inverting the Laplace transform of the integral of the payoff with the resolvent kernel.

[9] Both the *Mathematica* and *Maple* software packages include all special functions appearing in these models as built-in functions.

[10] If $\tilde{Q}(x)=Q(x)+c$, then we clearly have $p^{\tilde{Q}}(t;x,y)=e^{-ct}p^{Q}(t;x,y)$. Thus, we consider potentials up to an additive constant. If the process $\hat{X}$ leads to the Schrödinger potential of the form included in the table plus a constant c, the transition density needs an extra discounting with e^{-ct}.

[11] In the table we provide the maximal interval (e_1,e_2) on which the potential is continuous. For any smaller interval $(e_1',e_2')\subset(e_1,e_2)$, one can also consider a sub-process, imposing boundary conditions at the regular endpoints (in this section we only consider killing boundary conditions; see Linetsky, 2005 for spectral expansions for processes with reflection).

Table 1.
Summary of analytically tractable Schrödinger potentials, diffusion processes, and related financial models.

Schrödinger potential	Diffusion processes	Financial models
Constant potential	Arithmetic Brownian motion	Bachelier
$Q(x) = \text{const}$	$a(x) = \text{const}, b(x) = \text{const}$	
$I = (-\infty, \infty)$	Geometric Brownian motion	Black–Scholes
	$a(x) = \sigma x, b(x) = (r - q)x$	
	Bounded diffusion on (a, b)	Bounded FX, IR
	$a(x) = \sigma(b - a)^{-1}(x - a)(b - x), b(x) = 0$	
	Quadratic volatility	Volatility smile
	$a(x) = ax^2 + bx + c, b(x) = 0$	
Harmonic oscillator	Ornstein–Uhlenbeck process	Vasicek model
$Q(x) = ax^2 + bx$	$a(x) = \sigma, b(x) = \kappa(\theta - x)$	Quadratic model
$I = (-\infty, \infty)$		
Radial harmonic oscillator	Radial OU process	
$Q(x) = ax^2 + bx^{-2}$	or Bessel process with linear drift	
$I = (0, \infty)$	$a(x) = 1, b(x) = (\nu + 1/2)x^{-1} + \mu x$	
	Related diffusions	
	$a(x) = \sigma x^{1/2}, b(x) = \kappa(\theta - x)$	CIR model
	$a(x) = \sigma x^{1+\beta}, b(x) = (r - q)x$	CEV model
	$a(x) = \sigma x^{1+\beta}, b(x) = (r - q + b + c\sigma^2 x^{2\beta})x$	JDCEV model
	$a(x) = \sigma x^{3/2}, b(x) = \kappa(\theta - x)x$	3/2 model
Coulomb potential	Bessel process with constant drift	
$Q(x) = ax^{-2} + bx^{-1}$	$a(x) = 1, b(x) = (\nu + 1/2)x^{-1} + \mu$	
$I = (0, \infty)$	Related diffusions	
	$a(x) = \sigma x^2, b(x) = \kappa(\theta - x)x^2$	Non-affine models
	$a(x) = \sigma x^{3/2}, b(x) = \kappa(\theta - x^{1/2})x^{3/2}$	Black–Scholes
	$a(x) = \sigma x, b(x) = (r - q + b + \alpha \ln^{-1}(x/K))x$	with default
Morse potential	GBM with affine drift	Asian options
$Q(x) = ae^{-2\gamma x} + be^{-\gamma x}$	$a(x) = \sigma x, b(x) = Ax + B$	Black–Scholes
$I = (-\infty, \infty)$	Related models	with default
	$a(x) = 2x, b(x) = 2(\nu + 1)x + 1$	Cash dividends
	$a(x) = \sigma x, b(x) = (r - q + b + \alpha x^{-p})x$	Merton's IR model
	$a(x) = \sigma x, b(x) = rx - \delta$	GARCH diffusion
	$a(x) = \sigma x, b(x) = \kappa(\theta - x)x$	Spot energy model
	$a(x) = \xi x, b(x) = \kappa(\theta - x)$	Brennan–Schwartz
Pöschl–Teller potential	Jacobi diffusion	FX target zones
$Q(x) = \frac{a}{\cos^2(\gamma x)} + \frac{b \sin(\gamma x)}{\cos^2(\gamma x)}$	$a(x) = A\sqrt{1 - x^2}, b(x) = \kappa(\gamma - x)$	
$I = (0, 2\pi/\gamma)$		
Hyperbolic barrier,	Hypergeometric diffusion	Volatility skew
Modified Pöschl–Teller	$a(x) = \sqrt{Ax^2 + Bx + C}, b(x) = \mu x$	
$Q(x) = \frac{a}{\cosh^2(\gamma x)} + \frac{b \sinh(\gamma x)}{\cosh^2(\gamma x)}$		
$I = (-\infty, \infty)$		

To summarize, when faced with a diffusion process on some interval with given volatility, drift, and killing (discount) rate, do the Liouville transformation, reduce the problem to Brownian motion on a transformed interval and killed at the rate (3.32)–(3.33) that is constructed from the infinitesimal parameters of the original diffusion, and look it up in the table of analytically tractable Schrödinger potentials. If it is analytically tractable, you will immediately obtain the transition density for the original diffusion from this density by undoing the Liouville transformation via Eq. (4.8).

4.2 Constant potential, Brownian motion, and related financial models

We first consider Brownian motion on the real line without drift and without killing. The SL problem is

$$-\frac{1}{2}u''(x) = \lambda u(x), \quad x \in (-\infty, \infty)$$

with both boundaries $-\infty$ and $+\infty$ natural and oscillatory with cutoff $\Lambda = 0$. We are in the Spectral Category III with $\Lambda_1 = \Lambda_2 = 0$ with non-simple purely absolutely continuous spectrum $[0, \infty)$. The speed and scale densities are constant, $m(x) = 2$, and $s(x) = 1$. The solutions

$$\psi_\alpha(x) = e^{\sqrt{2\alpha}x}, \qquad \phi_\alpha(x) = e^{-\sqrt{2\alpha}x}$$

are exponentials in this case, and the Green's function with respect to the speed measure is:

$$G_\alpha(x, y) = \frac{e^{-\sqrt{2\alpha}|x-y|}}{2\sqrt{2\alpha}}.$$

Regarded as a function of the complex variable $\alpha \in \mathbb{C}$, it has a branching point at zero, and we place a branch cut from $\alpha = 0$ to $\alpha \to -\infty$ along the negative real axes. It is convenient to parameterize the branch cut $\{\alpha = -\rho^2/2, \rho \geqslant 0\}$. The jump across the cut is:

$$G_{\frac{1}{2}\rho^2 e^{i\pi}}(x, y) - G_{\frac{1}{2}\rho^2 e^{-i\pi}}(x, y) = -\frac{i}{\rho}\cos\big(\rho(x-y)\big).$$

The Bromwich Laplace transform inversion (3.11) can now be accomplished by applying the Cauchy Residue Theorem (see Titchmarsh, 1962 for details). Since in this case the Green's function does not have any poles, the Bromwich integral (3.11) reduces to:

$$\begin{aligned} p_m(t; x, y) &= -\frac{1}{2\pi i}\int_0^\infty e^{-\frac{\rho^2 t}{2}}\big(G_{\frac{1}{2}\rho^2 e^{i\pi}}(x, y) - G_{\frac{1}{2}\rho^2 e^{-i\pi}}(x, y)\big)\rho\,\mathrm{d}\rho \\ &= \frac{1}{2\pi}\int_0^\infty e^{-\frac{\rho^2 t}{2}}\cos\big(\rho(x-y)\big)\,\mathrm{d}\rho \end{aligned}$$

$$= \frac{1}{2\pi} \int_0^\infty e^{-\frac{\rho^2 t}{2}} \big(\sin(\rho x)\sin(\rho y) + \cos(\rho x)\cos(\rho y)\big)\, \mathrm{d}\rho,$$

yielding the spectral representation of the Brownian motion transition density in the form (3.12). In this case, the integral with respect to the spectral parameter can be calculated in closed form to yield the familiar Gaussian density:

$$p(t; x, y) = 2p_m(t; x, y) = \frac{1}{\pi} \int_0^\infty e^{-\frac{\rho^2 t}{2}} \cos\big(\rho(x-y)\big)\, \mathrm{d}\rho$$

$$= \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}}.$$

Furthermore, for any diffusion process with volatility $a(x)$, drift $b(x)$, and killing rate $k(x)$ such that the Liouville transformation reduces it to Brownian motion with $a(x) = 1$, $b = 0$, and constant potential (killing rate) $Q = const$, we immediately obtain the transition density via Eq. (4.8) (note that a constant c added to the potential simply shifts the spectrum by c, i.e., discounts the transition density with e^{-ct}). For Brownian motion with drift $\mu \in \mathbb{R}$, the potential is constant, $Q = \mu^2/2$, and Eq. (4.8) reduces to the familiar Cameron–Martin–Girsanov formula:

$$p^\mu(t; x, y) = e^{\mu(y-x) - \mu^2 t/2} p^0(t; x, y),$$

where p^μ is the transition density of Brownian motion with drift μ. Brownian motion was first employed to model stock prices by Louis Bachelier in his celebrated thesis (see Schachermayer and Teichmann, 2006 for an illuminating discussion). The problem with Bachelier's model is that Brownian motion lives on the whole real line, while stock prices should stay non-negative.

To insure positive prices, an alternative is to take an exponential and consider geometric Brownian motion as in the Black–Scholes (1973) and Merton (1973) model. For geometric Brownian motion of the Black–Scholes–Merton model,

$$a(x) = \sigma x, \qquad b(x) = (r - q)x,$$

the Liouville transformation

$$y = \sigma^{-1} \ln x, \qquad u(x) = \sigma^{\frac{1}{2}} x^{-\frac{\nu}{\sigma}} v\big(y(x)\big)$$

reduces the SL equation to the Schrödinger equation with constant potential:

$$Q = \frac{\nu^2}{2}, \quad \nu := \frac{r-q}{\sigma} - \frac{\sigma}{2},$$

and the lognormal risk-neutral transition density for the Black–Scholes–Merton model is immediately recovered from (4.8).

Next consider a bounded foreign exchange rate model of Ingersoll (1996) and Rady (1997), where the (forward) foreign exchange rate is assumed to follow the process:

$$\mathrm{d}X_t = \sigma \frac{(X_t - a)(b - X_t)}{(b - a)}\,\mathrm{d}B_t, \quad X_0 = x \in (a, b).$$

Both a and b are inaccessible natural boundaries. The Liouville transformation for this process with quadratic volatility in the bounded interval (a, b) is:

$$y = \frac{1}{\sigma}\ln\left(\frac{x - a}{b - x}\right), \qquad u(x) = \sqrt{\sigma(b - a)^{-1}(x - a)(b - x)}v\big(y(x)\big).$$

The resulting potential function (3.30) turns out to be a constant $Q = \sigma^2/2$. Thus, this process is reducible to Brownian motion, and we immediately obtain its transition density from the one for Brownian motion by inverting the Liouville transformation (which was, in effect, done in Ingersoll, 1996 and Rady, 1997 without explicitly referring to the Liouville transformation). This diffusion has also been applied to interest rate modeling by Rady and Sandmann (1994) and Miltersen et al. (1997) with $a = 0$ and $b = 1$.

One can also consider a diffusion with *quadratic volatility* and no drift,

$$a(x) = ax^2 + bx + c, \qquad b(x) = 0,$$

where one assumes that the quadratic does not have any roots in $(0, \infty)$. If the process does not have roots in $[0, \infty)$, then the process is killed at the origin. If $a > 0$ and $c = 0$, then the process has a root at $x = 0$ and the origin is a natural boundary. These quadratic volatility processes on $(0, \infty)$ have been used to model volatility smiles by Zuhlsdorff (2001) (see also Albanese et al., 2001). However, one must be careful with this process as it is in general a strictly local martingale. Infinity is an entrance boundary for this process, and the global process dynamics is such that it rapidly falls from high positive values. The result is that the process is a strict supermartingale. This is similar to the CEV process with $\beta > 0$. Zuhlsdorff (2001) and Albanese et al. (2001) used this process to value double-barrier options. There is no problem in this case as the process is killed at some upper barrier. But special care must be taken when considering this process on $(0, \infty)$ (one possibility is to regularize the process similar to the Andersen and Andreasen, 2000 regularization for the CEV process with $\beta > 0$). Carr et al. (2003) investigate more general classes of local volatility models reducible to Brownian motion (they consider more general transformations that also depend on time).

So far we have considered Brownian motion on the real line. Brownian motion on the half-line (a, ∞) killed at a, as well as on the finite interval (a, b) killed at both a and b, is treated similarly. In the latter case, the SL problem is

$$-\frac{1}{2}u''(x) = \lambda u(x), \quad x \in (a, b),\ u(a) = u(b) = 0.$$

Both endpoints are regular killing boundaries and, hence, we are in the Spectral Category I. The fundamental solution entire in α can be taken in the form:

$$\psi_\alpha(x) = \frac{\sinh(\sqrt{2\alpha}(x-a))}{\sqrt{2\alpha}}, \qquad \phi_\alpha(x) = \frac{\sinh(\sqrt{2\alpha}(b-x))}{\sqrt{2\alpha}}.$$

The Wronskian

$$w_\alpha = -\frac{\sinh(\sqrt{2\alpha}(b-a))}{\sqrt{2\alpha}}$$

is entire in α with simple positive zeros $\alpha = -\lambda_n$:

$$\lambda_n = \frac{n^2\pi^2}{2(b-a)^2}, \quad n = 1, 2, \ldots. \tag{4.10}$$

At $\alpha = -\lambda_n$ the two solutions become linearly dependent:

$$\phi_{-\lambda_n}(x) = (-1)^{n+1}\psi_{-\lambda_n}(x),$$

the Wronskian derivative at $\alpha = -\lambda_n$ is

$$w'_{-\lambda_n} = (-1)^{n+1}\frac{(b-a)^3}{n^2\pi^2},$$

the eigenfunctions (3.32) normalized in $L^2((a,b), 2dx)$ are given by

$$\varphi_n(x) = \frac{1}{\sqrt{b-a}}\sin\left(n\pi\frac{x-a}{b-a}\right), \tag{4.11}$$

and we arrive at the familiar spectral representation for the transition density of Brownian motion between two killing barriers:

$$\begin{aligned} p(t;x,y) &= 2p_m(t;x,y) \\ &= \frac{2}{b-a}\sum_{n=1}^{\infty} e^{-\lambda_n t}\sin\left(n\pi\frac{x-a}{b-a}\right)\sin\left(n\pi\frac{y-a}{b-a}\right). \end{aligned}$$

We illustrate with the application to the pricing of double-barrier options (see Davydov and Linetsky, 2003 for more details and further references). Assume that under the risk-neutral probability measure the underlying asset price follows a geometric Brownian motion with the initial price $S_0 = x$, constant volatility $\sigma > 0$, constant risk-free rate $r \geqslant 0$ and constant dividend yield $q \geqslant 0$. Consider a double-barrier call option with the strike price K, expiration date T, and two knock-out barriers L and U, $0 < L < K < U$. The knock-out provision renders the option worthless as soon as the underlying price leaves the price range (L, U) (it is assumed that $S_0 \in (L, U)$). The double barrier call payoff is $\mathbf{1}_{\{\mathcal{T}_{(L,U)}>T\}}(S_T - K)^+$, where $\mathcal{T}_{(L,U)} = \inf\{t \geqslant 0\colon S_t \notin (L, U)\}$ is the first exit time from the range (L, U). The Liouville transformation reduces the

SL equation to the Schrödinger equation on the interval $(0, \ln(U/L)/\sigma)$ with the constant potential $Q = r+\nu^2/2$, where $\nu = (r-q-\sigma^2/2)/\sigma$. The eigenvalues and eigenfunctions for the problem with zero potential and interval (a, b) are given by (4.10) and (4.11). The additional constant in the potential simply shifts the eigenvalues up by Q. Inverting the Liouville transformation yields the eigenfunctions and eigenvalues of the original problem:

$$\varphi_n(x) = \frac{\sigma x^{-\frac{\nu}{\sigma}}}{\sqrt{\ln(U/L)}} \sin\left(\frac{\pi n \ln(x/L)}{\ln(U/L)}\right),$$
$$\lambda_n = r + \frac{\nu^2}{2} + \frac{\sigma^2\pi^2 n^2}{2\ln^2(U/L)}, \quad n = 1, 2, \ldots.$$

The call option value function is given by the eigenfunction expansion (3.21):

$$C(t, x) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t}\varphi_n(x).$$

The expansion coefficients are calculated in closed form in this case:

$$c_n = \left((\cdot - K)^+, \varphi_n\right) = \frac{L^{\frac{\nu}{\sigma}}}{\sqrt{\ln(U/L)}}\left[L\psi_n(\nu+\sigma) - K\psi_n(\nu)\right],$$

where

$$\psi_n(a) := \frac{2}{\omega_n^2 + a^2}\left[e^{ak}\left(\omega_n \cos(\omega_n k) - a\sin(\omega_n k)\right) - (-1)^n \omega_n e^{au}\right],$$

and

$$\omega_n := \frac{n\pi}{u}, \quad k := \frac{1}{\sigma}\ln\left(\frac{K}{L}\right), \quad u := \frac{1}{\sigma}\ln\left(\frac{U}{L}\right).$$

A practically important observation is that the eigenvalues increase as n^2 to facilitate fast convergence of the eigenfunction expansion. The longer the time to expiration, the faster the eigenfunction expansion converges. The option delta and gamma can be obtained by differentiating the eigenfunction expansion term-by-term:

$$\Delta(t, x) = C_x(t, x) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t}\varphi_n'(x),$$
$$\Gamma(t, x) = C_{xx}(t, x) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t}\varphi_n''(x).$$

To illustrate, Table 2 shows convergence of the spectral expansions for double-barrier call option prices, deltas and gammas with $T = 1/12$ and one year to expiration and $S_0 = K = 100$, $L = 80$, $U = 130$, $r = 0.1$, $q = 0$,

Table 2.
Convergence of spectral expansions for double-barrier call prices, deltas and gammas with $T = 1/12$ (one month) and one year.

N	Price	Delta	Gamma
	Double-Barrier Call $T = 1/12$ *Years*		
1	7.65069	−0.020999	−0.028245
2	5.78096	0.932450	0.032490
3	2.60902	0.720920	0.143401
4	3.11276	0.480955	0.102269
5	3.35713	0.536457	0.078607
6	3.29822	0.561105	0.088242
7	3.28703	0.554994	0.090335
8	3.28941	0.554203	0.089673
9	3.28958	0.554412	0.089623
10	3.28953	0.554422	0.089643
11	3.28953	0.554419	0.089643
	Double-Barrier Call $T = 1$ *Years*		
1	2.03207	−0.0055774	−0.0075021
2	2.01848	0.0013543	−0.0070606
3	2.01842	0.0013505	−0.0070586
4	2.01842	0.0013505	−0.0070586

The number in the N column indicates how many terms are included in the truncated expansion (e.g., $N = 2$ means that the first two terms with $n = 1$ and $n = 2$ are included in the eigenfunction expansions). Parameters: $S_0 = K = 100$, $L = 80$, $U = 130$, $r = 0.1$, $q = 0$, $\sigma = 0.25$.

$\sigma = 0.25$. For one-year options, the first three terms are enough to converge to five significant digits. For one-month options ($T = 1/12$), the first ten terms are required to achieve this level of accuracy. There is no loss of accuracy in computing delta and gamma. This is in contrast to numerical option pricing methods such as lattices, numerical PDE schemes and simulation. This basic example illustrates the two key characteristics of the spectral method: *numerical convergence improves with increasing maturity, and the Greeks* (*delta and gamma*) *are obtained at no additional computational cost by direct differentiation of the expansion*.

4.3 Harmonic oscillator, Ornstein–Uhlenbeck process, and related models

4.3.1 Harmonic oscillator potential

Our next example is the quadratic potential:

$$Q(x) = ax^2 + bx + \frac{b^2}{4a} = \frac{\kappa^2}{2}(x + \beta)^2,$$

$$\kappa > 0,\ a = \frac{\kappa^2}{2},\ b \in \mathbb{R},\ \beta = \frac{b}{2a},\ x \in \mathbb{R}$$

known as the *harmonic oscillator* potential in quantum mechanics (e.g., Morse and Feshbach, 1953, p. 1641). Consider Brownian motion killed at the rate $Q(x)$. Since $Q(x) \to +\infty$ as $x \to \pm\infty$, both $-\infty$ and $+\infty$ are non-oscillatory natural boundaries and the spectrum is purely discrete. This is in contrast with standard Brownian motion that has oscillatory natural boundaries and purely absolutely continuous spectrum. Intuitively, the killing rate increases fast as the process wanders away from β, rapidly increasing the killing probability. This "localization" relative to the free Brownian motion results in the discrete spectrum, similar to the discrete spectrum of Brownian motion killed at the endpoints of a finite interval.

For simplicity set $\beta = 0$. The fundamental solutions are expressed in terms of the Weber–Hermite parabolic cylinder function:

$$\psi_\alpha(x) = D_{-\frac{\alpha}{\kappa}-\frac{1}{2}}(-x\sqrt{2\kappa}), \qquad \phi_\alpha(x) = D_{-\frac{\alpha}{\kappa}-\frac{1}{2}}(x\sqrt{2\kappa})$$

with the Wronskian

$$w_\alpha = \frac{2\sqrt{\pi\kappa}}{\Gamma(\alpha/\kappa + 1/2)}.$$

The zeros of the Wronskian are (for notational convenience here we label the eigenvalues and eigenfunctions starting from $n = 0$):

$$\alpha = -\lambda_n, \quad \lambda_n = \kappa(n + 1/2), \quad n = 0, 1, \ldots.$$

At an eigenvalue, $\alpha = -\lambda_n$, the Weber–Hermite functions D_ν degenerate into Hermite polynomials H_n, the eigenfunctions are expressed in terms of the latter, and the spectral representation of the transition density of Brownian motion killed at the quadratic rate is:

$$\begin{aligned} p(t; x, y) &= \frac{\partial}{\partial y}\mathbb{E}_x^W\left[e^{-\frac{\kappa^2}{2}\int_0^t W_u^2\,\mathrm{d}u}\mathbf{1}_{\{W_t \leqslant y\}}\right] \\ &= \sum_{n=0}^{\infty} e^{-\kappa(n+1/2)t}\frac{1}{2^n n!}\left(\frac{\kappa}{\pi}\right)^{\frac{1}{2}} e^{-\frac{\kappa}{2}(x^2+y^2)} H_n\big(x\sqrt{\kappa}\big)H_n\big(y\sqrt{\kappa}\big). \end{aligned} \tag{4.12}$$

Applying Mehler's formula (Eq. (22) in Erdelyi, 1953, p. 194)

$$\sum_{n=0}^{\infty}\frac{(z/2)^n}{n!}H_n(x)H_n(y) = (1-z^2)^{-\frac{1}{2}}\exp\left\{\frac{2xyz - (x^2+y^2)z^2}{1-z^2}\right\},$$

the spectral representation reduces to the familiar expression (e.g., Borodin and Salminen, 2002, p. 168, Eq. (1.9.7))

$$p(t; x, y) = \sqrt{\frac{\kappa}{2\pi\sinh(\kappa t)}}\exp\left\{-\frac{(x^2+y^2)\kappa\cosh(\kappa t) - 2xy\kappa}{2\sinh(\kappa t)}\right\}. \tag{4.13}$$

4.3.2 Ornstein–Uhlenbeck process

The mean-reverting Ornstein–Uhlenbeck (OU) process has the infinitesimal parameters

$$a(x) = \sigma, \qquad b(x) = \kappa(\theta - x),$$

where $\kappa > 0$, θ, and $\sigma > 0$ are the rate of mean reversion, the long-run level, and volatility, respectively (in finance applications, typically $\theta > 0$). The Liouville transformation (3.28) with $x_0 = \theta$ reduces the SL equation (3.22) to the Schrödinger equation with quadratic potential:

$$Q(x) = \frac{1}{2}\kappa^2 x^2 - \frac{1}{2}\kappa.$$

The spectral representation of the OU transition density immediately follows from (4.12) by Eq. (4.8). This spectral representation in terms of Hermite polynomials is well known (e.g., Wong, 1964; Karlin and Taylor, 1981, p. 333; Schoutens, 2000). The eigenvalues are $\lambda_n = \kappa n$. The principal eigenvalue is zero, $\lambda_0 = 0$. Hence, the first term of the eigenfunction expansion of the transition density gives the Gaussian stationary density of the OU process:

$$\pi(x) = \sqrt{\frac{\kappa}{\pi\sigma^2}} e^{-\frac{\kappa(x-\theta)^2}{\sigma^2}}.$$

The eigenfunction expansion can be summed up, resulting in the familiar Gaussian density of the OU process (obtained from (4.12) by (4.8)).

If the OU process is considered on some interval $I \subset \mathbb{R}$ with killing or reflection at finite boundaries, then the eigenfunctions are expressed in terms of the Weber-Hermite functions, rather than reduce to Hermite polynomials (see Linetsky, 2004e and Alili et al., 2005 for the case of killing and associated results on hitting times of the OU process, and Linetsky, 2005 for the case of reflection).

4.3.3 Vasicek, quadratic, and Black's interest rate models

If we take the short rate $r(x) = x$, we obtain the state-price density of the Vasicek (1977) interest rate model. In the Vasicek model the short rate r can get negative. Consequently, the pricing semigroup is *not*, in general, a contraction semigroup. This leads to serious economic problems (see Gorovoi and Linetsky, 2004). If we take

$$r(x) = ax^2 + bx + c, \quad \text{with } c \geqslant b^2/4a,$$

we obtain the state-price density of a non-negative interest rate model where the short rate is a quadratic of the OU process (this quadratic model was proposed by Beaglehole and Tenney, 1992; see also Jamshidian, 1996; Leippold and Wu, 2002, and Chen et al., 2004). For further details and references see Gorovoi and Linetsky (2004), where an alternative non-negative interest rate model with $r(x) = x^+ \equiv \max\{x, 0\}$, the so-called *Black's model of interest*

rates as options (Black, 1995), is solved analytically via the spectral expansion method and calibrated to Japanese Government Bond data. This model can handle low interest rate regimes and has been recently adopted by the Bank of Japan (Bank of Japan, "Financial Markets Report: Developments during the First Half of 2005," p. 7).

4.4 Radial harmonic oscillator, radial OU and Bessel processes, and related models

4.4.1 Radial OU process and Bessel processes

Next we consider a d-dimensional Ornstein–Uhlenbeck process $\{X_t, t \geqslant 0\}$ with the infinitesimal generator

$$\frac{1}{2}\Delta - \mu x \cdot \nabla,$$

where Δ is the standard d-dimensional Laplacian (infinitesimal generator of d-dimensional Brownian motion) and $\mu \in \mathbb{R}$. The radial part of this process, $\{R_t = |X_t|, t \geqslant 0\}$ (here $|x| = \sqrt{x \cdot x}$ is the Euclidean norm), turns out to be a one-dimensional diffusion process with volatility and drift

$$a(x) = 1, \qquad b(x) = \frac{\nu + 1/2}{x} - \mu x, \quad x \in (0, \infty),$$

where $\nu = d/2 - 1$. In fact, the process can be considered for all $\nu \in \mathbb{R}$. When $\mu \neq 0$, it is called the *radial Ornstein–Uhlenbeck process* (Shiga and Watanabe, 1973; Pitman and Yor, 1982; Goeing-Jaeschke and Yor, 2003; Borodin and Salminen, 2002). When $\mu = 0$, it is the *Bessel process of index* ν (Revuz and Yor, 1999; Borodin and Salminen, 2002). The boundary classification at the origin is independent of the drift parameter μ. For all $\mu \in \mathbb{R}$, 0 is entrance for $\nu \geqslant 0$, regular for $-1 < \nu < 0$ (both killing and reflecting boundary conditions arise in finance applications) and exit for $\nu \leqslant -1$. For all ν and μ infinity is a natural boundary.

The Liouville transformation with $x_0 = 0$ reduces the radial OU SL equation to the Schrödinger equation with the *radial harmonic oscillator potential* (Morse and Feshbach, 1953, p. 1661; Grosche and Steiner, 1998):

$$Q(x) = \frac{1}{2}\big(\nu^2 - 1/4\big)x^{-2} + \frac{1}{2}\mu^2 x^2 - \mu(\nu + 1), \quad x \in (0, \infty).$$

For $\mu \neq 0$ and all $\nu \in \mathbb{R}$, $Q(x) \to \infty$ as $x \to +\infty$, and $+\infty$ is non-oscillatory natural. Thus, the spectrum is purely discrete (Spectral Category I). For $\mu = 0$, the process reduces to the Bessel process of index ν, $Q(x) \to 0$ as $x \to +\infty$, $+\infty$ is oscillatory with cutoff $\Lambda = 0$. Zero is not an eigenvalue and the Bessel process has a simple and purely absolutely continuous spectrum $\sigma(-\mathcal{G}) = \sigma_{ac}(-\mathcal{G}) = [0, \infty)$ (Spectral Category II).

A property of this process important for practical calculations is that if we kill the process at the rate $k(x) = \frac{\beta^2}{2}x^{-2} + \frac{\gamma^2}{2}x^2$, the process $\hat{R}$ with killing

is just as analytically tractable as the process R with $\beta = \gamma = 0$. Indeed, both processes lead to the same potential function:

$$Q(x) = \frac{1}{2}\big(\nu^2 + \beta^2 - 1/4\big)x^{-2} + \frac{1}{2}\big(\mu^2 + \gamma^2\big)x^2 - \mu(\nu + 1), \quad x \in (0, \infty). \tag{4.14}$$

Only the coefficients in front of x^{-2} and x^2 change.

For $\mu \neq 0$, the solutions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ are expressed in terms of the Whittaker functions (or, equivalently, in terms of the Kummer and Tricomi confluent hypergeometric functions). Explicit expressions can be found on pp. 139–140 in Borodin and Salminen (2002). At an eigenvalue, $\lambda = \lambda_n$, the Whittaker functions degenerate into the generalized Laguerre polynomials, and the eigenfunctions are expressed in terms of the latter. If the process is considered on an interval with killing or reflection at finite boundaries, then the eigenfunctions are expressed in terms of the Whittaker functions. For $\mu = 0$, the solutions $\psi_\alpha(x)$ and $\phi_\alpha(x)$ are expressed in terms of the modified Bessel function (explicit expressions can be found on p. 133 in Borodin and Salminen, 2002), and the spectral representation with absolutely continuous spectrum (3.33) has an integral form (e.g., Karlin and Taylor, 1981, p. 338 for $\nu = d/2 - 1$).

4.4.2 The CIR model

Let $\{R_t, t \geqslant 0\}$ be a radial OU process with $\nu \geqslant 0$ and $\mu > 0$. For $\sigma > 0$, the squared process $\{X_t = \frac{\sigma^2}{4}R_t^2, t \geqslant 0\}$ is a Feller's (1951) square-root diffusion on $(0, \infty)$ with the infinitesimal parameters (see also Wong, 1964):

$$a(x) = \sigma\sqrt{x}, \qquad b(x) = \kappa(\theta - x), \quad \text{where } \theta := \frac{\sigma^2}{4\mu}(\nu + 1) > 0,\ \kappa := 2\mu > 0.$$

Here $\kappa > 0$, $\theta > 0$, and $\sigma > 0$ are the rate of mean reversion, the long-run level, and volatility, respectively. This diffusion is known in finance as the CIR process (Cox et al., 1985) and is widely used as a model of interest rates (the CIR term structure model), stochastic volatility (Heston, 1993), and credit spreads (Duffie and Singleton, 2003).

The scale and speed densities of the CIR diffusion are:

$$s(x) = x^{-\beta}e^{ax}, \qquad m(x) = \frac{2}{\sigma^2}x^{\beta-1}e^{-ax}, \quad \text{where } a := \frac{2\kappa}{\sigma^2},\ \beta := \frac{2\kappa\theta}{\sigma^2}.$$

For $\beta \geqslant 1$, zero is an inaccessible entrance boundary, and $+\infty$ is a non-attracting natural boundary. The fundamental solutions are:

$$\psi_\alpha(x) = M(\alpha/\kappa, \beta, ax), \qquad \phi_\alpha(x) = U(\alpha/\kappa, \beta, ay),$$

where $M(a,b,z)$ and $U(a,b,z)$ are the Kummer and Tricomi confluent hypergeometric functions (see Slater, 1960 and Buchholz, 1969). The Wronskian is ($\Gamma(z)$ is the Gamma function):

$$w_\alpha = \frac{\Gamma(\beta)}{\Gamma(\alpha/\kappa)} a^{-\beta+1}.$$

The zeros of the Wronskian are $\alpha = -\lambda_n$, $\lambda_n = \kappa n$, $n = 0, 1, \ldots$. The spectrum is purely discrete with the eigenvalues $\lambda_n = \kappa n$ (for notational convenience here we label the eigenvalues starting from zero). At an eigenvalue, $\alpha = -\kappa n$, the confluent hypergeometric functions M and U become linearly dependent and degenerate into the generalized Laguerre polynomials. The normalized eigenfunctions (3.32) are:

$$\lambda_n = \kappa n, \qquad \varphi_n(x) = \sqrt{\frac{n!\kappa}{\Gamma(\beta+n)}} a^{\frac{\beta-1}{2}} L_n^{(\beta-1)}(ax), \quad n = 0, 1, \ldots,$$

where $L_n^{(\alpha)}(x)$ are the generalized Laguerre polynomials. Note that the principal eigenvalue is zero, $\lambda_0 = 0$. Hence, the first term of the eigenfunction expansion of the transition density gives the stationary density of the CIR process:

$$\pi(x) = \frac{\kappa a^{\beta-1}}{\Gamma(\beta)} m(x) = \frac{a^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-ax},$$

which is a gamma density. Note that in this case the speed measure is a finite measure and, hence, the stationary density exists and is equal to the speed density normalized to integrate to one.

The symmetric transition density has the eigenfunction expansion (3.20) (here we count the eigenvalues starting from $n = 0$). Applying the Hille–Hardy formula (Eq. (20) in Erdelyi, 1953, p. 189; $I_\alpha(x)$ is the modified Bessel function of order α)

$$\sum_{n=0}^{\infty} \frac{z^n n!}{\Gamma(n+\alpha+1)} L_n^{(\alpha)}(x) L_n^{(\alpha)}(y) = (1-z)^{-1} \exp\left\{-z\frac{x+y}{1-z}\right\} (xyz)^{-\frac{\alpha}{2}} I_\alpha\left(\frac{2\sqrt{xyz}}{1-z}\right),$$

the spectral representation for the symmetric CIR transition density with respect to the speed measure m reduces to the closed-form expression in terms of the modified Bessel function:

$$p_m(t;x,y) = \frac{\kappa}{1-e^{-\kappa t}} (xye^{-\kappa t})^{\frac{1-\beta}{2}} \exp\left\{-\frac{\alpha(x+y)}{e^{\kappa t}-1}\right\} \times I_{\beta-1}\left(\frac{2\alpha\sqrt{xye^{-\kappa t}}}{1-e^{-\kappa t}}\right).$$

The transition density with respect to the Lebesgue measure is $p(t; x, y) = p_m(t; x, y)m(y)$.

The state-price density of the CIR term structure model can be interpreted as the transition density for the CIR process killed at the rate $k(x) = r(x) = x$. This is equivalent to killing the Bessel process with linear drift at the rate $k(x) = \frac{\sigma^2}{4}x^2$. From Eq. (4.14) we see that adding the killing only changes the coefficient in front of the x^2 term in the corresponding Schrödinger potential, and the spectral representation for the CIR term structure model follows immediately. A detailed spectral analysis of the CIR model can be found in Davydov and Linetsky (2003, Section 4) and Gorovoi and Linetsky (2004, Section 5). Applications to modeling mortgages with prepayment can be found in Gorovoi and Linetsky (2006).

4.4.3 *The 3/2 model*

Let $\{R_t, t \geqslant 0\}$ be a radial OU process with $\nu > 1$ and $\mu > 0$. For $\sigma > 0$, the *reciprocal* squared process $\{X_t := 4\sigma^{-2}R_t^{-2}, t \geqslant 0\}$ is a diffusion on $(0, \infty)$ with infinitesimal parameters

$$a(x) = \sigma x^{3/2}, \qquad b(x) = \kappa(\theta - x)x,$$
$$\kappa := \frac{\sigma^2}{2}(\nu - 1) > 0, \ \theta := \frac{4\mu}{\sigma^2(\nu - 1)} > 0.$$

This diffusion with non-linear drift and infinitesimal variance $\sigma^2 x^3$ is the reciprocal of the square-root CIR process. This process was proposed by Cox et al. (1985, p. 402, Eq. (50)) as a model for the inflation rate in their three-factor inflation model. They were able to solve the three-factor valuation PDE for the real value of a nominal bond (their Eqs. (53–54)). More recently this diffusion appeared in Lewis (1998, 2000), Heston (1997), Ahn and Gao (1999) in different contexts. Heston (1997) and Lewis (2000) apply this process in the context of stochastic volatility models. The latter reference provides a detailed study of a stochastic volatility model where the instantaneous asset price variance follows this process. Lewis (1998) and Ahn and Gao (1999) propose this process as a model for the nominal short rate. They show that the 3/2 model is more empirically plausible than the square-root model. The accumulated empirical evidence estimating short rate models with the diffusion parameter $\sim r^\gamma$ suggests that empirically $\gamma > 1$, thus contradicting the square-root specification. Furthermore, recent empirical studies suggest that the short rate drift is substantially non-linear in the short rate. Lewis (1998) (see also Ahn and Gao, 1999) obtains an analytical solution for the zero-coupon bond price by directly solving the PDE. This solution is, in fact, contained in a more general solution given in Eq. (54) of Cox et al. (1985) for their three-factor inflation model. The spectral analysis of the 3/2 model is given in Linetsky (2004a).

4.4.4 The CEV model

Let $\{X_t, t \geqslant 0\}$ be a radial OU process with index $\nu < 0$ and drift parameter $m \in \mathbb{R}$. If $\nu \in (-1, 0)$, the origin is regular and we send the process to the cemetery state Δ at the first hitting time of zero, T_0. If $\nu \leqslant -1$, the origin is exit. Thus, in both cases the lifetime of the process is $\zeta = T_0$. If $m = 0$, then X is the Bessel process of index $\nu < 0$ and killed at the origin. For some $\delta > 0$ define a new process $\{S_t, t \geqslant 0\}$ as follows:

$$S_t := \begin{cases} (\frac{\delta}{2|\nu|} X_t)^{-2\nu}, & 0 \leqslant t < \zeta, \\ \Delta, & t \geqslant \zeta. \end{cases}$$

This process is a diffusion on $(0, \infty)$ with infinitesimal parameters

$$a(x) = \sigma(x)x = \delta x^{1+\beta}, \qquad b(x) = \mu x,$$
$$\text{where } \beta := \frac{1}{2\nu} < 0, \ \mu := 2\nu m \in \mathbb{R},$$

and killing (default) at the first hitting time of zero. This is a *constant elasticity of variance* (CEV) model of Cox (1975) (see also Schroder, 1989; Delbaen and Shirakawa, 2002a, 2002b; Andersen and Andreasen, 2000; Davydov and Linetsky, 2001, 2003; Linetsky, 2004c and references therein).

The CEV specification nests the absolute diffusion ($\beta = -1$) and square-root ($\beta = -1/2$) models of Cox and Ross (1976) as particular cases. For $\beta < 0$, the local volatility $\sigma(x) = \delta x^{\beta}$ is a decreasing function of the asset price. We have two model parameters β and δ; β is the elasticity of the local volatility function and δ is the scale parameter. For $\beta < 0$, $+\infty$ is a natural boundary; attracting for $\mu > 0$ and non-attracting for $\mu \leqslant 0$ (the risk-neutral drift is $\mu = r - q$, where $r \geqslant 0$ and $q \geqslant 0$ are the constant risk-free rate and dividend yield, respectively). For $-1/2 \leqslant \beta < 0$, the origin is an exit boundary. For $-\infty < \beta < -1/2$, the origin is a regular boundary and is specified as killing, by sending the stock process to the default state Δ. In this paper we focus on the CEV process with $\beta < 0$. This process is used to model the volatility skew in the equity options market. The CEV process can also be considered for $\beta > 0$. In this case when $\mu = 0$ the process is a strict local martingale (similar situation to the quadratic volatility process discussed above). We refer to Davydov and Linetsky (2001) for further discussion and references. Davydov and Linetsky (2003) and Linetsky (2004c) provide detailed spectral analysis of the CEV process and obtain analytical solutions for barrier and lookback options in terms of spectral expansions. Here we will consider the case $\mu \geqslant 0$ ($r \geqslant q$). The negative drift case $\mu < 0$ is treated similarly.

The scale and speed densities of the CEV process are

$$s(x) = e^{-ax^{-2\beta}}, \qquad m(x) = 2\delta^{-2}x^{-2\beta-2}e^{ax^{-2\beta}}, \quad \text{where } a := \frac{\mu}{\delta^2|\beta|}.$$

For $\mu > 0$, the fundamental solutions and their Wronskian are:

$$\psi_\alpha(x) = e^{-\frac{a}{2}x^{-2\beta}} x^{\beta+\frac{1}{2}} M_{k(\alpha),\frac{\nu}{2}}\left(ax^{-2\beta}\right),$$

$$\phi_\alpha(x) = e^{-\frac{a}{2}x^{-2\beta}} x^{\beta+\frac{1}{2}} W_{k(\alpha),\frac{\nu}{2}}\big(ax^{-2\beta}\big),$$
$$w_\alpha = \frac{2\mu\Gamma(\nu+1)}{\delta^2\Gamma(\alpha/\gamma+1)},$$
$$\text{where } \nu := \frac{1}{2|\beta|},\ \gamma := 2\mu|\beta|,\ k(\alpha) := \frac{\nu-1}{2} - \frac{\alpha}{\gamma},$$

and $M_{k,m}(x)$ and $W_{k,m}(x)$ are the Whittaker functions (which are related to the confluent hypergeometric functions; see Slater, 1960 and Buchholz, 1969). For notational convenience in what follows, here we re-define $\nu = 1/(2|\beta|)$ as the *absolute value* of the index ν of the radial OU process X we started with. The spectrum is purely discrete and the symmetric transition density admits an eigenfunction expansion (3.20) with the eigenvalues and normalized eigenfunctions

$$\lambda_n = \gamma(n+1),$$
$$\varphi_n(x) = a^{\nu/2}\sqrt{\frac{n!\mu}{\Gamma(\nu+n+1)}}\, x e^{-ax^{-2\beta}} L_n^{(\nu)}\big(ax^{-2\beta}\big), \quad n = 0, 1, \ldots,$$

where $L_n^{(\nu)}(x)$ are the generalized Laguerre polynomials. Applying the Hille–Hardy formula as we did for the CIR process, we can show that the eigenfunction expansion collapses to the expression with the modified Bessel function:

$$p_m(t; x, y) = \frac{\mu(xy)^{1/2}}{e^{\gamma t}-1} \exp\left\{-\frac{a(x^{-2\beta}+y^{-2\beta})}{1-e^{-\gamma t}} + \frac{\mu t}{2}\right\} \times I_\nu\left(\frac{2a(xy)^{-\beta}e^{-\gamma t/2}}{1-e^{-\gamma t}}\right).$$

Now consider the driftless case with $\mu = 0$ and $\beta < 0$. This case is important for pricing options on futures (recall that futures prices have zero drift under the risk-neutral measure). This case is related to the Bessel processes with negative index and killing at the origin. The fundamental solutions and their Wronskian are (here $\nu := 1/(2|\beta|)$):

$$\psi_\alpha(x) = \sqrt{x} I_\nu\big(Ax^{-\beta}\sqrt{2\alpha}\big), \qquad \phi_\alpha(x) = \sqrt{x} K_\nu\big(Ax^{-\beta}\sqrt{2\alpha}\big),$$
$$w_\alpha = |\beta|,\ A := \frac{1}{\delta|\beta|},$$

where $I_\nu(x)$ and $K_\nu(x)$ are the modified Bessel functions.

In the driftless case the spectrum is purely absolutely continuous and the symmetric transition density admits an absolutely continuous spectral representation (3.33) (with $\Lambda = 0$, $d\rho_{ac}(\lambda) = |\beta|^{-1}\,\mathrm{d}\lambda$, and no additional eigenvalues in this case):

$$p_m(t; x, y) = \int_0^\infty e^{-\lambda t}\psi_{-\lambda}(x)\psi_{-\lambda}(y)|\beta|^{-1}\,\mathrm{d}\lambda,$$

$$\text{where } \psi_{-\lambda}(x) = \sqrt{x} J_\nu\big(Ax^{-\beta}\sqrt{2\lambda}\big),$$

where J_ν is the Bessel function of order ν (recall that $J_\nu(z) = I_\nu(iz)$).

The continuous spectral representation of the driftless CEV density has the form of the Laplace transform. This Laplace transform can be explicitly calculated using the following integral identity (Erdelyi, 1953, vol. II, p. 53) (for all $\nu > -1$ and $t > 0$):

$$\frac{1}{t}\exp\left\{-\frac{u^2+v^2}{2t}\right\} I_\nu\left(\frac{uv}{t}\right) = \int_0^\infty e^{-\lambda t} J_\nu\big(u\sqrt{2\lambda}\big) J_\nu\big(v\sqrt{2\lambda}\big)\,\mathrm{d}\lambda,$$

yielding the following explicit expression:

$$p_m(t; x, y) = \frac{(xy)^{1/2}}{2|\beta|t}\exp\left\{-\frac{A^2}{2t}\big(x^{-2\beta}+y^{-2\beta}\big)\right\} I_\nu\left(\frac{A^2}{t}(xy)^{-\beta}\right).$$

Since zero is a killing boundary, the transition density is defective, and we obtain the survival probability:

$$P_t\big(x, (0,\infty)\big) = \int_0^\infty p(t; x, y)\,\mathrm{d}y = \frac{\gamma(\nu, \eta(t,x))}{\Gamma(\nu)} < 1,$$

where

$$\eta(t,x) := \begin{cases} ax^{-2\beta}/(1-e^{-\gamma t}), & \mu > 0, \\ Ax^{-2\beta}/(2t), & \mu = 0 \end{cases}$$

and $\gamma(\nu, z)$ is the incomplete Gamma function, defined $\gamma(\nu, z) = \int_0^z t^{\nu-1}e^{-t}\,\mathrm{d}t$. The probability of killing by hitting zero (default) up to time t is:

$$P_t\big(x, \{\Delta\}\big) = 1 - P_t\big(x, (0,\infty)\big) = \frac{\Gamma(\nu, \eta(t,x))}{\Gamma(\nu)},$$

where $\Gamma(\nu, z)$ is the complementary incomplete Gamma function $\Gamma(\nu, z) = \Gamma(\nu) - \gamma(\nu, z) = \int_z^\infty t^{\nu-1}e^{-t}\,\mathrm{d}t$ (Abramowitz and Stegun, 1972, p. 260). Thus, the CEV model is completely analytically tractable.

4.4.5 The jump-to-default extended CEV model

In the standard CEV model default happens when the stock price hits zero. However, for empirically realistic values of parameters β and δ the default probability is small. Moreover, default often happens as a surprise when the firm has a positive stock price. Carr and Linetsky (2006) extend the CEV model by introducing the possibility of default from a positive stock price. They introduce default intensity specified as a function of the underlying stock price as follows:

$$h(x) = b + c\sigma^2(x) = b + c\delta^2 x^{2\beta},$$

with default intensity parameters $b \geqslant 0$ and $c > 0$ (here we consider a constant parameter version of the model; the general version allows for deterministic time dependence of parameters). This default intensity is an affine function of the instantaneous stock variance $\sigma^2(x)$ (the greater the stock volatility, the greater the default intensity). Since in the CEV model the stock volatility is a negative power of the stock price (recall that the CEV diffusion coefficient is $a(x) = \sigma(x)x = \delta x^{1+\beta}$), the default intensity is also a negative power of the stock price (plus a constant). In order for the discounted gains process (including price changes, dividends, and possible default) to be a martingale under the risk-neutral measure, the risk-neutral drift of the process has to be adjusted to compensate the jump:

$$b(x) = \big(r - q + h(x)\big)x = \big(r - q + b + c\delta^2 x^{2\beta}\big)x.$$

The resulting model is called *Jump-to-Default extended CEV* (JDCEV for short). When $c \geqslant 1/2$, zero is an inaccessible boundary. As the killing rate (default intensity) increases fast, the process is always killed from a positive value (default occurs from a positive stock price via a jump-to-default), before the process has the opportunity to diffuse down to zero. When $c < 1/2$, zero is a killing boundary for the process (exit for $\beta \in [c - 1/2, 0)$ and regular specified as killing for $\beta < c - 1/2$), and the default event can occur either from a positive value by a jump-to-default, or via diffusion down to zero. The JDCEV model retains complete analytical tractability of the standard CEV model, with defaultable bond and stock option pricing formulas available in closed from (Carr and Linetsky, 2006).

4.5 Coulomb potential, Bessel processes with constant drift, and related models

4.5.1 Coulomb potential and Bessel processes with constant drift

Next we consider a d-dimensional ($d \geqslant 2$) *pole seeking Brownian motion* process $\{X_t, t \geqslant 0\}$, a diffusion process in $\mathbb{R}^d$ with the infinitesimal generator

$$\frac{1}{2}\Delta + \mu\frac{x}{|x|}\cdot\nabla$$

with $\mu \in \mathbb{R}$. The radial part of this process, $\{R_t = |X_t|, t \geqslant 0\}$, turns out to be a one-dimensional diffusion process with

$$a(x) = 1, \qquad b(x) = \frac{\nu + 1/2}{x} + \mu, \quad x \in (0, \infty),$$

where $\nu = d/2 - 1$. For $d = 2$ and $\mu < 0$ this process was studied by Kendall (1972), where it was called *a pole seeking Brownian motion* (see also Pitman and Yor, 1981, pp. 362–364; Yor, 1984, p. 104; and DeLong, 1981 for related work). Here we consider this process with $\nu, \mu \in \mathbb{R}$ and call it *Bessel process of index ν with constant drift μ*. The boundary classification at 0 is independent of μ and is the same as for the standard Bessel process. For all $\nu \in \mathbb{R}$,

$+\infty$ is oscillatory natural with cutoff $\Lambda = \mu^2/2$. According to Theorem 3.3, $+\infty$ is non-oscillatory (oscillatory) for $\lambda = \Lambda = \mu^2/2$ if $\mu(\nu + 1/2) \geqslant 0$ ($\mu(\nu + 1/2) < 0$). Thus, we are in the Spectral Category II with purely absolutely continuous spectrum above $\mu^2/2$. If $\mu(\nu + 1/2) < 0$ we have an infinite sequence of eigenvalues clustering at $\mu^2/2$ (i.e., $\mu^2/2$ is the limit point of the point spectrum). If $\mu(\nu + 1/2) \geqslant 0$, there may only be a finite set of non-negative eigenvalues in $[0, \mu^2/2]$ (it turns out that this set is empty). The explicit form of the spectral representation for this process is given in Linetsky (2004d).

The Liouville transformation with $x_0 = 0$ reduces the SL equation to the Schrödinger equation with *Coulomb potential* (Morse and Feshbach, 1953, p. 1663):

$$Q(x) = \frac{1}{2}\big(\nu^2 - 1/4\big)x^{-2} + \mu(\nu + 1/2)x^{-1} + \frac{\mu^2}{2}, \quad x \in (0, \infty).$$

This is the celebrated Coulomb potential appearing in the quantum mechanical model of the hydrogen atom. The Schrödinger equation with Coulomb potential has the form of the Whittaker equation with the Whittaker functions as solutions.

A property of this process important for practical applications is that if we kill the process at the rate $k(x) = \frac{1}{2}\beta^2 x^{-2} + \gamma x^{-1}$, the process $\hat{R}$ with killing is just as analytically tractable as the process R without killing, $\beta = \gamma = 0$. Indeed, both processes lead to the same potential function:

$$Q(x) = \frac{1}{2}\big(\nu^2 + \beta^2 - 1/4\big)x^{-2} + \big(\mu(\nu + 1/2) + \gamma\big)x^{-1} + \frac{\mu^2}{2},$$
$$x \in (0, \infty).$$

Only the coefficients in front of x^{-2} and x^{-1} change.

Remark. Yor (1984) calls this process with unit volatility and drift $b(x) = (\nu + 1/2)x^{-1} - \delta$ Bessel process with "naive" drift $\delta > 0$ in order to avoid confusion with the diffusion obtained by taking the radial part of an $\mathbb{R}^d$-valued Brownian motion started at the origin and with some drift vector $\vec{\delta}$ and with the infinitesimal generator $\frac{1}{2}\Delta + \vec{\delta} \cdot \nabla$. This latter diffusion has unit volatility and drift

$$b(x) = \frac{\nu + 1/2}{x} + \frac{\delta I_{\nu+1}(\delta x)}{I_\nu(\delta x)},$$

where $\nu = d/2 - 1$, $\delta = |\vec{\delta}|$, and $I_\nu(z)$ is the Bessel function of order ν, and is usually called Bessel process with drift (Watanabe, 1975; Pitman and Yor, 1981, p. 310).

4.5.2 Two non-affine term structure models

Let $\{X_t, t \geqslant 0\}$ be a Bessel process with $\nu > 1$ and constant drift $\mu < 0$ and $\alpha = \beta = 0$. For $\sigma > 0$, the reciprocal squared process $\{4\sigma^{-2}R_t^{-2}, t \geqslant 0\}$ is a diffusion on $(0, \infty)$ with the infinitesimal parameters

$$a(x) = \sigma x^{3/2}, \qquad b(x) = \kappa(\theta - x^{1/2})x^{3/2},$$
$$\kappa := \frac{\sigma^2}{2}(\nu - 1) > 0, \ \theta := -\frac{2\mu}{\sigma(\nu - 1)} > 0.$$

This diffusion with infinitesimal variance $\sigma^2 x^3$ is similar to the 3/2 model of Section 4.4.3 but has a different specification of non-linear drift. The spectral representation of the state-price density in the term structure model with the short rate following this diffusion is obtained by killing the process at the linear rate $r(x) = x$. This corresponds to killing the Bessel process with constant drift at the rate $r(x) = 4\sigma^{-2}x^{-2}$. Transforming to the Liouville normal form, only the coefficient in front of the x^{-2} is modified. Thus, in this term structure model the spectral representation for the state-price density follows immediately from the spectral representation for the transition density of the Bessel process with constant negative drift.

Let $\{X_t, t \geqslant 0\}$ be a Bessel process with $\nu \geqslant 1/2$ and constant drift $\mu < 0$ and $\alpha = \beta = 0$. For some $\sigma > 0$, the reciprocal process $\{(\sigma X_t)^{-1}, t \geqslant 0\}$ is a diffusion on $(0, \infty)$ with the infinitesimal parameters

$$a(x) = \sigma x^2, \qquad b(x) = \kappa(\theta - x)x^2,$$
$$\kappa := \sigma^2(\nu - 1/2) > 0, \ \theta := -\frac{\mu}{\sigma(\nu - 1/2)} > 0.$$

This diffusion has the CEV infinitesimal variance $\sigma^2 x^{2\gamma}$ with $\gamma = 2$ in contrast to the 3/2 model. Its drift is also more non-linear that the 3/2 model. In light of the recent empirical evidence on non-linearity of the drift and high positive values for γ, this model may be expected to outperform the 3/2 model in empirical tests. The spectral representation of the state-price density is obtained by killing the process at the rate $r(x) = x$. This corresponds to killing the Bessel process with constant drift at the rate $r(x) = (\sigma x)^{-1}$. Transforming to the Liouville normal form, only the coefficient in front of x^{-1} is modified. Thus, in this term structure model the spectral representation for the state-price density follows immediately from the spectral representation for the transition density of the Bessel process with constant negative drift. More details on Bessel processes with constant drift and the associated non-affine term structure models can be found in Linetsky (2004d).

4.5.3 A jump-to-default extended Black–Scholes model

Consider the following extension of the Black–Scholes model that includes default. Start with the standard Black–Scholes geometric Brownian motion

process for the stock price and assume that the default intensity has the form:

$$h(x) = \frac{c}{\ln(x/B)},$$

where x is the stock price of the underlying firm and B is some threshold default barrier. This intensity tends to infinity as the stock price falls towards the barrier, and tends to zero as the stock price goes to infinity. This specification of default intensity as a function of the stock price is similar to the one used by Madan and Unal (1998).

In order for the discounted gain process (including stock price changes, dividends, and possible default) to be a martingale under the risk-neutral measure, the risk-neutral drift of the process has to be adjusted as follow (the volatility is assumed lognormal in the Madan and Unal model, $a(x) = \sigma x$):

$$b(x) = \big(r - q + h(x)\big)x = \left(r - q + \frac{c}{\ln(x/B)}\right)x.$$

We now show that this model reduces to the killed Bessel process with constant drift associated with the Coulomb potential and is, thus, analytically tractable. Let $\{S_t, t \geqslant 0\}$ be a diffusion process with the above drift and lognormal volatility $\sigma > 0$. Introduce a new process $\{R_t = \sigma^{-1}\ln(S_t/B)\}$. This process is a Bessel process with $\nu = c/\sigma^2 - 1/2$ and constant drift $\mu = (r - q - \sigma^2/2)/\sigma$. The associated Schrödinger potential is the Coulomb potential. Furthermore, introducing default by killing the original process at the rate $h(x) = c/\ln(x/B)$ (equivalently, killing the Bessel process with drift at the rate $h(x) = c\sigma^{-1}x^{-1}$) only modifies the constant in front of the term with x^{-1} in the Coulomb potential, and we obtain the analytical state-price density in this model of default.

4.6 Morse potential, geometric Brownian motion with affine drift, and related models

4.6.1 Morse potential and geometric Brownian motion with affine drift

Consider a diffusion process with infinitesimal parameters

$$a(x) = \sigma x, \qquad b(x) = Ax + B,$$
$$x \in (0, \infty),\ \sigma > 0,\ A, B \in \mathbb{R},\ B \neq 0. \tag{4.15}$$

For $B = 0$ this is a geometric Brownian motion. For $B \neq 0$ we call this process *geometric Brownian motion with affine drift*. This process was studied by Wong (1964) who obtained a spectral representation for $B > 0$ and $A < \sigma^2/2$ (this process also appeared in Shiryaev, 1961 in the context of quickest detection problems; see also Peskir, 2006). For all $A, B \in \mathbb{R}$, $+\infty$ is a natural boundary. Adding the constant $B \neq 0$ in the drift drastically changes the behavior of the process near the origin. For all $A \in \mathbb{R}$, 0 is exit (entrance) for $B < 0$ ($B > 0$).

For $B > 0$ and $A < \sigma^2/2$ (the case studied by Wong, 1964), the process has a stationary distribution with the reciprocal Gamma density:

$$\pi(x) \sim x^{\frac{2A}{\sigma^2}-2} e^{-\frac{2B}{\sigma^2}x^{-1}}.$$

Let $\zeta := T_0$ be the lifetime of the process ($\zeta = \infty$ if $B > 0$). The process $\{\ln X_t, 0 \leqslant t < \zeta\}$ is a diffusion with

$$a(x) = \sigma, \qquad b(x) = A - \frac{1}{2}\sigma^2 + Be^{-x}.$$

The Liouville transformation reduces the corresponding SL equation to the Schrödinger equation with *Morse potential* (Morse, 1929; Morse and Feshbach, 1953, p. 1671):

$$\begin{aligned} &Q(x) = c_0 + c_1 e^{-\gamma x} + c_2 e^{-2\gamma x}, \quad x \in (0, \infty), \\ &c_0 = \frac{1}{2\sigma^2}\left(A - \frac{1}{2}\sigma^2\right)^2, \quad c_1 = B\left(\frac{A}{\sigma^2} - 1\right), \quad c_2 = \frac{B^2}{2\sigma^2}, \quad \gamma = 2\sigma. \end{aligned} \tag{4.16}$$

The Schrödinger equation with potential of the form $ae^{-2\gamma x} + be^{-\gamma x}$ first appeared in Morse (1929). The origin is non-oscillatory and $+\infty$ is oscillatory with cutoff $\Lambda = c_0$. The cutoff value is non-oscillatory. Thus we have absolutely continuous spectrum above c_0 plus a finite set of non-negative eigenvalues below c_0.

Consider a standardized version of the process with $B > 0$ that has the following standardized parameters $\sigma = 2$, $A = 2(\nu + 1)$, and $B = 1$. In what follows this process will be denoted $X^{(\nu)}$. This process was studied in Donati-Martin et al. (2001) and Linetsky (2004b) in connection with Asian options and in Linetsky (2006) in connection with the jump-to-default extended Black–Scholes model. The fundamental solutions and their Wronskian for this process are:

$$\begin{aligned} &\psi_s(x) = x^{\frac{1-\nu}{2}} e^{\frac{1}{4x}} W_{\frac{1-\nu}{2},\mu(\alpha)}\left(\frac{1}{2x}\right), \\ &\phi_s(x) = x^{\frac{1-\nu}{2}} e^{\frac{1}{4x}} \mathcal{M}_{\frac{1-\nu}{2},\mu(\alpha)}\left(\frac{1}{2x}\right), \qquad w_\alpha = \frac{1}{2\Gamma(\mu(\alpha) + \nu/2)}, \end{aligned}$$

where $\mu(\alpha) = \frac{1}{2}\sqrt{2\alpha + \nu^2}$, and $\mathcal{M}_{\kappa,\mu}(z)$ and $W_{\kappa,\mu}(z)$ are the Whittaker functions ($\mathcal{M}_{\kappa,\mu}(z) = M_{\kappa,\mu}(z)/\Gamma(2\mu + 1)$ is the regularized Whittaker function).

For $\nu < 0$, the Green's function (3.10) considered in the complex α-plane has some poles in the interval $\alpha \in [-\nu^2/2, 0]$ (the poles of the Gamma function in the denominator of w_α) and a branch cut from $\alpha = -\nu^2/2$ to $-\infty$ placed along the negative real axes. It is convenient to parameterize the branch cut as follows: $\{\alpha = -\nu^2/2 - \rho^2/2, \rho \in [0, \infty)\}$. Applying the Cauchy Residue Theorem, the Laplace inversion produces the spectral representation of the

symmetric transition density:

$$p_m^{(\nu)}(t;x,y) = \mathbf{1}_{\{\nu<0\}} \sum_{n=0}^{[|\nu|/2]} e^{-\lambda_n t} \mathrm{Res}_{\alpha=-\lambda_n} G_\alpha(x,y) - \frac{1}{2\pi i} \int_0^\infty e^{-\frac{(\nu^2+\rho^2)t}{2}} \{G_{\frac{1}{2}(\nu^2+\rho^2)e^{i\pi}}(x,y) - G_{\frac{1}{2}(\nu^2+\rho^2)e^{-i\pi}}(x,y)\} \rho \, \mathrm{d}\rho.$$

The poles give the eigenvalues, and the residues at the poles give the corresponding contributions to the density from the eigenfunctions. The integral along the branch cut produces the continuous spectrum. For $x, y > 0$ and $\nu \in \mathbb{R}$, the transition density has the following spectral representation:

$$\begin{aligned} p^{(\nu)}(t;x,y) &= \mathbf{1}_{\{\nu<0\}} \pi(y) \\ &\quad + \mathbf{1}_{\{\nu<-2\}} \sum_{n=1}^{[|\nu|/2]} e^{-2n(|\nu|-n)t} \frac{2(|\nu|-2n)n!}{\Gamma(1+|\nu|-n)} \\ &\quad \times e^{-\frac{1}{2y}} (2x)^n (2y)^{n-1-|\nu|} L_n^{(|\nu|-2n)}\left(\frac{1}{2x}\right) L_n^{(|\nu|-2n)}\left(\frac{1}{2y}\right) \\ &\quad + \frac{1}{2\pi^2} \int_0^\infty e^{-\frac{(\nu^2+\rho^2)t}{2}} e^{\frac{1}{4x}-\frac{1}{4y}} \left(\frac{y}{x}\right)^{\frac{\nu-1}{2}} W_{\frac{1-\nu}{2},\frac{i\rho}{2}}\left(\frac{1}{2x}\right) \\ &\quad \times W_{\frac{1-\nu}{2},\frac{i\rho}{2}}\left(\frac{1}{2y}\right) \left|\Gamma\left(\frac{\nu+i\rho}{2}\right)\right|^2 \sinh(\pi\rho)\rho \, \mathrm{d}\rho, \end{aligned}$$

where $L_n^{(\alpha)}(x)$ are the generalized Laguerre polynomials, $[x]$ denotes the integer part of x, $\mathbf{1}_{\{\cdot\}}$ is the indicator function, and

$$\pi(x) = \frac{2^\nu}{\Gamma(-\nu)} x^{\nu-1} e^{-\frac{1}{2x}}$$

is the stationary density of the process (reciprocal Gamma). When $x = 0$ (recall that the process can be started at zero (zero is an entrance boundary)), $y > 0$, and $\nu \in \mathbb{R}$,

$$\begin{aligned} p^{(\nu)}(t;0,y) &= \frac{1}{2\pi^2} \int_0^\infty e^{-\frac{(\nu^2+\rho^2)t}{2}} e^{-\frac{1}{4y}} (2y)^{\frac{\nu-1}{2}} W_{\frac{1-\nu}{2},\frac{i\rho}{2}}\left(\frac{1}{2y}\right) \\ &\quad \times \left|\Gamma\left(\frac{\nu+i\rho}{2}\right)\right|^2 \sinh(\pi\rho)\rho \, \mathrm{d}\rho \\ &\quad + \mathbf{1}_{\{\nu<0\}} \frac{2^\nu}{\Gamma(-\nu)} x^{\nu-1} e^{-\frac{1}{2x}} \end{aligned}$$

$$+ \mathbf{1}_{\{\nu<-2\}} \sum_{n=1}^{[|\nu|/2]} e^{-2n(|\nu|-n)t} \frac{(-1)^n 2(|\nu|-2n)}{\Gamma(1+|\nu|-n)} \times e^{-\frac{1}{2y}} (2y)^{n-1-|\nu|} L_n^{(|\nu|-2n)}\left(\frac{1}{2y}\right).$$

4.6.2 A jump-to-default extended Black–Scholes model

Linetsky (2006) studies the following extension of the Black–Scholes model that includes default. Start with the standard Black–Scholes geometric Brownian motion process for the stock price and assume that the default intensity has the form:

$$h(x) = cx^{-p},$$

for some $p > 0$ and $c > 0$. That is, the default intensity is the negative power of the underlying stock price. When the stock tends to zero, the intensity tends to infinity. When the stock tends to infinity, the intensity asymptotically goes to zero. In order for the discounted gains process (including stock price changes, dividends, and possible default) to be a martingale under the risk-neutral measure, the risk-neutral drift of the process has to be adjusted to compensate for the possible default (the volatility is assumed lognormal, $a(x) = \sigma x$):

$$b(x) = \big(r - q + h(x)\big)x = \big(r - q + cx^{-p}\big)x.$$

Let $\{S_t, t \geqslant 0\}$ be a diffusion process with this drift and volatility. Introduce a new process $\{Z_t = \beta S_t^p, t \geqslant 0\}$, where $\beta = p\sigma^2/(4\alpha)$. This process is a diffusion with volatility and drift:

$$a(x) = \gamma x, \qquad b(x) = \alpha x + \beta,$$
$$\text{with } \alpha = p\big(r - q + (p+1)\sigma^2/2\big),\ \beta = p^2\sigma^2/4,\ \gamma = p\sigma.$$

Thus, we are in the class (4.15) with the analytically tractable transition density, and closed-form expressions are available for both corporate bonds and stock options in this model.

4.6.3 Arithmetic Asian options

The process $X^{(\nu)}$ and its density $p^{(\nu)}$ are also closely related to the problem of pricing arithmetic Asian options. Assume that, under the risk-neutral measure, the underlying asset price follows a geometric Brownian motion process $\{S_t = S_0 \exp(\sigma B_t + (r - q - \sigma^2/2)t), t \geqslant 0\}$. For $t > 0$, let $\mathcal{A}_t$ be the continuous arithmetic average price,

$$\mathcal{A}_t = \frac{1}{t}\int_0^t S_u \,\mathrm{d}u.$$

An Asian put option with strike $K > 0$ and expiration $T > 0$ delivers the payoff $(K - \mathcal{A}_T)^+$ at T (Asian calls can be obtained by the call-put parity for

Asian options; we only consider puts here). After standardizing the problem (see Geman and Yor, 1993), it reduces to computing the expectation of the form

$$P^{(\nu)}(k,\tau) = \mathbb{E}\big[\big(k - A_\tau^{(\nu)}\big)^+\big],$$

where $\tau = \sigma^2 T/4$, $\nu = 2(r - q - \sigma^2/2)/\sigma^2$, $k = \tau K/S_0$, and $A_\tau^{(\nu)}$ is the so-called *Brownian exponential functional* (see Yor, 2001)

$$A_\tau^{(\nu)} = \int_0^\tau e^{2(B_u+\nu u)}\,\mathrm{d}u.$$

Dufresne's identity in law (Dufresne, 1989, 1990; see also Donati-Martin et al., 2001) states that, for each fixed $t > 0$,

$$A_t^{(\nu)} \overset{(law)}{=} X_t^{(\nu)},$$

where $X_t^{(\nu)}$ is starting at the origin. Since the spectral representation for this diffusion is known, we immediately obtain the spectral expansion for the arithmetic Asian put (Linetsky, 2004b):

$$\begin{aligned}
P^{(\nu)}(k,\tau) = {} & \frac{1}{8\pi^2}\int_0^\infty e^{-\frac{(\nu^2+\rho^2)\tau}{2}}(2k)^{\frac{\nu+3}{2}}e^{-\frac{1}{4k}}W_{-\frac{\nu+3}{2},\frac{i\rho}{2}}\left(\frac{1}{2k}\right)\\
& \times\left|\Gamma\left(\frac{\nu+i\rho}{2}\right)\right|^2\sinh(\pi\rho)\rho\,\mathrm{d}\rho\\
& + \mathbf{1}_{\{\nu<0\}}\frac{1}{2\Gamma(|\nu|)}\big\{2k\Gamma\big(|\nu|,1/(2k)\big) - \Gamma\big(|\nu|-1,1/(2k)\big)\big\}\\
& + \mathbf{1}_{\{\nu<-2\}}e^{-2(|\nu|-1)\tau}\frac{(|\nu|-2)}{2\Gamma(|\nu|)}\Gamma\big(|\nu|-2,1/(2k)\big)\\
& + \mathbf{1}_{\{\nu<-4\}}\sum_{n=2}^{[|\nu|/2]}e^{-2n(|\nu|-n)\tau}\frac{(-1)^n(|\nu|-2n)}{2n(n-1)\Gamma(1+|\nu|-n)}\\
& \times(2k)^{\nu+n+1}e^{-\frac{1}{2k}}L_{n-2}^{(|\nu|-2n)}\left(\frac{1}{2k}\right).
\end{aligned}$$

This explicit expression constitutes an analytical inversion of the celebrated Geman and Yor (1993) Laplace transform result for arithmetic Asian options originally obtained via the theory of Bessel processes (see also Donati-Martin et al., 2001 for an alternative derivation of the Geman and Yor Laplace transform via Dufresne's identity).

4.6.4 Merton's cash dividends model

Let $\{S_t, t \geqslant 0\}$ be the risk-neutral price process for an asset with constant volatility $\sigma > 0$ that continuously pays cash dividends at the rate of $D > 0$ dollars per year. It is a diffusion with infinitesimal parameters ($r > 0$ is the constant risk-free rate):

$$a(x) = \sigma x, \qquad b(x) = rx - D. \tag{4.17}$$

Similar to the arithmetic Asian option application, it is a diffusion of the general form (4.15). The difference is that here $B = -D < 0$ and, hence, zero is exit (the asset price hits zero in finite time and is sent to the bankruptcy state with probability one), while in the Asian option application $B = 1 > 0$, and zero is entrance. The problem of pricing options on assets that continuously pay cash dividends has been first studied by Merton (1973) who obtained an asymptotic result for the call option price with infinite time to maturity. Lewis (1998) has recently obtained closed-form solutions for call and put options with finite time to maturity in terms of spectral expansions. As mentioned in the remark in Section 3, in this case neither call nor put payoffs are square-integrable with the speed density, and the analysis requires special care. The close connection between the Asian option problem and the cash dividends problem was also emphasized by Lipton (1999).

4.6.5 GARCH stochastic volatility model, spot energy model, and Brennan and Schwartz interest rate model

Under a GARCH diffusion specification (Nelson, 1990; see also Lewis, 2000) the infinitesimal variance $V_t = \sigma_t^2$ of the asset return is assumed to follow a positive mean-reverting diffusion process solving the SDE

$$\mathrm{d}V_t = \kappa(\theta - V_t)\,\mathrm{d}t + \xi V_t\,\mathrm{d}B_t, \quad V_0 = \sigma_0^2,$$

where $\theta > 0$ is the long-run variance level, $\xi > 0$ is the volatility of variance, and $\kappa > 0$ is the rate of mean-reversion. This process belongs to the family (4.15) with $A < 0$ and $B > 0$. It possesses a stationary distribution with the reciprocal Gamma density. Re-scaling and time-changing, the process $\{X_t := \alpha V_{4t/\xi^2}, t \geqslant 0\}$, where $\alpha = \xi^2/(4\kappa\theta)$, is a diffusion $X^{(\nu)}$ with $\nu = -1 - 4\kappa/(\xi^2) < -1$ and starting at $X_0 = \alpha V_0 > 0$. Hence, the spectral representation for the transition density of the GARCH stochastic volatility process is immediately obtained from the spectral representation for the process $X^{(\nu)}$.

In the energy markets the spot price is also often modeled as a mean-reverting positive diffusion (e.g., Pilipović, 1998, p. 64)

$$\mathrm{d}S_t = \kappa(L - S_t)\,\mathrm{d}t + \sigma S_t\,\mathrm{d}B_t, \quad S_0 > 0,$$

where $L > 0$ is the long-run equilibrium spot price level, $\sigma > 0$ is the spot price volatility and $\kappa > 0$ is the rate of mean-reversion. This is the same process as in the GARCH diffusion model.

The same diffusion is also known in the interest rate literature as the Brennan–Schwartz (1979) short rate model. While the spectral representation for the transition density for this model is available in closed form, there is no analytical solution for the state-price density and, hence, zero-coupon bonds. If we discount at the linear rate $r(x) = x$, the resulting SL equation leads to the Schrödinger equation with a *three-term* potential of the form $c_1 e^{-\gamma y} + c_2 e^{-2\gamma y} + c_3 e^{\gamma y}$. To the best of our knowledge, there is no closed-form solution in terms of classical special functions.

4.6.6 *Two non-affine term structure models*

For $A < \sigma^2/2$ and $B > 0$ let $\{X_t, t \geqslant 0\}$ be a diffusion with the infinitesimal parameters (4.15) and consider a reciprocal process $\{(\sigma X_t)^{-1}, t \geqslant 0\}$. It is a diffusion on $(0, \infty)$ with the infinitesimal parameters:

$$a(x) = \sigma x, \qquad b(x) = \kappa(\theta - x)x,$$
$$\kappa := \sigma B > 0, \ \theta := (\sigma B)^{-1}\big(\sigma^2 - A\big) > 0.$$

This process has been deduced by Merton (1975) as a model for the short rate in his economic growth model. Lewis (1998) has recently obtained closed-form solutions for zero-coupon bonds in this model in terms of spectral expansions.

Consider a squared reciprocal process $\{(\sigma X_t)^{-2}, t \geqslant 0\}$. It is a diffusion on $(0, \infty)$ with infinitesimal parameters:

$$a(x) = 2\sigma x, \qquad b(x) = \kappa\big(\theta - x^{1/2}\big)x,$$
$$\kappa := 2\sigma B > 0, \ \theta := (\sigma B)^{-1}\big(3\sigma^2/2 - A\big) > 0.$$

This process provides another analytically tractable specification for the short rate with non-affine drift.

4.7 *Modified Pöschl–Teller potential, hypergeometric diffusions, and volatility skew*

4.7.1 *Hypergeometric diffusions*

Our next example is a diffusion with the infinitesimal parameters

$$a(x) = \sqrt{Ax^2 + Bx + C}, \qquad b(x) = \mu x,$$
$$A, C \geqslant 0, \ B^2 < 4AC, \ \mu \in \mathbb{R}. \tag{4.18}$$

For $B^2 < 4AC$ the parabola $Ax^2 + Bx + C$ is above the x-axis for all real x. Hence, a diffusion $\{X_t, t \geqslant 0\}$ with the infinitesimal parameters (4.18) can be defined on the whole real line. We call this process *hypergeometric diffusion* since solutions to the associated SL equation are expressed in terms of the Gauss hypergeometric function. This diffusion was first studied by Wong (1964) in a special case. The process $\{Z_t := (2AX_t + B)(4AC - B^2)^{-1/2}, t \geqslant 0\}$

is a diffusion on $\mathbb{R}$ with the infinitesimal parameters:

$$a(x) = \sqrt{A(x^2+1)}, \qquad b(x) = \mu(x-L), \quad L := \frac{B}{\sqrt{4AC-B^2}}.$$

The process $\{Y_t := \operatorname{arcsinh}(Z_t),\, t \geqslant 0\}$ is a diffusion on $\mathbb{R}$ with the infinitesimal parameters:

$$a(x) = \sqrt{A}, \qquad b(x) = \left(\mu - \frac{1}{2}A\right)\tanh x - \frac{\mu L}{\cosh x}.$$

The Liouville transformation reduces the associated SL equation to the Schrödinger equation with potential:

$$\begin{aligned} Q(x) &= c_0 + \frac{c_1}{\cosh^2(Ax)} + c_2\frac{\sinh(Ax)}{\cosh^2(Ax)}, \\ c_0 &:= \frac{1}{2}\left(\frac{\mu}{A} - \frac{1}{2}\right)^2, \quad c_1 := \frac{\mu^2L^2}{2A} - \frac{1}{2A}\left(\mu - \frac{A}{2}\right)\left(\mu - \frac{3A}{2}\right), \\ c_2 &:= \frac{\mu L}{A}(A-\mu). \end{aligned}$$

This is known as a *hyperbolic barrier potential* closely related to the *modified Pöschl–Teller potential* (Grosche and Steiner, 1998, p. 251). Solutions are expressed in terms of the Gauss hypergeometric function. Both $-\infty$ and $+\infty$ are oscillatory with cutoff $\Lambda_1 = \Lambda_2 = c_0$ ($Q(x) \to \infty$ as $y \to \pm\infty$) and we have non-simple purely absolutely continuous spectrum above c_0 (similar to the standard Brownian motion on the real line) plus a finite set of eigenvalues below c_0 (Spectral Category III).

4.7.2 A volatility skew model

For the purpose of modeling non-negative asset prices, we restrict the process with the infinitesimal parameters (4.18) to the positive half-line $(0, \infty)$ and make 0 a killing boundary by sending the process to a cemetery state Δ at its lifetime $\zeta = T_0$. The resulting process $\{S_t,\, t \geqslant 0\}$ is a hybrid of the Black–Scholes–Merton geometric Brownian motion, square-root, and absolute diffusion specifications. With three volatility parameters A, B, and C, this model can be calibrated to a variety of implied volatility skew shapes. To facilitate calibration, it is convenient to parameterize the diffusion coefficient as follows:

$$a(x) = \sigma_K\sqrt{\frac{x^2 + \alpha K x + \beta K^2}{1+\alpha+\beta}}, \quad \sigma_K > 0,\ K > 0,\ \alpha \in \mathbb{R},\ \beta > \frac{\alpha^2}{4},$$

so that $a(K) = \sigma_K K$. Here K is some reference asset price level (e.g., the asset price level at the time of calibration), σ_K is the local volatility at the reference level (ATM volatility), and α and β are two model parameters governing the

shape of the *local volatility function*

$$\sigma(x) := \frac{a(x)}{x} = \sigma_K \sqrt{\frac{1 + \alpha(K/x) + \beta(K/x)^2}{1 + \alpha + \beta}}.$$

4.8 Pöschl–Teller-type potentials, Jacobi diffusion, and related models

Larsen and Sorensen (2007) propose an analytically tractable model for a foreign exchange rate in a target zone enforced by the central banks and applied it to the European Monetary System from 1979 until the introduction of the Euro in 1999. They model a spot exchange rate as a diffusion process S that may not be more than $a \times 100$ percent over the central parity μ for any of the two currencies, i.e., $\mu/(1+a) < S_t < \mu(1+a)$. Then $X_t = \ln(S_t)$ must satisfy that $m - z < X_t < m + z$, where $m = \ln \mu$ and $z = \ln(1+a)$. The logarithm of the spot exchange rate is modeled as a diffusion

$$\mathrm{d}X_t = \beta\big[(m + \gamma z) - X_t\big]\,\mathrm{d}t + \sigma\sqrt{z^2 - (X_t - m)^2}\,\mathrm{d}B_t,$$

where $\beta > 0$ and $\gamma \in (-1, 1)$. This process reverts to its mean $m + \gamma z$. The parameter γ is an asymmetry parameter that expresses whether one currency is stronger than the other. The process is ergodic on the interval $(m - z, m + z)$ if and only if $\beta \geqslant \sigma^2$ and $-1 + \sigma^2/\beta \leqslant \gamma \leqslant 1 - \sigma^2/\beta$. The stationary distribution is a shifted and re-scaled Beta distribution. Both boundaries at $m - z$ and $m + z$ are unattainable entrance boundaries. A standardized process $Z_t := (X_t - m)/z$ follows a *Jacobi diffusion*

$$\mathrm{d}Z_t = \kappa(\gamma - Z_t)\,\mathrm{d}t + A\sqrt{1 - Z^2}\,\mathrm{d}B_t, \quad \kappa := \beta z,\ A := \sigma z,$$

on the interval $(-1, 1)$. The Liouville transformation reduces the associated SL equation to the Schrödinger equation with potential of the form:

$$Q(x) = c_0 + \frac{c_1}{\cos^2(\gamma x)} + c_2 \frac{\sin(\gamma x)}{\cos^2(\gamma x)}, \quad \gamma := \sigma z.$$

This potential is of the Pöschl–Teller type (Grosche and Steiner, 1998, Section 6.5, pp. 240–241). The spectrum is purely discrete and the eigenfunctions are expressed in terms of Jacobi polynomials. The explicit form of the spectral representation is given in Wong (1964) and Karlin and Taylor (1981, p. 335).

Delbaen and Shirakawa (2002a) study an interest rate model based on the Jacobi diffusion. The short rate is assumed to stay within a range between some lower and upper boundaries. See also Borodin and Salminen (2004) for some applications of Jacobi diffusions and further references.

4.9 Concluding comments on one-dimensional diffusion models

In this section we have cataloged some of the most important analytically tractable diffusions that appear in financial applications. Further analytically

tractable examples of Schrödinger potentials can be found in the literature on the Schrödinger equation (see Grosche and Steiner, 1998 and references therein). In particular, the general form of potential functions leading to solutions in terms of hypergeometric and confluent hypergeometric equations have been classified (the so-called *Natanzon potentials*, Natanzon, 1979). With each of these potentials one can associate diffusion processes with transition densities explicitly given in terms hypergeometric functions (see, e.g., Albanese and Lawi, 2005 for some recent work in this direction).

For simplicity, in this survey we only discussed the case of continuous infinitesimal parameters, assuming $a \in C^2(I)$, $b \in C^1(I)$, and $k \in C(I)$. However, the methods described can be used to study diffusions with discontinuous coefficients. Decamps et al. (2006) consider a class of models called *self-exciting threshold models*, which are diffusion counterparts of discrete-time regime switching time series models. In these models drift, volatility, or discount rate can have discontinuities, changing when the process crosses a certain level. For example, drift can have the form:

$$b(x) = \mathbf{1}_{\{x<k\}} b_1(x) + \mathbf{1}_{\{x\geqslant k\}} b_2(x),$$

where k is some threshold. Below k the drift b_1 is in effect. Above k the drift b_2 is in effect. The spectral method is well suited to such models, yielding the spectral representation of the transition density (Decamps et al., 2006).

We hope that the survey in this section provides enough ammunition for our reader, whenever faced with a diffusion process, to be able to determine whether or not the process is analytically tractable in the sense that an explicit analytical expression for its transition density can be obtained in terms of classical special functions and, if the answer is positive, to find that expression.

5 Symmetric multi-dimensional diffusions

5.1 Drift of a symmetric diffusion

We now come back to the set-up of Section 1.1 and consider a d-dimensional diffusion process $\hat{X}$ in $D \subseteq \mathbb{R}^d$, with $D = \mathbb{R}^d$ or an open domain in $\mathbb{R}^d$. The infinitesimal generator $\mathcal{G}$ of $\hat{X}$ is a 2nd-order differential operator defined by

$$\mathcal{G}f(x) := \frac{1}{2} \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{d} b_i(x) \frac{\partial f}{\partial x_i}(x) - k(x) f(x) \quad (5.1)$$

for $f \in C_c^2(D)$ (twice-continuously differentiable functions with compact support in D). Here $a_{ij}(x)$ is a diffusion matrix, $b_i(x)$ is a drift vector, and $k(x) = r(x) + h(x)$ is the killing rate. For simplicity, we assume that $a_{ij}(x) \in C^2(D)$, and $b_i(x) \in C^1(D)$, $k(x) \in C(D)$, and $a_{ij}(x)$ is positive definite for each

$x \in D$. If the process can reach the boundary of D when started from the interior, it is killed at the first exit time τ_D.

We further assume that a measure m on D is given and has a density with respect to the Lebesgue measure, $m(dx) = m(x)\,dx$, $m(x) > 0$ on D. For simplicity we assume that $m(x) \in C^1(D)$. While the symmetry condition was automatically satisfied for one-dimensional diffusions, for multi-dimensional diffusions it imposes the following restriction on the form of the drift.

Theorem 5.1. *The symmetry condition* $(\mathcal{G}f, g) = (f, \mathcal{G}g)$ *is satisfied for all* $f, g \in C_c^2(D)$ (*twice-differentiable functions with compact support*) *if and only if the drift vector has the form*:

$$b_i(x) = \frac{1}{2}\sum_{j=1}^{d}\frac{\partial a_{ij}}{\partial x_j}(x) + \frac{1}{2}\sum_{j=1}^{d} a_{ij}(x)\frac{\partial \ln m}{\partial x_j}(x). \tag{5.2}$$

Proof. For $f, g \in C_c^2(D)$ we have ($\partial_i = \partial/\partial x_i$):

$$\begin{aligned}(\mathcal{G}f, g) - (f, \mathcal{G}g) &= \frac{1}{2}\sum_{i,j=1}^{d}\int_D a_{ij}(g\partial_i\partial_j f - f\partial_i\partial_j g)m\,dx \\ &\quad + \sum_{i=1}^{d}\int_D b_i(g\partial_i f - f\partial_i g)m\,dx \\ &= \sum_{i=1}^{d}\int_D \left\{b_i - \frac{1}{2}\sum_{j=1}^{d}(\partial_j a_{ij} + a_{ij}\partial_j \ln m)\right\} \\ &\quad \times (g\partial_i f - f\partial_i g)m\,dx,\end{aligned}$$

where we integrated by parts. In order for this to vanish for every $f, g \in C_c^2(D)$, the drift must satisfy

$$b_i - \frac{1}{2}\sum_{j=1}^{d}(\partial_j a_{ij} + a_{ij}\partial_j \ln m) = 0.$$

□

This restriction on the drift insures that the infinitesimal generator can be written in the divergence form:

$$\mathcal{G}f(x) = \frac{1}{2m(x)}\sum_{i,j=1}^{d}\frac{\partial}{\partial x_i}\left(m(x)a_{ij}(x)\frac{\partial f}{\partial x_j}(x)\right) - k(x)f(x). \tag{5.3}$$

Thus, a symmetric diffusion process is parameterized by a diffusion matrix a and a scalar m.

If D is a compact domain in $\mathbb{R}^d$ with a boundary ∂D and the diffusion process is killed at the first hitting time of the boundary, then it is well known

that the infinitesimal generator $\mathcal{G}$ with the Dirichlet boundary condition is a self-adjoint operator in $L^2(D, m)$ with purely discrete spectrum and, hence, the eigenfunction expansion (2.11) is valid. More generally, if the state space is a compact d-dimensional manifold (with or without a boundary; in the former case the process is killed on the boundary), the spectrum is also purely discrete. When the state space is not compact, there may be some essential spectrum.

From the point of view of finance applications, we are interested in the following question. Can a rich enough catalog of analytically tractable symmetric multi-dimensional diffusions be developed (similar to Table 1 for one-dimensional diffusions) to facilitate building analytically tractable multi-factor models in financial applications?

We start by considering the simplest example of the unit diffusion matrix, $a_{ij} = \delta_{ij}$. The drift restriction reduces to:

$$b_i(x) = \frac{1}{2}\frac{\partial \ln m}{\partial x_i}(x),$$

i.e., the drift is the gradient of a scalar, and the infinitesimal generator is (Δ is the Laplacian)

$$\frac{1}{2}\Delta + \nabla\phi \cdot \nabla - k,$$

where $\phi(x) := \frac{1}{2}\ln m(x)$. In particular, consider the case of $D = \mathbb{R}^d$, $k(x) \equiv 0$, and the following two specifications for the scalar $\phi(x)$:

$$\phi_1(x) = \sum_{i=1}^{d} \mu_i x_i, \qquad \phi_2(x) = \sum_{i=1}^{d} \mu_i x_i - \frac{1}{2}\sum_{i,j=1}^{d} \kappa_{ij} x_i x_j,$$

where μ is a constant vector, and κ is a symmetric positive-definite constant matrix. In the first case, the process is a d-dimensional Brownian motion with constant drift μ. In the second case, the process is a d-dimensional Ornstein–Uhlenbeck process with mean-reverting drift $b_i(x) = \mu_i - \sum_{j=1}^{d} \kappa_{ij} x_j$. Both cases are analytically tractable. In the first case there is a continuous spectrum, as in the case of one-dimensional Brownian motion. In the second case the spectrum is purely discrete, as in the case of one-dimensional OU processes, and the eigenfunctions are expressed in terms of Hermite polynomials. Both of these cases are Gaussian. Furthermore, if we kill either of the processes at the rate

$$k(x) = \sum_{i,j=1}^{d} A_{ij} x_i x_j + \sum_{i=1}^{d} B_i x_i + C,$$

i.e., a discount (killing) rate is quadratic-affine in the state variables, we are still in the Gaussian class. The corresponding quadratic term structure models (QTSM) are multi-dimensional version of the one-dimensional QTSM considered in Section 4.3.3, where one can find further references. The eigenfunction

expansion method has been applied to the pricing of interest rate derivatives in QTSM by Boyarchenko and Levendorskiy (2006), who study a more difficult non-symmetric case with the non-symmetric matrix κ_{ij}.

Another classical example of a symmetric diffusion in $\mathbb{R}^d$ is a pole-seeking Brownian motion with the generator $\frac{1}{2}\Delta + \mu\frac{x}{|x|}\cdot\nabla$ considered in Section 4.5.1. This is obtained by choosing $\phi(x) = \mu|x|$.

5.2 Diffusions on symmetric Riemannian manifolds

What about multi-dimensional diffusions with general (non-diagonal) state-dependent diffusion matrices? Are there rich enough classes of analytically tractable processes? The answer to this question comes from harmonic analysis on symmetric Riemannian spaces. We can relate the diffusion matrix a to a Riemannian metric g on the state space and take the state space of the diffusion process to be a Riemannian manifold $(\mathcal{M}, g)$. Our symmetric diffusion process is then interpreted as a process on this Riemannian manifold. Analytical tractability arises if this Riemannian manifold possesses an isometry group G, a Lie group of transformations that leave the metric g invariant. If the state space is a symmetric Riemannian space, then the powerful representation theory of Lie groups and algebras can be brought to bear and construct the spectral expansion of the transition density (heat kernel) explicitly (e.g., Anker et al., 2002; Anker and Ostellari, 2004 and references therein). This provides a rich supply of analytically tractable multi-dimensional diffusions. To the best of our knowledge, these classes of processes have not been systematically explored in mathematical finance so far.

6 Introducing jumps and stochastic volatility via time changes

6.1 Bochner's subordination of Markov processes and semigroups

In this section we follow Carr, Linetsky and Mendoza (2007). Let $\{T_t, t \geqslant 0\}$ be a *subordinator*, i.e., a non-decreasing Lévy process with the Laplace transform ($\lambda \geqslant 0$)

$$\mathbb{E}\left[e^{-\lambda T_t}\right] = e^{-t\phi(\lambda)}$$

with the Laplace exponent

$$\phi(\lambda) = \gamma\lambda + \int_{(0,\infty)} \left(1 - e^{-\lambda s}\right)\nu(\mathrm{d}s)$$

with the Lévy measure $\nu(\mathrm{d}s)$ satisfying

$$\int_{(0,\infty)} (s \wedge 1)\nu(\mathrm{d}s) < \infty,$$

non-negative drift $\gamma \geqslant 0$, and transition kernel $\pi_t(\mathrm{d}s)$,

$$\int_{[0,\infty)} e^{-\lambda s}\pi_t(\mathrm{d}s) = e^{-t\phi(\lambda)}.$$

The standard references on subordinators include Bertoin (1996) and Sato (1999).

Let $\hat{X}$ be a Markov process with lifetime $\hat{\zeta}$ (we are in the framework of Section 1.1). Consider a time-changed (also called *subordinated*) process $\{\hat{X}_t^{\phi}, t \geqslant 0\}$ defined by:

$$\hat{X}_t^{\phi} = \begin{cases} \hat{X}_{T_t}, & T_t < \hat{\zeta}, \\ \Delta, & T_t \geqslant \hat{\zeta} \end{cases}$$

(the process T is assumed to be independent of the process $\hat{X}$; the superscript ϕ indicates that the process $\hat{X}_t^{\phi}$ is a subordinated process with the subordinator with Laplace exponent ϕ). The idea of time changing a stochastic process with a subordinator is originally due to S. Bochner (1948, 1955). The subordinator is also called the directing process. The following fundamental theorem due to R.S. Phillips (1952) (see Sato, 1999, Theorem 32.1, p. 212) characterizes the time-changed transition semigroup and its infinitesimal generator.

Theorem 6.1. *Let $\{T_t, t \geqslant 0\}$ be a subordinator with Lévy measure ν, drift γ, Laplace exponent $\phi(\lambda)$, and transition kernel $\pi_t(ds)$. Let $\{\mathcal{P}_t, t \geqslant 0\}$ be a strongly continuous contraction semigroup of linear operators in the Banach space* **B** *with infinitesimal generator $\mathcal{G}$. Define (the superscript ϕ refers to the subordinated quantities with the subordinator with the Laplace exponent ϕ):*

$$\mathcal{P}_t^{\phi} f := \int_{[0,\infty)} (\mathcal{P}_s f)\pi_t(\mathrm{d}s), \quad f \in \mathbf{B}. \tag{6.1}$$

Then $\{\mathcal{P}_t^{\phi}, t \geqslant 0\}$ is a strongly continuous contraction semigroup of linear operators on **B**. *Denote its infinitesimal generator by $\mathcal{G}^{\phi}$. Then* $\mathrm{Dom}(\mathcal{G}) \subset \mathrm{Dom}(\mathcal{G}^{\phi})$, $\mathrm{Dom}(\mathcal{G})$ *is a core of $\mathcal{G}^{\phi}$, and*

$$\mathcal{G}^{\phi} f = \gamma \mathcal{G} f + \int_{(0,\infty)} (P_s f - f)\nu(\mathrm{d}s), \quad f \in \mathrm{Dom}(\mathcal{G}). \tag{6.2}$$

In our case here the semigroup $\{\mathcal{P}_t, t \geqslant 0\}$ is the transition semigroup of the diffusion process $\hat{X}$ with generator (5.1). We assume that $\hat{X}$ has a density $p(t; x, y)$. From Eq. (6.1) the subordinated process $\hat{X}^{\phi}$ has a density:

$$p^{\phi}(t; x, y) = \int_{[0,\infty)} p(s; x, y)\pi_t(\mathrm{d}s). \tag{6.3}$$

The subordinated process is a Markov process with the generator (6.2). We can re-write this generator in the *Lévy-type* form:

$$\mathcal{G}^{\phi}f(x) := \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}^{\phi}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{d} b_i^{\phi}(x)\frac{\partial f}{\partial x_i}(x) - k^{\phi}(x)f(x)$$
$$+ \int_{D\setminus\{x\}} \left(f(y) - f(x) - \mathbf{1}_{\{\|y-x\|\leqslant 1\}} \sum_{i=1}^{d} (y_i - x_i)\frac{\partial f}{\partial x_i}(x) \right)$$
$$\times \Pi^{\phi}(x, \mathrm{d}y) \tag{6.4}$$

for any $f \in C_c^2(D)$. Here $a_{ij}^{\phi}(x)$, $b_i^{\phi}(x)$, and $k^{\phi}(x)$ are the diffusion matrix, drift vector, and killing rate of the subordinated process, and $\Pi^{\phi}(x, \mathrm{d}y)$ is a *jump measure* (state-dependent *Lévy measure*). Intuitively, for any $x \in D$ and a Borel set $A \subset D\setminus\{x\}$ bounded away from x, the Lévy measure $\Pi(x, A)$ gives the arrival rate of jumps from the state x into the set A, i.e.,

$$\mathbb{P}_x(X_t \in A) \sim \Pi(x, A)t \quad \text{as } t \to 0.$$

If Π is a finite measure with $\lambda(x) := \Pi(x, D) < \infty$ for every $x \in D$, then the process has a finite number of jumps in any finite time interval and $\lambda(x)$ is the (state-dependent) jump arrival rate. If the Lévy measure Π is infinite, then the process X has infinite activity jumps. If $a^{\phi} \equiv 0$, then $\hat{X}^{\phi}$ is a jump process with drift and killing.

6.2 *Lévy's local characteristics of the subordinated process*

We can explicitly identify the local Lévy characteristics a_{ij}^{ϕ}, b_i^{ϕ}, Π^{ϕ}, and k^{ϕ} of the subordinated process by re-writing the operator (6.2) in the form (6.4) (references on the Lévy characteristics of the subordinated process include Okura, 2002, Theorem 2.1 and Chen and Song, 2005a, Section 2).

Theorem 6.2. *The Lévy characteristics a_{ij}^{ϕ}, b_i^{ϕ}, Π^{ϕ}, and k^{ϕ} of the subordinated process are*:

$$a_{ij}^{\phi}(x) = \gamma a_{ij}(x),$$
$$b_i^{\phi}(x) = \gamma b_i(x) + \int_{(0,\infty)} \left(\int_{\{y\in D:\|y-x\|\leqslant 1\}} (y_i - x_i)\, p(s; x, y)\, \mathrm{d}y \right) \nu(\mathrm{d}s),$$
$$\Pi^{\phi}(x, \mathrm{d}y) = \int_{(0,\infty)} p(s; x, y)\nu(\mathrm{d}s)\, \mathrm{d}y,$$
$$k^{\phi}(x) = \gamma k(x) + \int_{(0,\infty)} P_s\big(x, \{\Delta\}\big)\nu(\mathrm{d}s),$$

where

$$P_s\big(x, \{\Delta\}\big) = 1 - P_s(x, D) = 1 - \int_D p(s; x, y)\,\mathrm{d}y$$

is the probability for the original process $\hat{X}$ to end up in the cemetery state Δ by time s.

Thus, Bochner's subordination of a diffusion process scales the diffusion matrix with the constant γ, the subordinated process acquires a jump component with the Lévy measure with the Lévy density $\int_{(0,\infty)} p(s; x, y)\nu(\mathrm{d}s)$ determined by the Lévy measure of the subordinator and the transition density of the original diffusion process, the killing rate is modified by scaling the original killing rate with γ and adding an additional term $\int_{(0,\infty)} P_s(x, \{\Delta\})\nu(\mathrm{d}s)$ determined by the killing probability of the original process and the Lévy measure of the subordinator, and the drift is scaled with γ, as well as acquires an additional term due to regularization.

Suppose $\gamma > 0$ and, without loss of generality, set $\gamma = 1$. Subordination adds the jump component so that the process $\hat{X}^{\phi}$ is now a jump-diffusion process with the same diffusion component as the original process $\hat{X}$ plus jumps with the (generally state-dependent) Lévy density. If the original process is a diffusion with killing, then the time-changed process is a jump-diffusion process with killing with the modified killing rate. Thus, the subordination procedure allows us to add jumps to any pure diffusion model. If $\gamma = 0$, then the time changed process is a pure jump process.

If the original process is a Lévy process, then the subordinated process is also a Lévy process with state-independent local characteristics. In this case, Theorem 6.2 specializes to Theorem 30.1 of Sato (1999, p. 196) on the Lévy characteristics of a subordinated Lévy process. In particular, many Lévy processes popular in finance (such as VG, CGMY, NIG) can be obtained from standard Brownian motion with drift by subordination with an appropriate subordinator.

The key observation for our purposes here is that when the original process is not a Brownian motion but a more general space-inhomogeneous diffusion process, then the time-changed process is a either a jump-diffusion or a pure jump process with *state-dependent* Lévy density. This allows us to introduce jumps and default in many prominent asset pricing models, such as CIR, CEV, etc. The results are Markov processes with diffusion, jumps, and killing (discounting and default) with state-dependent local characteristics (as opposed to space-homogeneous Lévy processes).

6.3 Subordinating symmetric Markov processes and spectral expansions

So far we have considered subordinating a general diffusion process. We now study subordination of *symmetric* Markov processes and semigroups. We have the following key result.

Theorem 6.3. *Let $\hat{X}$ be a symmetric Markov process with lifetime $\hat{\zeta}$ and symmetry measure m and T a subordinator with Laplace exponent ϕ. The time-changed process $\hat{X}^{\phi}$ is also m-symmetric.*

Proof. The proof follows from Theorem 6.1 with $\mathbf{B} = \mathcal{H} := L^2(D, m)$. From Eq. (6.1):

$$\big(\mathcal{P}_t^{\phi} f, g\big) = \int\limits_{[0,\infty)} (\mathcal{P}_s f, g)\pi_t(\mathrm{d}s) = \int_{[0,\infty)} (f, \mathcal{P}_s g)\pi_t(\mathrm{d}s) = \big(f, \mathcal{P}_t^{\phi} g\big),$$
$$f, g \in \mathcal{H}.$$

□

Then the infinitesimal generator $\mathcal{G}^{\phi}$ is self-adjoint in $\mathcal{H}$ and we obtain its spectral representation and the spectral representation of the subordinated semigroup from that for the process $\hat{X}$ and the subordinator ϕ (see also Okura, 2002 and Chen and Song, 2005a, 2005b for subordination of symmetric Markov processes).

Theorem 6.4. *Let $\hat{X}$, T, $\hat{X}^{\phi}$ be the processes defined previously. We have*:

$$\mathcal{P}_t^{\phi} f = e^{t\mathcal{G}^{\phi}} f = \int\limits_{[0,\infty)} e^{-t\phi(\lambda)} E(\mathrm{d}\lambda) f, \quad f \in \mathcal{H},$$

$$\mathcal{G}^{\phi} f = -\phi(-\mathcal{G}) f = -\int\limits_{[0,\infty)} \phi(\lambda) E(\mathrm{d}\lambda) f, \quad f \in \mathrm{Dom}(\mathcal{G}^{\phi}),$$

$$\mathrm{Dom}\big(\mathcal{G}^{\phi}\big) = \left\{ f \in \mathcal{H} : \int\limits_{[0,\infty)} \phi^2(\lambda)\big(E(\mathrm{d}\lambda) f, f\big) < \infty \right\}.$$

Proof. From Theorem 6.1, Eq. (6.1), the Spectral Representation Theorems 2.1 and 2.2, Eq. (2.5), we have:

$$\mathcal{P}_t^{\phi} f = \int\limits_{[0,\infty)} (\mathcal{P}_s f)\pi_t(\mathrm{d}s) = \int_{[0,\infty)} \left(\int\limits_{[0,\infty)} e^{-\lambda s} E(\mathrm{d}\lambda) f \right) \pi_t(\mathrm{d}s)$$
$$= \int\limits_{[0,\infty)} \left(\int\limits_{[0,\infty)} e^{-\lambda s} \pi_t(\mathrm{d}s) \right) E(\mathrm{d}\lambda) f = \int\limits_{[0,\infty)} e^{-t\phi(\lambda)} E(\mathrm{d}\lambda) f,$$

and similarly for the generator. □

The consequences of this theorem for asset pricing are profound. For any symmetric pricing semigroup for which we know its spectral representation, we can construct a new family of tractable asset pricing models by subordinating it with Lévy subordinators. The resulting pricing semigroup is again

analytically tractable, as long as we know the characteristic exponent $\phi(\lambda)$ of the subordinator in closed form. In particular, this allows us to introduce jumps and default in all the diffusion processes and corresponding asset pricing models considered in Section 4 and summarized in Table 1, such as CIR, CEV, etc. The results are general Markov processes with diffusion, jump, and killing (discounting and default) with state-dependent local characteristics (as opposed to space-homogeneous Lévy processes) that are nevertheless analytically tractable. The local Lévy characteristics of the subordinated process are given by Theorem 6.2.

As an immediate corollary, if the spectrum of the original semigroup is purely discrete, the spectrum of the subordinated semigroup is also purely discrete and the eigenfunction expansion reads:

$$\mathcal{P}_t^{\phi} f = \sum_n e^{-t\phi(\lambda_n)} c_n \varphi_n,$$

where $c_n = (f, \varphi_n)$ are the expansion coefficients of the payoff $f \in \mathcal{H}$ in the eigenfunction basis $\{\varphi_n\}$. The crucial observation is that the subordinated process $\hat{X}^{\phi}$ shares the eigenfunctions $\varphi_n(x)$ with the original process $\hat{X}$, and the eigenvalues of the (negative of) the infinitesimal generator of the subordinated process are

$$-\mathcal{G}^{\phi}\varphi_n = \lambda_n^{\phi}\varphi_n, \quad \lambda_n^{\phi} = \phi(\lambda_n),$$

where λ_n are the eigenvalues of $-\mathcal{G}$ and $\phi(\lambda)$ is the Laplace exponent of the subordinator T. Hence, for the subordinated semigroup we have:

$$\mathcal{P}_t^{\phi}\varphi_n = e^{-\lambda_n^{\phi} t}\varphi_n.$$

Thus, if we know the eigenfunction expansion of the original process $\hat{X}$ (the original pricing semigroup), we also know it for the subordinated process. This makes the spectral representation very convenient for time changes. This was already observed by S. Bochner (1948) in his pioneering paper introducing the concept of time changes for stochastic processes.

6.4 *Stochastic volatility via time changes*

One can also take the time change to be the integral of another positive process:

$$T_t = \int_0^t V_u \, \mathrm{d}u.$$

As long as the Laplace transform

$$\mathbb{E}\left[e^{-\lambda \int_0^t V_u \, \mathrm{d}u}\right]$$

is known analytically and the process V is independent of the process X, the time changed process $Y_t := X_{T_t}$ is analytically tractable. This idea is used in Carr et al. (2003) and Carr and Wu (2004) to build financial models based on time changed Lévy processes (see the Chapter by Wu in this volume for a survey), and in Carr, Linetsky and Mendoza (2007) to build financial models based on time changed Markov processes. In the asset pricing applications, the rate of time change V_u can be interpreted as stochastic volatility of the time-changed process Y_t. In particular, when X is Brownian motion and V is a CIR process, the time change leads to Heston's stochastic volatility model. Carr, Linetsky and Mendoza (2007) construct more general models with stochastic volatility, such as CEV and JDCEV with stochastic volatility.

7 Conclusion

In this Chapter we surveyed the spectral expansion approach to the valuation of derivatives when the underlying state follows a symmetric Markov process. In the Markovian framework, the key object is the pricing operator mapping (possibly defaultable) future payments into present values. The pricing operators indexed by time form a pricing semigroup in an appropriate payoff space, which can be interpreted as the transition semigroup of the Markov process with killing at the rate equal to the default-free interest rate plus default intensity. In applications it is important to have a tool kit of analytically tractable Markov processes with known transition semigroups that lead to closed-form pricing formulas for derivative assets. When the Markov process is symmetric in the sense that there is a measure m on the state space D and the semigroup is symmetric in the Hilbert space $L^2(D, m)$, we apply the Spectral Representation Theorem to obtain spectral representations for the semigroup and value functions of derivative assets. In this Chapter we surveyed the spectral method in general, as well as those classes of symmetric Markov processes for which the spectral representation can be obtained in closed form, thus generating closed-form solutions to derivative pricing problems. This Chapter supplies a tool kit of analytically tractable Markovian models that can be adapted to a wide range of financial engineering applications, as well as a framework to solve new problems.

References

Abramowitz, M., Stegun, I.A. (1972). *Handbook of Mathematical Functions*. Dover, New York.

Ahn, D.-H., Gao, B. (1999). A parametric nonlinear model of term structure dynamics. *Review of Financial Studies* 12, 721–762.

Ait-Sahalia, Y., Hansen, L.P., Scheinkman, J.A. (2004). Operator methods for continuous-time Markov processes. In: Ait-Sahalia, Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*, Elsevier, Amsterdam, in press.

Albanese, C., Lawi, S. (2005). Laplace transforms for integrals of stochastic processes. *Markov Processes and Related Fields*, in press.

Albanese, C., Kuznetsov, A. (2004). Unifying the three volatility models. *Risk* 17 (3), 94–98.
Albanese, C., Kuznetsov, A. (2005). Discretization schemes for subordinated processes. *Mathematical Finance*, in press.
Albanese, C., Campolieti, G., Carr, P., Lipton, A. (2001). Black–Scholes goes hypergeometric. *Risk* (December), 99–103.
Alili, L., Patie, P., Pedersen, J. (2005). Representations of the first hitting time density of an Ornstein–Uhlenbeck process. *Stochastic Models* 21 (4).
Amrein, W.O., Hinz, A.M., Pearson, D.B. (Eds.) (2005). *Sturm–Liouville Theory*. Birkhäuser, Basel.
Andersen, L., Andreasen, J. (2000). Volatility skews and extensions of the LIBOR market model. *Applied Mathematical Finance* 7 (1), 1–32.
Anker, J.-P., Ostellari, P. (2004). The heat kernel on symmetric spaces. In: Gindikin, S. (Ed.), *Lie Groups and Symmetric Spaces: In Memory of F.I. Karpelevich*. In: *American Mathematical Society Transl. Ser. 2, vol. 210*. Amer. Math. Soc., pp. 27–46.
Anker, J.-P., Bougerol, Ph., Jeulin, T. (2002). The infinite Brownian loop on a symmetric space. *Revista Matematica Iberoamericana* 18, 41–97.
Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus*. Cambridge Univ. Press, Cambridge.
Beaglehole, D.R., Tenney, M. (1992). A non-linear equilibrium model of the term structure of interest rates: Corrections and additions. *Journal of Financial Economics* 32, 345–354.
Bertoin, J. (1996). *Lévy Processes*. Cambridge Univ. Press, Cambridge.
Bibby, B.M., Jacobsen, M., Sorensen, M. (2004). Estimating functions for discretely sampled diffusion-type models. In: Ait-Sahalia, Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*, Elsevier, Amsterdam. In press.
Bielecki, T., Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*. Springer, Berlin.
Black, F. (1995). Interest rates as options. *Journal of Finance* 50, 1371–1376.
Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–659.
Bochner, S. (1948). Diffusion equation and stochastic processes. *Proceedings of the National Academy of Sciences of the United States of America* 35, 368–370.
Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. Univ. of California Press.
Borodin, A.N., Salminen, P. (2002). *Handbook of Brownian Motion*, second ed. Birkhäuser, Boston.
Borodin, A.N., Salminen, P. (2004). On some exponential integral functionals of BM(μ) and BES(3). *Zapiski Nauchnyx Seminarov (POMI)* 311, 51–78.
Boyarchenko, N., Levendorskiy, S. (2006). The eigenfunction expansion method in multi-factor quadratic term structure models. *Mathematical Finance*, in press.
Brennan, M., Schwartz, E. (1979). A continuous-time approach to the pricing of bonds. *Journal of Banking and Finance* 3, 133–155.
Buchholz, H. (1969). *The Confluent Hypergeometric Function*. Springer, Berlin.
Carr, P., Linetsky, V. (2006). A jump-to-default extended constant elasticity of variance model: An application of Bessel processes. *Finance and Stochastics* 10 (3), 303–330.
Carr, P., Wu, L. (2004). Time-changed Levy processes and option pricing. *Journal of Financial Economics* 71 (1), 113–141.
Carr, P., Geman, H., Madan, D.B., Yor, M. (2003). Stochastic volatility for Levy processes. *Mathematical Finance* 13 (3), 345–382.
Carr, P., Linetsky, V., Mendoza, R. (2007). Time changed Markov processes in credit equity modeling. Working paper.
Chen, L., Filipovic, D., Poor, H.V. (2004). Quadratic term structure models for risk-free and defaultable rates. *Mathematical Finance* 14, 515–536.
Chen, Z.Q., Song, R. (2005a). Two-sided eigenvalue estimates for subordinate processes in domains. *Journal of Functional Analysis* 226, 90–113.
Chen, Z.Q., Song, R. (2005b). Spectral properties of subordinate processes in domains. Preprint.
Chung, K.L., Zhao, Z. (1995). *From Brownian Motion to Schrödinger's Equation*. Springer, Berlin.
Coddington, E., Levinson, N. (1955). *Theory of Ordinary Differential Equations*. McGraw–Hill, New York.

Cox, J.C. (1975). Notes on option pricing I: Constant elasticity of variance diffusions. Working paper. Stanford University. (Reprinted in *Journal of Portfolio Management* 22 (1996) 15–17.)
Cox, J.C., Ross, S. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3, 145–166.
Cox, J.C., Ingersoll, J.E., Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica* 53, 385–407.
Davies, E.B. (1980). *One-Parameter Semigroups*. Academic Press, London.
Davydov, D., Linetsky, V. (2001). The valuation and hedging of barrier and lookback options under the CEV process. *Management Science* 47, 949–965.
Davydov, D., Linetsky, V. (2003). Pricing options on scalar diffusions: An eigenfunction expansion approach. *Operations Research* 51, 185–209.
Decamps, M., Goovaerts, M., Schoutens, W. (2006). A self-exciting threshold term structure model. *International Journal of Theoretical and Applied Finance*, in press.
Delbaen, F., Shirakawa, H. (2002a). An interest rate model with upper and lower bounds. *Financial Engineering and the Japanese Markets* 9 (3–4), 191–209.
Delbaen, F., Shirakawa, H. (2002b). A note on option pricing for constant elasticity of variance model. *Financial Engineering and the Japanese Markets* 9 (2), 85–99.
DeLong, D.M. (1981). Crossing Probabilities for a Square-root Boundary by a Bessel Process. *Communication in Statistics Theory and Methods A* 10 (21), 2197–2213.
Demuth, M., van Casteren, J.A. (2000). *Stochastic Spectral Theory for Self-adjoint Feller Operators*. Birkhäuser, Basel.
Donati-Martin, C., Ghomrasni, R., Yor, M. (2001). On certain Markov processes attached to exponential functionals of Brownian motion: Applications to Asian options. *Revista Matemática Iberoamericana* 17 (1), 179–193.
Duffie, D., Singleton, K. (2003). *Credit Risk*. Princeton Univ. Press, Princeton, NJ.
Dufresne, D. (1989). Weak convergence of random growth processes with applications to insurance. *Insurance: Mathematics and Economics* 8, 187–201.
Dufresne, D. (1990). The distribution of a perpetuity, with applications to risk theory and pension funding. *Scandinavian Actuarial Journal* 1990, 39–79.
Dunford, N., Schwartz, J. (1963). *Linear Operators Part II: Spectral Theory (Self-Adjoint Operators in Hilbert Spaces)*. Wiley, NJ.
Ethier, S.N., Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.
Erdelyi, A. (1953). *Higher Transcendental Functions, vol. II*. MacGraw–Hill, New York.
Everitt, W.N. (2005). A Catalog of Sturm–Liouville differential equations. In: Amrein, W.O., Hinz, A.M., Pearson, D.B. (Eds.), *Sturm–Liouville Theory*. Birkhäuser, Basel, pp. 271–331.
Feynman, R. (1948). Space–time approach to non-relativistic quantum mechanics. *Review of Modern Physics* 20, 367–387.
Feller, W. (1951). Two singular diffusion problems. *Annals of Mathematics* 54, 173–182.
Florens, J.-P., Renault, E., Touzi, N. (1998). Testing for embeddability by stationary reversible continuous-time Markov processes. *Econometric Theory* 14, 744–769.
Fukushima, M., Oshima, Y., Takeda, M. (1994). *Dirichlet Forms and Symmetric Markov Processes*. W. de Gruyter, Berlin.
Fulton, C., Pruess, S., Xie, Y. (1996). The automatic classification of Sturm–Liouville problems. Preprint http://www.mines.edu/fs_home/spruess/papers/class.ps.
Garman, M. (1985). Towards a semigroup pricing theory. *Journal of Finance* 40 (3), 847–861.
Geman, H., Yor, M. (1993). Bessel processes, Asian options and perpetuities. *Mathematical Finance* 3, 349–375.
Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer.
Glazman, I. (1965). *Direct Methods of Qualitative Spectral Analysis of Singular Differential Operators*. (English Translation). Monson, Jerusalem.
Goeing-Jaeschke, A., Yor, M. (2003). A survey and some generalizations of Bessel processes. *Bernoulli* 9, 313–349.
Gorovoi, V., Linetsky, V. (2004). Black's model of interest rates as options, eigenfunction expansions and Japanese interest rates. *Mathematical Finance* 14, 49–78.

Gorovoi, V., Linetsky, V. (2006). Intensity-based mortgage valuation: An analytically tractable model. *Mathematical Finance*, in press.

Grosche, C., Steiner, F. (1998). *Handbook of Feynman Path Integrals*. Springer, Berlin.

Hansen, L.P., Scheinkman, J.A. (2002). Semigroup asset pricing. Working paper.

Hansen, L.P., Scheinkman, J.A., Touzi, N. (1998). Spectral methods for identifying scalar diffusions. *Journal of Econometrics* 86, 1–32.

Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6, 327–343.

Heston S.L. (1997). A simple new formula for options with stochastic volatility. Working paper. Washington University.

Hille, E., Phillips, R.S. (1957). *Functional Analysis and Semigroups*. American Mathematical Society, Providence, RI.

Ingersoll, J. (1996). Valuing foreign exchange rate derivatives with a bounded exchange process. *Review of Derivatives Research* 1 (2), 159–181.

Ito, K., McKean, H. (1974). *Diffusion Processes and their Sample Paths*, second printing. Springer, Berlin.

Jamshidian, F. (1996). Bond, futures and option evaluation in the quadratic interest rate model. *Applied Mathematical Finance* 3, 93–115.

Jeanblanc, M., Pitman, J., Yor, M. (1997). Feynman–Kac formula and decompositions of Brownian paths. *Computational and Applied Mathematics* 16, 27–52.

Kac, M. (1951). On some connections between probability theory and differential and integral equations. In: *Proc. 2nd Berkeley Symposium on Mathematics, Statistics and Probability*. Univ. of California Press, pp. 189–215.

Kac, M. (1959). *Probability and Related Topics in Physical Sciences*. Interscience, New York.

Karlin, S., Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, San Diego.

Kendall, D.G. (1972). Pole-seeking Brownian motion and bird navigation. *Journal of the Royal Statistical Society, Series B* 36, 365–417.

Lando, D. (2004). *Credit Risk Modeling*. Princeton Univ. Press, Princeton, NJ.

Langer, H., Schenk, W.S. (1990). Generalized second-order differential operators, corresponding gap diffusions and superharmonic transformations. *Mathematische Nachrichten* 148, 7–45.

Larsen, K.S., Sorensen, M. (2007). Diffusion models for exchange rates in a target zone. Working paper. University of Copenhagen.

Leippold, M., Wu, L. (2002). Asset pricing under the quadratic class. *Journal of Financial and Quantitative Analysis* 37 (2), 271–295.

Levinson, N. (1951). A simplified proof of the expansion theorem for singular second order differential operators. *Duke Mathematical Journal* 18, 57–71.

Levitan, B.M. (1950). *Expansion in Characteristic Functions of Differential Equations of the Second Order*. Gostekhizdat, Moscow (in Russian).

Levitan, B.M., Sargsjan, I.S. (1975). *Introduction to Spectral Theory*. American Mathematical Society, Providence, RI.

Lewis, A. (1998). Applications of eigenfunction expansions in continuous-time finance. *Mathematical Finance* 8, 349–383.

Lewis, A. (2000). *Option Valuation under Stochastic Volatility*. Finance Press, CA.

Linetsky, V. (2004a). The spectral decomposition of the option value. *International Journal of Theoretical and Applied Finance* 7 (3), 337–384.

Linetsky, V. (2004b). Spectral expansions for Asian (average price) options. *Operations Research* 52, 856–867.

Linetsky, V. (2004c). Lookback options and diffusion hitting times: A spectral expansion approach. *Finance and Stochastics* 8 (3), 373–398.

Linetsky, V. (2004d). The spectral representation of Bessel processes with constant drift: Applications in queueing and finance. *Journal of Applied Probability* 41 (2), 327–344.

Linetsky, V. (2004e). Computing hitting time densities for CIR and OU diffusions: Applications to mean-reverting models. *Journal of Computational Finance* 7 (4), 1–22.

Linetsky, V. (2005). On the transitions densities of reflected diffusions. *Advances in Applied Probability* 37, 1–26.
Linetsky, V. (2006). Pricing equity derivatives subject to bankruptcy. *Mathematical Finance* 16 (2), 255–282.
Lipton, A. (1999). Similarities via self-similarities. *Risk* (September), 101–105.
Lipton, A. (2001). *Mathematical Methods for Foreign Exchange*. World Scientific, Singapore.
Lipton, A. (2002). The volatility smile problem. *Risk* (February), 61–65.
Lipton, A., McGhee, W. (2002). Universal barriers. *Risk* (May), 81–85.
Madan, D., Unal, H. (1998). Pricing the risk of default. *Review of Derivatives Research* 2, 121–160.
McKean, H. (1956). Elementary solutions for certain parabolic partial differential equations. *Transactions of the American Mathematical Society* 82, 519–548.
Merton, R.C. (1973). Theory of rational options pricing. *Bell Journal of Economics and Management Science* 4, 141–183.
Merton, R.C. (1975). An asymptotic theory of growth under uncertainty. *Review of Economic Studies* 42, 375–393.
Miltersen, K., Sandmann, K., Sondermann, D. (1997). Closed-form solutions for term structure derivatives with lognormal interest rates. *Journal of Finance* 52, 409–430.
Morse, P.M. (1929). Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Physical Review* 34, 57–64.
Morse, P.M., Feshbach, H. (1953). *Methods of Theoretical Physics. Part II*. McGraw–Hill.
Natanzon, G.A. (1979). General properties of potentials for which the Schrödinger equation can be solved by means of hypergeometric functions. *Theoretical Mathematical Physics* 38, 146–153.
Nelson, D.B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics* 45, 7–39.
Okura, H. (2002). Recurrence and transience criteria for subordinated symmetric Markov processes. *Forum Mathematicum* 14, 121–146.
Peskir, G. (2006). On the fundamental solution of the Kolmogorov–Shiryaev equation. In: *The Shiryaev Festschrift*. Springer, Berlin, pp. 535–546.
Phillips, R.S. (1952). On the generation of semigroups of linear operators. *Pacific Journal of Mathematics* 2, 343–369.
Pilipović, D. (1998). *Energy Risk*. McGraw–Hill, New York.
Pitman, J.W., Yor, M. (1981). Bessel processes and infinitely divisible laws. In: Williams, D. (Ed.), *Stochastic Integrals*. In: *Lecture Notes in Mathematics*, vol. 851. Springer.
Pitman, J.W., Yor, M. (1982). A decomposition of Bessel bridges. *Zeit. Wahrsch. Geb.* 59, 425–457.
Rady, S. (1997). Option pricing in the presence of natural boundaries and a quadratic diffusion term. *Finance and Stochastics* 1 (4), 331–344.
Rady, S., Sandmann, K. (1994). The direct approach to debt option pricing. *Review of Futures Markets* 13, 461–514.
Reed, M., Simon, B. (1980). *Functional Analysis*. Academic Press, San Diego.
Revuz, D., Yor, M. (1999). *Continuous Martingales and Brownian Motion*, third ed. Springer, Berlin.
Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Univ. Press, Cambridge.
Schachermayer, W., Teichmann, J. (2006). How close are the option pricing formulas of Bachelier and Black–Merton–Scholes? *Mathematical Finance*, in press.
Schoutens, W. (2000). *Stochastic Processes and Orthogonal Polynomials*. Springer, Berlin.
Schroder, M. (1989). Computing the constant elasticity of variance option pricing formula. *Journal of Finance* 44 (March), 211–219.
Shiga, T., Watanabe, S. (1973). Bessel diffusions as a one-parameter family of diffusion processes. *Zeitschrift fur Wahrscheinlichkeitstheorie Verw. Geb.* 27, 37–46.
Shiryaev, A.N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathematics Doklady* 2, 795–799.
Slater, L.J. (1960). *Confluent Hypergeometric Functions*. Cambridge Univ. Press.
Titchmarsh, E.C. (1962). *Eigenfunction Expansions Associated with Second-order Differential Equations*. Clarendon, Oxford.
Vasicek, O.A. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* 5, 177–188.

Watanabe, S. (1975). On time inversion of one-dimensional diffusion processes. *Zeitschrift fur Wahrscheinlichkeitstheorie* 31, 115–124.

Weidmann, J. (1987). *Spectral Theory of Ordinary Differential Operators*. *Lecture Notes in Mathematics*, vol. 1258. Springer, Berlin.

Wong, E. (1964). The construction of A Class of stationary Markov processes. In: Bellman, R. (Ed.), *Sixteenth Symposium in Applied Mathematics – Stochastic Processes in Mathematical Physics and Engineering*. American Mathematical Society, Providence, RI, pp. 264–276.

Yor, M. (1984). On square-root boundaries for Bessel processes, and pole-seeking Brownian motion. In: Truman, A., Williams, D. (Eds.), *Stochastic Analysis and Applications Proceedings*. Swansea 1993. In: *Lecture Notes in Mathematics*, vol. 1095, pp. 100–107.

Yor, M. (2001). *Exponential Functionals of Brownian Motion and Related Processes*. Springer, Berlin.

Zuhlsdorff, C. (2001). The pricing of derivatives on assets with quadratic volatility. *Applied Mathematical Finance* 8 (4), 235–262.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15007-6

Chapter 7

Variational Methods in Derivatives Pricing*

Liming Feng

Department of Industrial and Enterprise Systems Engineering,
University of Illinois at Urbana-Champaign, 117 Transportation Building MC-238,
104 South Mathews Avenue, Urbana, IL 61801, USA
E-mail: fenglm@uiuc.edu

Pavlo Kovalov

Quantitative Risk Management, Inc., 181 West Madison Street, 41st Floor, Chicago,
Illinois 60602, USA
E-mail: pavlo.kovalov@qrm.com

Vadim Linetsky

Department of Industrial Engineering and Management Sciences,
McCormick School of Engineering and Applied Sciences, Northwestern University,
2145 Sheridan Road, Evanston, IL 60208, USA
E-mail: linetsky@iems.northwestern.edu
url: http://users.iems.northwestern.edu/∼linetsky

Michael Marcozzi

Department of Mathematical Sciences, University of Nevada Las Vegas,
4505 Maryland Parkway, Box 454020, Las Vegas, NV 89154-4020, USA
E-mail: marcozzi@unlv.nevada.edu

Abstract

When underlying financial variables follow a Markov jump-diffusion process, the value function of a derivative security satisfies a partial integro-differential equation (PIDE) for European-style exercise or a partial integro-differential variational inequality (PIDVI) for American-style exercise. Unless the Markov process has a special structure, analytical solutions are generally not available, and it is necessary to solve the PIDE or the PIDVI numerically. In this chapter we briefly survey a computational method for the valuation of options in jump-diffusion models based on: (1) converting the PIDE or PIDVI to a variational (weak) form; (2) discretizing the weak formulation spatially by the Galerkin finite element method to obtain a system of ODEs; and (3) integrating the resulting system of ODEs in time. To introduce the method, we start with the basic examples of European, barrier, and American

* This research was supported by the National Science Foundation under grants DMI-0422937 and DMI-0422985.

options in the Black–Scholes–Merton model, then describe the method in the general setting of multi-dimensional jump-diffusion processes, and conclude with a range of examples, including Merton's and Kou's one-dimensional jump-diffusion models, Duffie–Pan–Singleton two-dimensional model with stochastic volatility and jumps in the asset price and its volatility, and multi-asset American options.

1 Introduction

When underlying financial variables follow a Markov jump-diffusion process, the value function of a derivative security satisfies a partial integro-differential equation (PIDE) for European-style exercise or a partial integro-differential variational inequality (PIDVI) for American-style exercise. Unless the Markov process has a special structure (as discussed in the previous chapter), analytical solutions are generally not available, and it is necessary to solve the PIDE or the PIDVI numerically. Numerical solution of initial and boundary value problems for partial differential equations (PDE) of diffusion-convection-reaction type (the type arising in Markov process models when the underlying state variable follows a diffusion process with drift and killing or discounting) on bounded domains is standard in two and three spatial dimensions. Such PDE problems arise in a wide variety of applications in physics, chemistry, and various branches of engineering. A variety of standard (both free and commercial) software implementations are available for this purpose.

However, PDE problems that arise in finance in the context of derivatives pricing in Markov process models have a number of complications: (1) diffusion models often have more than three state variables, resulting in multi-dimensional PDE formulations; (2) Markov process often has a jump component in addition to the diffusion component, resulting in a nonlocal integral term in the evolution equation (making it into a partial *integro*-differential equation (PIDE)); (3) the state space is often an *unbounded* domain in $\mathbb{R}^n$, resulting in PDE and PIDE problems on unbounded domains, which need to be localized to bounded domains in order to be solved numerically; (4) American-style early exercise is often permitted (early exercise in American options, conversion and call features in convertible bonds, etc.), leading to free-boundary problems that can be formulated as partial differential (or integro-differential if jumps are present) variational inequalities (PDVI or PIDVI); (5) payoffs are often nonsmooth (e.g., call and put option payoffs have a kink, digital option payoffs have discontinuities), creating additional challenges for numerical solution methods.

In this chapter we briefly survey a general computational method for the valuation of derivative securities in jump-diffusion models. The method is based on: (1) converting the PIDE or PIDVI to a variational (weak) form; (2) discretizing the weak formulation spatially by the *Galerkin finite element*

method to obtain a system of ODEs (this framework is called the *finite element method-of-lines*); and (3) integrating the resulting system of ODEs in time by applying appropriate time stepping schemes. To introduce the method, we start with the basic examples of European, barrier, and American options in the Black–Scholes–Merton model, then describe the method in the general setting of multi-dimensional jump-diffusion processes, and conclude with a range of examples, including Merton's and Kou's one-dimensional jump-diffusion models, Duffie–Pan–Singleton two-dimensional model with stochastic volatility and jumps in the asset price and its volatility, and multi-asset American options.

Historically, binomial and trinomial trees (lattices) have been the numerical method of choice in financial applications (e.g., Hull, 2006). Their main advantage securing their popularity in financial applications is the ease of implementation for low-dimensional diffusion problems (one or two state variables), as well as the ease of incorporating American-style early exercise via dynamic programming. However, tree methods have serious limitations. Since trinomial trees can be interpreted as first-order fully explicit finite-difference schemes for the pricing PDE, they are only first-order accurate in time and have stability restrictions of the form $\Delta t \leqslant C\Delta x^2$, where Δt is the time step and Δx is the step size of the spatial discretization. This may require one to take prohibitively large number of time steps to converge to reasonable accuracies. This disadvantage becomes particularly serious for multi-dimensional diffusion problems with more than two state variables and/or for jump-diffusion processes (where one also needs to deal with nonlocal integral terms, that present significant challenges for tree methods). In contrast, implicit finite-difference methods avoid the undesirable stability restrictions on the size of the time step relative to the size of the spatial discretization. Furthermore, higher order time stepping schemes are available, such as Crank–Nicolson, backward differentiation formulae based schemes (BDF), etc. Surveys of finite-difference methods in derivatives pricing can be found in Tavella and Randall (2000) and Wilmott et al. (1993).

In this chapter we focus on the *finite element method*. The finite element method is a general technique for the numerical solution of differential equations in science and engineering. The method has its origins in structural engineering in the late 50's and early 60's. It soon became clear that the method was quite general with roots in the variational methods in mathematics introduced in the beginning of the 20th century. Continuing development of the finite element method has resulted in a general purpose method for the numerical solution of partial differential equations.

The basic idea in any numerical method for a differential equation is to discretize a given continuous problem in order to obtain a discrete problem with only finitely many unknowns, which may be solved using a computer. The classical numerical method for partial differential equations is the finite difference method. Here, the discrete problem is obtained by replacing derivatives with difference quotients involving values at finitely many points.

The discretization process using a finite element method is quite different. We start not from the given partial differential equation, but from its equivalent reformulation in terms of a *variational problem*. In order to obtain a discrete problem, we consider solutions consisting only of simple functions. That is, if the solution to the continuous problem u is an element of the infinite-dimensional function space V, we consider a discrete solution $u_h \in V_h$, where the space of simple functions V_h has finite dimension. The hope is that the solution u_h of the discrete problem is a sufficiently good approximation of the solution u of the original partial differential equation. If one chooses $V_h \subset V$, then the discrete problem corresponds to the classical *Ritz–Galerkin method*. The special feature of the finite element method as a particular Ritz–Galerkin method is that the functions in V_h are chosen to be *piecewise polynomial* (in particular, piecewise affine functions in the case of linear finite elements). Specifically, the state space domain of the problem is partitioned into nonoverlapping *elements* (intervals in one dimension, triangles or rectangles in two dimensions, etc.). Thus, the state space domain is covered with a mesh called a *triangulation*. The maximal degree p of the polynomial approximation is chosen (e.g., $p = 1$ for linear finite elements). The space V_h consists of functions of V whose restrictions to the elements are polynomials of degree $\leqslant p$. For derivatives pricing problems, the approximation to the value function $u(t, x)$ solving the pricing PDE or PIDE is written in the form

$$u_h(t, x) = \sum_{i=1}^{m} u_i(t)\phi_{h,i}(x),$$

where $\phi_{h,i}(x)$ are some basis functions in V_h (e.g., piecewise affine functions in the case of the linear finite element method) and $u_i(t)$ are time-dependent coefficients to be determined by numerically solving a system of ODEs, the so-called *finite element method-of-lines*.

The advantage of the finite element method is that complicated geometry of the state space, general boundary conditions, nonlinear and nonlocal equations can all be handled relatively easily in the same general framework. The finite element method is particularly suited for nonlocal integro-differential equations appearing in jump-diffusion models in finance. The finite element method has a functional analytic grounding which provides added reliability and in many cases makes it possible to mathematically analyze and estimate the error in the approximate finite element solution. In addition, the clear structure and versatility of the finite element method make it possible to construct general purpose software for applications. See the monographs Ciarlet (1978), Hundsdorfer and Verwer (2003), Johnson (1987), Larsson and Thomee (2003), Quarteroni and Valli (1997), Thomee (1997, 2001) for general introduction to finite element methods for parabolic problems. A recent monograph Achdou and Pironneau (2005) is devoted to applications of the finite element method to options pricing. In this chapter we give a brief survey of the finite element method in derivatives pricing problems.

2 European and barrier options in the Black–Scholes–Merton model

2.1 PDE formulation

In the Black–Scholes–Merton model, the asset price is assumed to follow a geometric Brownian motion process under an equivalent martingale measure (EMM):

$$S_t = Ke^{X_t}, \quad X_t = x + \mu t + \sigma B_t, \ t \geqslant 0,$$

where $\{X_t, \ t \geqslant 0\}$ is a Brownian motion with drift $\mu = r - q - \sigma^2/2$ and volatility $\sigma > 0$ (here B is a standard Brownian motion), $r \geqslant 0$ is the risk-free interest rate, $q \geqslant 0$ is the continuous dividend yield, and $K > 0$ is some reference asset price level. One typically sets $K = S_0$, the initial asset price at time zero. This corresponds to starting the process X_t at the origin, $X_0 = 0$. Then X_t has the interpretation of the continuously compounded return process net of dividends. Alternatively, when pricing call and put options, it will be convenient for us to set K equal to the strike price of the option contract to be priced. This corresponds to starting the process X_t at

$$X_0 = x = \ln(S_0/K),$$

where K is the strike price. The drift μ is such that after discounting at the risk-free rate the total gains process, including price appreciation and dividends, is a martingale under the EMM. The infinitesimal generator of the Markov process X is

$$\mathcal{G}f(x) = \frac{1}{2}\sigma^2 f_{xx}(x) + \mu f_x(x).$$

Consider a European-style option contract that delivers a payoff $F(S_T)$ at expiration $T > 0$. The payoff function F is assumed to depend on the underlying asset price at expiration. To be specific, we will call the underlying asset a stock. The price of the option at time $t \in [0, T]$ is given by its expected discounted payoff, where the expectation is taken under the EMM:

$$V(t, x) = e^{-r(T-t)}\mathbb{E}_{t,x}\big[\psi(X_T)\big]. \tag{2.1}$$

Here we substituted $S_T = Ke^{X_T}$ and defined the payoff function $\psi(X_T) := F(Ke^{X_T})$ in terms of the variable $X_T = \ln(S_T/K)$. The subscript in the conditional expectation operator $\mathbb{E}_{t,x}$ signifies that at time t the state of the Markov process X is known, $X_t = x$. For call and put options, we set K equal to the strike price, and the payoffs are $\psi_{\text{call}}(x) = K(e^x - 1)^+$ and $\psi_{\text{put}}(x) = K(1 - e^x)^+$, where $x^+ \equiv \max\{x, 0\}$.

The value function $V = V(t, x)$ can be characterized as the solution of the following *fundamental pricing PDE* (the backward Kolmogorov equation for the expectation (2.1)):

$$V_t + \mathcal{G}V - rV = 0, \quad t \in [0, T),$$

with the terminal (payoff) condition $V(T, x) = \psi(x)$. For future convenience, we transform the terminal value problem into an initial value problem (IVP) by defining $U(t, x) := V(T-t, x)$ and include the discounting in the definition of the operator:

$$\mathcal{A}U := \frac{1}{2}\sigma^2 U_{xx} + \mu U_x - rU.$$

Then we need to solve the following parabolic PDE:

$$U_t - \mathcal{A}U = 0, \quad t \in (0, T],$$

with the initial condition $U(0, x) = \psi(x)$.

We will also be interested in *knock-out options*. A knock-out option contract is canceled (knocked out) if the underlying state variable exits a pre-specified open domain Ω in the state space. Let τ be the first hitting time of the boundary $\partial\Omega$ (here we consider diffusions, and so the process will always hit the boundary due to the path continuity). The knock-out option is then canceled (declared null and void) at time τ if $\tau \leqslant T$, where T is the option expiration. Otherwise, the option holder receives the payoff at expiration. In some cases the option holder receives a *rebate* $R(X_\tau)$ at time τ if $\tau \leqslant T$. It can be constant or depend on the state of the underlying process at time τ (more generally, it can also depend on time, i.e., the rebate is $R(\tau, X_\tau)$ if the process hits $\partial\Omega$ at time τ and is in state $X_\tau \in \partial\Omega$ at τ, but for simplicity we assume it only depends on the state and is independent of time). In particular, there are six types of knock-out options: calls and puts with lower barriers, calls and puts with upper barriers, and calls and puts with both lower and upper barriers. A down-and-out call (put) delivers a call (put) payoff at T if the stock does not fall to or below a lower barrier L, $0 < L < S_0$. An up-and-out call (put) delivers a call (put) payoff prior to and including expiration T if the stock does not increase to or above an upper barrier U, $S_0 < U < \infty$, prior to and including expiration T. A double-barrier call (put) delivers a call (put) payoff at T if the stock does not exit from an open interval (L, U) prior to and including expiration T. Here $0 < L < S_0 < U < \infty$ are lower and upper barriers. In terms of the process $X_t = \ln(S_t/K)$, the lower and upper barriers are $\underline{x} = \ln(L/K)$ and $\bar{x} = \ln(U/K)$. In terms of the PDE, the value function of the knock-out option satisfies the following *knock-out condition with rebate*:

$$U(t, x) = R(x), \quad x \in \partial\Omega, \ t \in [0, T],$$

where $\partial\Omega = \{\underline{x}\}$, $\partial\Omega = \{\bar{x}\}$, and $\partial\Omega = \{\underline{x}, \bar{x}\}$ for down-and-out, up-and-out, and double-barrier options, respectively. If there is no rebate, $R = 0$, then the value function vanishes on the boundary $\partial\Omega$.

2.2 Localization to bounded domains

For double-barrier options the state space is bounded, $\Omega = (\underline{x}, \bar{x})$. For European and single-barrier calls and puts, the state space Ω is unbounded

($\Omega = (\underline{x}, \infty)$ for down-and-out options, $\Omega = (-\infty, \bar{x})$ for up-and-out options, and $\Omega = \mathbb{R}$ for European options without barriers). In order to develop numerical approximations, we localize the original problem by considering an increasing exhausting sequence of bounded open domains $\{\Omega_k\}$ such that $\Omega_k \subset \Omega_{k+1}$ and $\bigcup_k \Omega_k = \Omega$ (i.e., the sequence exhausts the state space). Then the value function U of the original problem on the unbounded domain Ω is realized as the limit of a sequence of functions U_k which solve the PDEs on bounded domains Ω_k:

$$U_{k,t} - \mathcal{A}U_k = 0, \quad t \in (0, T], \ x \in \Omega_k \tag{2.2}$$

with the initial condition:

$$U_k(0, x) = \psi(x), \quad x \in \Omega_k. \tag{2.3}$$

An *artificial boundary condition* is imposed on the boundary $\partial\Omega_k$:

$$U_k(t, x) = R(x), \quad x \in \partial\Omega_k, \ t \in [0, T], \tag{2.4}$$

where $R(x)$ is the *artificial rebate*. In other words, we approximate the original option contract with a contract that knocks out when the process exits an open bounded domain Ω_k and pays a rebate $R(X_\tau)$ at the first hitting time τ of the boundary $\partial\Omega_k$. The economics of the problem often suggests an appropriate choice of the artificial boundary condition. For European options, the payoff function provides a reasonable choice for the artificial boundary condition: $R(x) = \psi(x)$ for $x \in \partial\Omega_k$, where $\partial\Omega_k = \{\underline{x}_k, \bar{x}_k\}$ and $\underline{x}_k \to -\infty$, $\bar{x}_k \to \infty$ as $k \to \infty$. For down-and-out put options with the lower barrier L and without rebate, the vanishing boundary condition $R(x) = 0$, $x \in \partial\Omega_k$, provides a reasonable choice. Here $\partial\Omega_k = \{\underline{x}, \bar{x}_k\}$, where $\underline{x} = \ln(L/K)$ is determined by the contractual lower barrier L and $\bar{x}_k \to \infty$ as $k \to \infty$ is the artificial upper barrier. For $x = \underline{x}$, this is the knock-out condition specified by the option contract. For $x = \bar{x}_k$, it provides a reasonable choice for the artificial boundary condition, since the value function of the down-and-out put rapidly decreases towards zero for high stock prices. Other types of knock-out options are treated similarly. For the localized problem, we have (see Bensoussan and Lions, 1984 for general results on localization to bounded domains):

$$\max_{t\in[0,T]} \left\| U(t, \cdot) - U_k(t, \cdot) \right\|_{L^\infty(G)} \to 0 \quad \text{as } k \to \infty$$

for any fixed compact set $G \subset \Omega_1$. The set G is referred to as the *approximation domain*, where we are interested in the value function U. The bounded domain Ω_k is referred to as the *computational domain* (see Marcozzi, 2001, for details). Kangro and Nicolaides (2000), Matache et al. (2004), and Hilber et al. (2005) show that the localization error decays exponentially in the size of the computational domain in the Black–Scholes setting, in the Lévy process setting, and in the stochastic volatility setting, respectively.

Other choices for artificial boundary conditions can be used to localize the problem to a bounded domain. In the present paper, we use the payoff function as the artificial rebate in the artificial boundary condition for European

options, down-and-out calls, and up-and-out puts ($R(x) = \psi(x)$) and vanishing boundary conditions for double-barrier call and put options, down-and-out puts, and up-and-out calls ($R(x) = 0$) (for simplicity in this paper we assume that the original knock-out contracts do not pay any contractual rebates). In what follows, we take a bounded computational domain Ω_k as given and solve the PDE on Ω_k with the artificial boundary condition on $\partial\Omega_k$ selected as discussed above. We drop the index k to lighten our notation.

2.3 *Variational formulation*

We consider a variational (weak) formulation of the problem (2.2)–(2.4) on a given bounded domain Ω. This PDE may have a nonhomogeneous boundary condition (with artificial rebate $R(x) \neq 0$, $x \in \partial\Omega$). To simplify future development, we homogenize it as follows. The artificial rebate $R(x)$ is defined on $\partial\Omega$. We extend it to $\overline{\Omega}$ and also denote the extended version by $R(x)$, $x \in \overline{\Omega}$. In this paper we either have $R(x) = 0$ or $R(x) = \psi(x)$ for $x \in \partial\Omega$. We extend it so that $R(x) = 0$ or $R(x) = \psi(x)$ for $x \in \overline{\Omega}$, respectively. Let $u := U - R$ (for European and American options, u can be interpreted as the excess option premium over the payoff; for European options it can be negative, as the option exercise is not allowed until expiration).

A *variational* (*weak*) *formulation* of the PDE problem is obtained by considering a space of test functions square-integrable on Ω, with their (weak) first derivatives square-integrable on Ω, and vanishing on the boundary $\partial\Omega$ (the Sobolev space $H_0^1(\overline{\Omega}) := \{f \in L^2(\overline{\Omega})\colon f_x \in L^2(\overline{\Omega}), f|_{\partial\Omega} = 0\}$). Multiplying the PDE with a test function $v = v(x)$, integrating over Ω, and integrating by parts, we arrive at the variational (weak) formulation of the PDE:

$$(u_t, v) + a(u, v) + a(R, v) = 0, \tag{2.5}$$

$$\big(u(0,\cdot), v\big) = (\psi - R, v), \tag{2.6}$$

where $(u, v) = \int_{\Omega} u(x)v(x)\,\mathrm{d}x$ is the inner product in $L^2(\Omega)$, and the bilinear form $a(\cdot,\cdot)$ is defined by

$$a(u, v) = \frac{1}{2}\sigma^2 \int_{\underline{x}}^{\bar{x}} u_x v_x\,\mathrm{d}x - \mu \int_{\underline{x}}^{\bar{x}} u_x v\,\mathrm{d}x + r \int_{\underline{x}}^{\bar{x}} uv\,\mathrm{d}x.$$

To solve the variational formulation, we seek a function $u = u(t, x)$ in an appropriate function space such that (2.5)–(2.6) hold for any test function $v \in H_0^1(\overline{\Omega})$. The solution u vanishes on the boundary $\partial\Omega$. The value function U of the problem with the inhomogeneous boundary conditions is then obtained by $U = R + u$. More details on the variational formulation of parabolic PDEs associated with diffusion processes can be found in Quarteroni and Valli (1997) and Thomee (1997), where the relevant functional analytic background can be found.

2.4 Galerkin finite element approximation

We now consider a spatial discretization of the variational formulation (2.5)–(2.6) by the Galerkin finite element method (see Ciarlet, 1978; Larsson and Thomee, 2003; Quarteroni and Valli, 1997; and Thomee, 1997, for textbook treatments of the finite element method). Consider a one-dimensional problem on a bounded domain $\Omega = (\underline{x}, \bar{x})$. We divide the interval $\overline{\Omega} = [\underline{x}, \bar{x}]$ into $m+1$ subintervals (*elements*), each having length of $h = (\bar{x} - \underline{x})/(m+1)$. Let $x_i = \underline{x} + i\,h$, $i = 0, \ldots, m+1$, be the nodes in $[\underline{x}, \bar{x}]$. We define the following piecewise linear finite element basis functions $\{\phi_{h,i}(x)\}_{i=1}^{m}$:

$$\phi_{h,i}(x) = \begin{cases} (x - x_{i-1})/h, & x_{i-1} \leqslant x \leqslant x_i, \\ (x_{i+1} - x)/h, & x_i < x \leqslant x_{i+1}, \\ 0, & x \notin [x_{i-1}, x_{i+1}]. \end{cases}$$

The ith basis function $\phi_{h,i}(x)$ is a hat function equal to one at the node x_i and zero at all other nodes, $\phi_i(x_j) = \delta_{ij}$, where $\delta_{ij} = 1$ (0) if $i = j$ ($i \neq j$). It is illustrated in Figure 1. If we define the hat function $\phi(x) := (x+1)\mathbf{1}_{\{-1 \leqslant x \leqslant 0\}} + (1-x)\mathbf{1}_{\{0 < x \leqslant 1\}}$, then $\phi_{h,i}(x) = \phi((x - x_i)/h)$. More generally, we can define $\phi_{h,i}(x)$ for all integer $i \in \mathbb{Z}$. The nodes $x_0 = \underline{x}$ and $x_{m+1} = \bar{x}$ are on the boundary $\partial\Omega$ and the nodes x_i with $i < 0$ or $i > m+1$ are outside of $\overline{\Omega} = [\underline{x}, \bar{x}]$ (we will need to consider nodes outside of $[\underline{x}, \bar{x}]$ when dealing with jumps).

We look for a finite-element approximation u_h to the solution u of the variational formulation (2.5)–(2.6) as a linear combination of the finite element basis functions with time-dependent coefficients:

$$u_h(t, x) = \sum_{i=1}^{m} u_i(t)\phi_{h,i}(x), \quad t \in [0, T]. \tag{2.7}$$

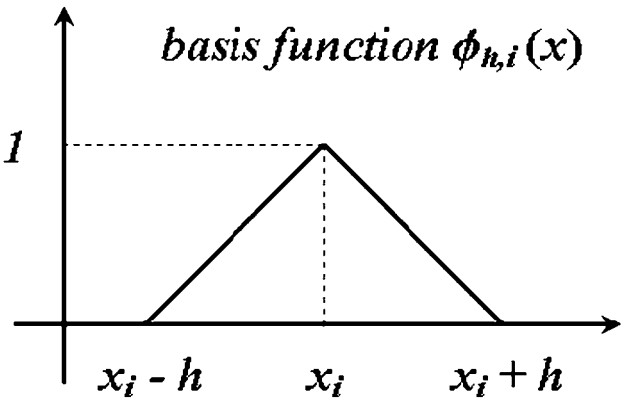


Figure 1. Hat function.

Note that, by construction, u_h vanishes on the boundary $\partial\Omega$ (since the basis functions vanish on the boundary). Thus, we look for an approximation u_h to the true solution u in the finite element basis space V_h spanned by the finite element basis functions $\{\phi_{h,i}\}_{i=1}^m$.

Denote by $\mathbf{u}(t) = (u_1(t), \ldots, u_m(t))^\top$ the m-dimensional vector of time-dependent coefficients to be determined. Substituting (2.7) into (2.5)–(2.6) and letting the test function v in (2.5)–(2.6) run through the set of all basis functions $\{\phi_{h,i}\}_{i=1}^m$ (i.e., we also approximate the test function by $v_h(x) = \sum_{i=1}^m v_i \phi_{h,i}(x)$), we obtain the following m-dimensional system of ODEs:

$$\mathbb{M}\mathbf{u}'(t) + \mathbb{A}\mathbf{u}(t) + \mathbf{F} = 0, \quad t \in (0, T], \tag{2.8}$$

with the initial condition:

$$\mathbb{M}\mathbf{u}(0) = \mathbf{C}. \tag{2.9}$$

Here $\mathbf{u}'(t) = (u_1'(t), \ldots, u_m'(t))^\top$, $u_i'(t) \equiv \mathrm{d}u_i(t)/\mathrm{d}t$, $\mathbb{M} = (m_{ij})_{i,j=1}^m$, where

$$m_{ij} = (\phi_j, \phi_i),$$

$\mathbb{A} = (a_{ij})_{i,j=1}^m$, where

$$a_{ij} = a(\phi_j, \phi_i),$$

$\mathbf{C} = (c_1, \ldots, c_m)^\top$, where

$$c_i = (\psi - R, \phi_i),$$

and $\mathbf{F} = (F_1, \ldots, F_m)^\top$, where

$$F_i = a(R, \phi_i)$$

(to lighten notation, we omit the index h in $\phi_{h,i}$; recall that $a(\cdot, \cdot)$ is the previously defined bilinear form). This ODE system is referred to as a *semi-discretization* of the variational problem (spatially discrete and continuous in time). The pricing problem is reduced to the integration of this ODE system. This is referred to as the *finite element method-of-lines* (MOL) ("lines" is a metaphor for the lines (x_i, t), $t \geqslant 0$ in the (x, t)-domain, x_i fixed, along which the approximations to the PDE solution are studied; see Hundsdorfer and Verwer, 2003). Due to the origins of the finite element method in structural engineering, $\mathbb{M}$ is referred to as the *mass matrix*, $\mathbb{A}$ as the *stiffness matrix*, and $\mathbf{F}$ as the *load vector*. For each t, on a bounded domain Ω the semi-discrete finite element approximation is known to be second order accurate in the spatial step size h:

$$\left\| u_h(t, \cdot) - u(t, \cdot) \right\| \leqslant Ch^2,$$

both in the $L^2(\Omega)$ norm and in the $L^\infty(\Omega)$ norm. Maximum norm error estimates available in the finite element method are particularly relevant in financial engineering as they give the worst case pricing error estimates.

The mass matrix $\mathbb{M}$ and the stiffness matrix $\mathbb{A}$ can be easily computed in closed form in this model with constant coefficients (numerical quadrature, such as the Gaussian quadrature, is used in more general models). For any $i, j \in \mathbb{Z}$, from the definition of the bilinear form $a(\cdot, \cdot)$, we have:

$$a(\phi_i, \phi_i) = a_0 = \frac{2}{3}rh + \frac{1}{h}\sigma^2,$$
$$a(\phi_i, \phi_{i\pm1}) = a_{\mp1} = \pm\frac{1}{2}\mu + \frac{1}{6}rh - \frac{1}{2h}\sigma^2,$$

and $a(\phi_i, \phi_j) = 0$ for $|i - j| > 1$. Moreover,

$$(\phi_i, \phi_i) = \frac{2}{3}h, \qquad (\phi_i, \phi_{i\pm1}) = \frac{1}{6}h, \qquad (\phi_i, \phi_j) = 0, \quad |i - j| > 1.$$

Therefore, both $\mathbb{A}$ and $\mathbb{M}$ are tri-diagonal $m \times m$ matrices with constant diagonals:

$$\mathbb{A} = \begin{pmatrix} a_0 & a_1 & & \\ a_{-1} & a_0 & \ddots & \\ & \ddots & \ddots & a_1 \\ & & a_{-1} & a_0 \end{pmatrix}, \qquad \mathbb{M} = \frac{h}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix}.$$

The load vector $\mathbf{F}$, $F_i = a(R, \phi_i)$, $i = 1, \ldots, m$, can be computed analytically in this model for $R = \psi$, where ψ is a call or put payoff, $K(e^x - 1)^+$ or $K(1 - e^x)^+$, respectively. Generally, it can be computed numerically (e.g., by Gaussian quadrature). Finally, the initial condition $\mathbb{M}\mathbf{u}(0) = \mathbf{C}$ with the vector $\mathbf{C}$, $c_i = (\psi - R, \phi_i)$, is treated as follows. For $R = \psi$, $\mathbf{C}$ vanishes identically, and we have a vanishing initial condition $\mathbf{u}(0) = 0$. For $R = 0$, $c_i = (\psi, \phi_i)$ can be computed analytically for simple payoffs or numerically by Gaussian quadrature in general. Then the initial vector $\mathbf{u}(0)$ is obtained by solving $\mathbb{M}\mathbf{u}(0) = \mathbf{C}$.

2.5 *Integrating the ODE system*

We have reduced the option pricing problem to the solution of the ODE system (2.8)–(2.9). This ODE system needs to be integrated numerically. We observe that $\mathbb{M} \sim O(h)$ and $\mathbb{A} \sim O(h^{-1})$. Hence, the system (2.8) is *stiff*. In particular, the term $\mathbb{A}\mathbf{u}$ resulting from the discretization of the diffusion operator generates stiffness. For stiff systems, explicit schemes are only conditionally stable and may require prohibitively small time steps when h is small via stability restrictions of the form $\Delta t \leqslant Ch^2$.

The simplest time discretization is provided by the so-called θ-scheme. Divide the time interval $[0, T]$ into N time steps, each having length $k = T/N$, and with the nodes $t_i = ik$, $i = 0, 1, \ldots, N$. Define $\mathbf{u}^i := \mathbf{u}(t_i)$, $i = 0, 1, \ldots, N$. Then the θ-scheme starts with the initial condition $\mathbb{M}\mathbf{u}^0 = \mathbf{C}$ (or $\mathbf{u}^0 = 0$ if $\mathbf{C} = 0$) and marches forward according to:

$$(\mathbb{M} + \theta k\mathbb{A})\mathbf{u}^i = \big(\mathbb{M} - (1 - \theta)k\mathbb{A}\big)\mathbf{u}^{i-1} - k\mathbf{F}, \quad i = 1, \ldots, N. \tag{2.10}$$

For $\theta = 0$, this is the fully explicit forward Euler scheme (that corresponds to replacing the time derivative at time t_i with the finite difference $\mathbf{u}'(t_i) \to k^{-1}(\mathbf{u}^i - \mathbf{u}^{i-1})$ in the ODE and evaluating the rest of the terms at the *previous* time level t_{i-1}). For $\theta = 1$, this is the fully implicit backward Euler scheme (that corresponds to replacing the time derivative with the finite difference $\mathbf{u}'(t_i) \to k^{-1}(\mathbf{u}^i - \mathbf{u}^{i-1})$ in the ODE and evaluating the rest of the terms at the *current* time level t_i). At each step, the linear system (2.10) is solved to determine the m-dimensional vector $\mathbf{u}^i$. The matrices $\mathbb{M}$ and $\mathbb{A}$ are tri-diagonal in this case, and the linear system can be solved directly by LU decomposition. For all $\theta \geqslant 1/2$, the θ-scheme is unconditionally stable. It is first-order accurate in time for $\theta \neq 1/2$, and it is second-order accurate in time for $\theta = 1/2$. The latter choice is the well-known Crank–Nicolson scheme popular in financial engineering. Its advantage is that it is second-order accurate in time and is very easy to implement. Its disadvantages are that it may generate spurious oscillations in the computed numerical solution if the time step exceeds twice the maximum stable explicit time step and, furthermore, the expected quadratic convergence may not be realized when initial conditions are not smooth (see Zvan et al., 1998a and Pooley et al., 2003 for discussions of these issues and suggested remedies).

Another way to attain second order accuracy in time is to approximate the time derivative in the ODE by the second-order backward differentiation formula. This results in the second-order backward differentiation (BDF) scheme:

$$\left(\frac{3}{2}\mathbb{M} + k\mathbb{A}\right)\mathbf{u}^i = 2\mathbb{M}\mathbf{u}^{i-1} - \frac{1}{2}\mathbb{M}\mathbf{u}^{i-2} - k\mathbf{F}, \quad i = 2, \ldots, N$$

($\mathbf{u}^1$ needed to launch the second-order BDF scheme to compute $\mathbf{u}^2$ is obtained by the backward Euler time stepping).

Generally, there are several major classes of time stepping schemes used to integrate systems of ODEs resulting from semi-discretization of parabolic PDEs in the finite element method-of-lines framework: schemes based on Runge–Kutta methods, schemes based on high order backward differentiation formulae (BDF), and schemes based on Richardson extrapolation. References include Deuflhard and Bornemann (2002), Hairer and Wanner (1996), Hundsdorfer and Verwer (2003), Quarteroni and Valli (1997), and Thomee (1997). There are a number of software packages available (both freely and commercially) based on these schemes that include adaptive automatic time step selection and adaptive automatic integration order selection in some packages. In Section 5 we present some example computations with the variable step-size and variable-order BDF-based package SUNDIALS (SUite of Nonlinear and DIfferential/ALgebraic equation Solvers) available from the Lawrence Livermore National Laboratory (http://www.llnl.gov/CASC/sundials/). The details of the adaptive order and step size selection can be found in the SUNDIALS documentation Hindmarsh et al. (2005, 2006).

We now present a simple to implement high order time stepping scheme based on applying a Richardson-type extrapolation procedure to the backward Euler scheme. Some computational examples are given in Section 5. The error of the backward Euler scheme is known to have an asymptotic expansion in the powers of the time step k:

$$\mathbf{u}(T) - \mathbf{u}^N = \mathbf{e}_1(T)k + \mathbf{e}_2(T)k^2 + \cdots. \tag{2.11}$$

Generally, the Euler scheme for the ODE $u' = G(u)$ has an asymptotic error expansion (2.11) if the right-hand side G is smooth. In our case G is linear and, hence, (2.11) holds. This asymptotic error expansion suggests applying extrapolation to cancel lower order terms in the error expansion and to increase the order of the scheme. References on extrapolation methods for stiff ODE systems include Deuflhard (1985), Deuflhard and Bornemann (2002, Section 6.4.2), and Hairer and Wanner (1996, Section IV.9). For applications of the extrapolation scheme to option pricing see Feng and Linetsky (2006b).

We now describe the extrapolation scheme based on the backward Euler scheme. We need to integrate the ODE system on the interval $[0, T]$. Assume that a *basic stepsize* H ($H = T/N$) and an *extrapolation stage number* $s \geqslant 1$ are given. Then one constructs a sequence of approximations to $\mathbf{u}(H)$ (at time H) using the backward Euler scheme with *internal stepsizes* $k_i = H/n_i$, $i = 1, 2, \ldots, s+1$, where $\{n_i\}_{i=1}^{s+1}$ is the *step number sequence*. For the Euler scheme, the harmonic sequence $\{1, 2, \ldots, s+1\}$ or the sequence $\{2, 3, \ldots, s+2\}$ are often used. Denoting the approximation obtained at time H (after one basic step) with internal stepsize k_i by $\mathbf{T}_{i,1} = \mathbf{u}(H; k_i)$, the *extrapolation tableau* is constructed as follows:

$$\mathbf{T}_{i,j} = \mathbf{T}_{i,j-1} + \frac{\mathbf{T}_{i,j-1} - \mathbf{T}_{i-1,j-1}}{(n_i/n_{i-j+1}) - 1}, \quad i = 2, \ldots, s+1,\ j = 2, \ldots, i. \tag{2.12}$$

The extrapolation tableau can be graphically depicted as follows:

$$\begin{array}{llll} \mathbf{T}_{1,1} & & & \\ \mathbf{T}_{2,1} & \mathbf{T}_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \mathbf{T}_{s+1,1} & \mathbf{T}_{s+1,2} & \cdots & \mathbf{T}_{s+1,s+1}. \end{array}$$

The value $\mathbf{T}_{s+1,s+1}$ after s extrapolation stages is taken as the approximation to $\mathbf{u}(H)$ and is used as the starting point to launch a new basic integration step over the next interval $[H, 2H]$. The procedure is continued in this way for N basic steps until the approximation of $\mathbf{u}(T)$ at time $T = NH$ is obtained. After one basic step, we have the following error estimate:

$$\mathbf{u}(H) - \mathbf{T}_{s+1,s+1} = O(k_1 k_2 \cdots k_{s+1}) = O\big(H^{s+1}/(n_1 n_2 \cdots n_{s+1})\big).$$

For the step number sequence $\{2, 3, \ldots, s+2\}$ we have in particular:

$$\mathbf{u}(H) - \mathbf{T}_{s+1,s+1} = O\big(H^{s+1}/(s+2)!\big). \tag{2.13}$$

The total number of time steps required to compute $\mathbf{T}_{s+1,s+1}$ (the total number of times the linear system (2.10) needs to be solved) is

$$\mathcal{N}_s = (s+4)(s+1)/2.$$

Recall that the factorial can be well approximated by $n! \approx \sqrt{2\pi(1+1/(6n))} \times n^{n+1/2}e^{-n}$ (a refinement of the Stirling formula). This approximation is very accurate even for small n (e.g., for $n = 2$ this gives 1.9974). For the sequence $\{2, 3, \ldots, s+2\}$, for fixed s the error of the extrapolation scheme is thus:

$$\begin{aligned}\mathbf{u}(H) - \mathbf{T}_{s+1,s+1} = O\big(\big\{2\pi\big(1+1/\big(6(s+2)\big)\big)\big\}^{-1/2} \\ \times \exp\big\{-(s+5/2)\ln(s+2) \\ + (s+1)(1+\ln H) + 1\big\}\big).\end{aligned}$$

To get some intuition on the dependence of the error on the number of time steps, recalling that the total number of time steps needed to integrate the ODE on the interval $[0, H]$ is $\mathcal{N}_s = (s+1)(s+4)/2$, we write the error estimate as follows:

$$\mathbf{u}(H) - \mathbf{T}_{s+1,s+1} = O\big(e^{-c\sqrt{\mathcal{N}_s}\ln\mathcal{N}_s}\big). \tag{2.14}$$

This suggests that the error decreases as $e^{-c\sqrt{\mathcal{N}_s}\ln\mathcal{N}_s}$ with the increasing number of time steps $\mathcal{N}_s$. We stress that the argument above is not rigorous. The error estimate only states that *for fixed* s and small $H \to 0$ the error is asymptotically $O(H^{s+1}/(s+2)!)$. Generally, it does not say anything about the behavior of the error with increasing s, as the constant C in the estimate $CH^{s+1}/(s+2)!$ may depend on s and, hence, on $\mathcal{N}_s$. If a hypothesis that the constant can be made independent of s (or increases slowly with s) holds, then Eq. (2.14) would, in fact, provide an error estimate in terms of the number of time steps. Unfortunately, it appears difficult to prove this hypothesis. However, in our numerical experiments with option pricing applications we do observe the rate of convergence suggested by the heuristic (2.14).

For a fixed basic step H, the total number of time steps needed to integrate the ODE on the time interval $[0, T]$ is: $\mathcal{N}_{T,H,s} = N(s+4)(s+1)/2$, where $N = T/H$. Due to rapid convergence of the extrapolation method, N and s are typically small in option pricing applications, resulting in the small total number of time steps $\mathcal{N}$ required to achieve desired accuracy.

At each time step, we need to solve the linear system (2.10). In one-dimensional models, we have a tri-diagonal system. This can be solved using the LU decomposition with forward/backward substitution in $O(m)$ operations for a system of order m (recall that here m is the number of elements in the finite element discretization). Hence, this time stepping scheme takes $O(\mathcal{N}m)$ floating point operations, where $\mathcal{N}$ is the total number of time steps.

So far we have taken the basic step size H and the number of extrapolation stages s as given. We now discuss selection of H and s in practice.

First suppose H is fixed, and select a local error tolerance $\epsilon > 0$. After j extrapolation stages, $\mathbf{T}_{j+1,j+1}$ approximates $\mathbf{u}(H)$. The error estimate is: $\mathcal{E}_j := \|\mathbf{T}_{j+1,j+1} - \mathbf{T}_{j+1,j}\|_{L^\infty}$ (the so-called *subdiagonal error estimate*, see Hairer and Wanner, 1996, p. 140). After each extrapolation stage with $j \geqslant 2$, we compare the estimated error $\mathcal{E}_j$ with the preceding error $\mathcal{E}_{j-1}$ and with the desired local error tolerance ϵ. Whenever $\mathcal{E}_j \leqslant \epsilon$ for some $j \leqslant s_{\max}$, we accept $\mathbf{T}_{j+1,j+1}$ as the approximation to $\mathbf{u}(H)$ and move on to compute the solution over the next basic step $[H, 2H]$ starting with $\mathbf{u}(H) = \mathbf{T}_{j+1,\,j+1}$ determined at the previous step. Alternatively, whenever $\mathcal{E}_j \geqslant \mathcal{E}_{j-1}$ (i.e., increasing extrapolation depth does not result in further error reduction) or if the desired error tolerance is not achieved in $s_{\max}$ stages, we restart the computation of the step with a smaller H, say $H_{\text{new}} = H_{\text{old}}/2$. In our numerical experiments we selected $s_{\max} = 10$, so that if the desired error tolerance is not achieved in ten extrapolation stages, we reduce the basic step size. This simple procedure allows us to select the basic step size H and the extrapolation stage s adaptively. The only user-specified parameter in addition to the desired error tolerance is the initial basic step size H. If the initial H is too large relative to the error tolerance, the adaptive procedure will reduce it and restart the computation with smaller H. If H is selected too small relative to the desired error tolerance, more time steps than necessary will be computed. In our computational experiments with options pricing problems, $H = 0.5$ year has proven adequate as a starting basic step size for error tolerances up to 10^{-5} (without the need to reduce the basic step size in most cases). For problems with maturities less than six months we set $H = T$. For faster computations with less precision (e.g., error tolerances up to 10^{-3}), the basic step $H = 1$ or even longer can be used as a starting step. Computational examples are given in Section 5 and in Feng and Linetsky (2006b). More sophisticated fully adaptive schemes that allow adaptive time step selection and adaptive extrapolation depth selection can be found in Deuflhard (1985), Deuflhard and Bornemann (2002, Section 6.4.2), and Hairer and Wanner (1996, Section IV.9).

3 American options in the Black–Scholes–Merton model

3.1 Optimal stopping, variational inequality, localization, discretization, linear complementarity problem

An American-style derivative security has a payoff $\psi = \psi(x)$, and the option holder can exercise at any time between the contract inception at time $t = 0$ and expiration $T > 0$ and receive the payoff. In general, the payoff can depend both on the time of exercise and the state of the underlying process at the time of exercise, $\psi = \psi(t, x)$, but to simplify notation we assume that it only depends on the state and not on time. Assuming the option holder follows a value-maximizing exercise strategy, the value function of the American option is given by the value function of the optimal stopping problem (Bensoussan,

1984; Karatzas, 1988):

$$V(t,x) = \sup_{\theta \in \Theta_{t,T}} \mathbb{E}_{t,x}\left[e^{-r(\theta-t)}\psi(X_\theta)\right], \tag{3.1}$$

where the supremum is taken over the set of all stopping times θ taking values in $[t, T]$, denoted by $\Theta_{t,T}$.

Based on the fundamental work of Bensoussan and Lions (1982, 1984) on the variational inequality formulation of optimal stopping problems for Markov processes and Glowinski et al. (1981) on numerical methods for variational inequalities, Jaillet et al. (1990) show that the value function of the American-style problem can be determined as the unique solution of the variational inequality (see also Lamberton and Lapeyre, 1996 for a textbook treatment):

$$\begin{aligned}
&V_t + \mathcal{A}V \leqslant 0, && t \in [0,T),\ x \in \mathbb{R},\\
&V \geqslant \psi, && t \in [0,T),\ x \in \mathbb{R},\\
&(V_t + \mathcal{A}V)\cdot(V-\psi) = 0, && t \in [0,T),\ x \in \mathbb{R},
\end{aligned}$$

with the terminal condition

$$V(T,x) = \psi(x), \quad x \in \mathbb{R}.$$

Introducing a time value (excess premium over the payoff) function $u(t,x) = V(T-t,x) - \psi(x)$, which is always positive for American options, the variational inequality is transformed into:

$$u_t - \mathcal{A}u - \mathcal{A}\psi \geqslant 0, \qquad t \in (0,T],\ x \in \mathbb{R}, \tag{3.2a}$$

$$u \geqslant 0, \qquad t \in (0,T],\ x \in \mathbb{R}, \tag{3.2b}$$

$$(u_t - \mathcal{A}u - \mathcal{A}\psi)\cdot u = 0, \quad t \in (0,T],\ x \in \mathbb{R}, \tag{3.2c}$$

with the homogeneous initial condition

$$u(0,x) = 0, \quad x \in \mathbb{R}. \tag{3.2d}$$

First we localize the problem (3.2) to a bounded domain $\Omega = (\underline{x}, \bar{x})$ by assuming that (3.2) hold on Ω and the function vanishes on the boundary,

$$u(t,x) = 0, \quad x \in \{\underline{x}, \bar{x}\},\ t \in [0,T].$$

This corresponds to the following artificial boundary condition for the value function: $V(t,x) = \psi(x)$, $x \in \{\underline{x}, \bar{x}\}$, $t \in [0,T]$.

Consider a space of test functions $v \in H_0^1(\overline{\Omega})$ (the previously introduced Sobolev space of functions vanishing on the boundary) and such that $v \geqslant 0$. By multiplying (3.2a) with v and integrating on Ω, we obtain:

$$(u_t, v) + a(u,v) + a(\psi,v) \geqslant 0, \quad \forall v \geqslant 0,\ v \in H_0^1(\overline{\Omega}), \tag{3.3}$$

where $a(u, v)$ is the previously defined bilinear form and (u, v) is the inner product in $L^2(\overline{\Omega})$. Integrating (3.2c) on Ω, we obtain:

$$(u_t, u) + a(u, u) + a(\psi, u) = 0. \tag{3.4}$$

Subtracting (3.4) from (3.3), we obtain:

$$(u_t, v - u) + a(u, v - u) + a(\psi, v - u) \geqslant 0, \quad \forall v \geqslant 0,\ v \in H_0^1(\overline{\Omega}). \tag{3.5}$$

We seek a solution $u = u(t, x)$ of (3.5) that vanishes on the boundary $\partial\Omega$, with the vanishing initial condition, $u(0, x) = 0$, and such that the nonnegativity constraint $u \geqslant 0$ is satisfied on $\overline{\Omega}$.

We now consider a fully discrete finite element approximation. We seek an approximate solution in the form (2.7). We also approximate test functions in the finite element basis, $v_h(x) = \sum_{i=1}^{m} v_i \phi_{h,i}(x)$, and denote by $\mathbf{v}$ the m-dimensional vector of nonnegative coefficients v_i. We discretize time by the fully implicit backward Euler method. Divide the time interval $[0, T]$ into N time steps, each having length $k = T/N$, and with the nodes $t_n = nk$, $n = 0, 1, \ldots, N$. Define $\mathbf{u}^n := \mathbf{u}(t_n)$, $n = 0, 1, \ldots, N$. Then the full discretization of Eq. (3.5) starts with the initial condition $\mathbf{u}^0 = 0$ and marches forward according to:

$$\begin{aligned} &\mathbf{u}^n \geqslant 0, \quad \left(\mathbf{v} - \mathbf{u}^n\right)^\top \left((\mathbb{M} + k\mathbb{A})\mathbf{u}^n - \mathbb{M}\mathbf{u}^{n-1} + k\mathbf{F}\right) \geqslant 0, \\ &\quad \forall \mathbf{v} \geqslant 0,\ n = 1, \ldots, N, \end{aligned} \tag{3.6}$$

where at each time step we need to determine an m-dimensional vector $\mathbf{u}^n$ such that all its elements are nonnegative, $\mathbf{u}^n \geqslant 0$, and (3.6) is satisfied for every m-dimensional vector $\mathbf{v}$ with all its elements nonnegative, $\mathbf{v}^n \geqslant 0$. The load vector $\mathbf{F}$ is given by $F_i = a(\psi, \phi_{h,i})$.

This is equivalent to solving a *linear complementarity problem* (LCP) at each step (one for each $n = 1, 2, \ldots, N$): determine an m-dimensional vector $\mathbf{u}^n$ such that (see Jaillet et al., 1990; Lamberton and Lapeyre, 1996 for the LCP formulation of the American option problem discretized by finite differences):

$$\left(\mathbf{u}^n\right)^\top \left((\mathbb{M} + k\mathbb{A})\mathbf{u}^n - \mathbb{M}\mathbf{u}^{n-1} + k\mathbf{F}\right) = 0, \tag{3.7a}$$

$$(\mathbb{M} + k\mathbb{A})\mathbf{u}^n - \mathbb{M}\mathbf{u}^{n-1} + k\mathbf{F} \geqslant 0, \tag{3.7b}$$

$$\mathbf{u}^n \geqslant 0. \tag{3.7c}$$

This LCP can be solved by, e.g., the projected successive over-relaxation (PSOR) algorithm (Cryer, 1971; Cottle et al., 1992). A textbook treatment of PSOR in the context of American option pricing in the finite difference framework can be found in Wilmott et al. (1993). While it is straightforward to implement, a disadvantage of this approach based on the discretization of the variational inequality and solution of the resulting LCP is that this approach is inherently only first-order accurate in time as it is based on the implicit Euler scheme.

3.2 Penalization and the nonlinear system of ODEs

An alternative approach is to apply the *penalization technique* to the variational inequality. Penalty methods approximate the variational inequality by a sequence of nonlinear PDEs (see Bensoussan and Lions, 1982, 1984; Glowinski et al., 1981; and Glowinski, 1984). The advantage of this approach is that, utilizing the finite element method-of-lines framework, one can apply available ODE solvers to obtain adaptive high-order integration in time. In the context of pricing American options, penalty methods have been applied by Zvan et al. (1998b), Forsyth and Vetzal (2002), d'Halluin et al. (2004) in the finite-difference framework and Sapariuc et al. (2004) and Kovalov and Linetsky (2007), Kovalov et al. (2007) in the finite-element framework.

We now briefly discuss the penalized formulation. The original variational inequality (3.2) on the bounded domain Ω is approximated with the nonlinear PDE problem with the penalty term approximating the constraint $u \geqslant 0$ ($x^- \equiv \max\{-x, 0\}$):

$$\frac{\partial u_\epsilon}{\partial t} - \mathcal{A}u_\epsilon - \frac{1}{\epsilon}(u_\epsilon)^- - \mathcal{A}\psi = 0, \quad t \in (0, T],\ x \in \Omega, \tag{3.8a}$$

$$u_\epsilon(t, x) = 0, \qquad x \in \partial\Omega,\ t \in (0, T], \tag{3.8b}$$

$$u_\epsilon(0, x) = 0, \qquad x \in \overline{\Omega}. \tag{3.8c}$$

The nonlinear penalty term $\frac{1}{\epsilon}(u_\epsilon)^-$ approximates the action of the early exercise constraint $u \geqslant 0$. According to Friedman (1976, Chapter 16), Bensoussan and Lions (1982, Chapter 3, Section 4), and Bensoussan and Lions (1984, Theorem 8.3, p. 155), the solution u_ϵ of the nonlinear penalized PDE problem (4.1)–(4.2) converges to the solution of the variational inequality (3.2) as $\epsilon \to 0$. In particular, the following penalization error estimate holds for the penalty approximation (e.g., Boman, 2001; Sapariuc et al., 2004):

$$\max_{t\in[0,T]} \left\| u_\epsilon(t, \cdot) - u(t, \cdot) \right\|_{L^\infty(\Omega)} \leqslant C\epsilon \tag{3.9}$$

for some $C > 0$ independent of ϵ, and

$$u_\epsilon \geqslant -C\epsilon.$$

The intuition behind this result is as follows. While the solution u of the variational inequality is constrained to stay nonnegative (which corresponds to the value function $V \geqslant \psi$, or not less than the payoff that can be obtained by early exercise), the solution of the nonlinear penalized PDE u_ϵ can fall below zero. However, while when $u_\epsilon \geqslant 0$ the penalty term vanishes, when $u_\epsilon < 0$, the penalty term $\frac{1}{\epsilon}(u_\epsilon)^-$ has a positive value that rapidly increases as the solution falls below zero, forcing the solution of the PDE back above zero. The smaller the value of ϵ, the larger the value of the coefficient $1/\epsilon$ in the penalty term and the more closely the penalty term approximates the action of the early exercise constraint. The penalized solution u_ϵ can fall below zero, but $u_\epsilon \geqslant -C\epsilon$ for some constant $C > 0$ independent of ϵ.

The variational formulation of this nonlinear PDE in the bounded domain Ω is

$$\left(\frac{\partial u_\epsilon}{\partial t}, v\right) + a(u_\epsilon, v) + \big(\pi_\epsilon(u_\epsilon), v\big) + a(\psi, v) = 0, \quad \forall v \in H_0^1(\overline{\Omega}),$$

with the initial condition (3.8c). Here

$$\pi_\epsilon(u_\epsilon) = -\frac{1}{\epsilon}(u_\epsilon)^- \tag{3.10}$$

is the penalty term. The semi-discrete finite element formulation then leads to the nonlinear ODE:

$$\mathbb{M}\mathbf{u}_\epsilon'(t) + \mathbb{A}\mathbf{u}_\epsilon(t) + \mathbb{M}\pi_\epsilon(\mathbf{u}_\epsilon)(t) + \mathbf{F} = 0, \quad t \in (0, T], \tag{3.11}$$

with the initial condition $\mathbf{u}_\epsilon(0) = 0$, where the nonlinear penalty vector $\pi_\epsilon(\mathbf{u}_\epsilon) = (\pi_\epsilon(u_{\epsilon,1}), \ldots, \pi_\epsilon(u_{\epsilon,m}))^\top$ has elements

$$\pi_\epsilon(u_{\epsilon,i})(t) = -\frac{1}{\epsilon}\big(u_{\epsilon,i}(t)\big)^-, \quad i = 1, 2, \ldots, m.$$

The problem is then reduced to integrating the nonlinear ODE system (3.11) with the vanishing initial condition. Note that due to the small parameter ϵ in the denominator, the penalty term adds stiffness to the ODE system. We thus have a tradeoff. From the estimate (3.9), the error of the penalty approximation is of the order ϵ. To reduce the approximation error, we wish to select small ϵ. However, excessively small ϵ makes the system stiff, and temporal integration methods for stiff ODEs need to be employed. If the penalty term is treated implicitly, then a nonlinear equation needs to be solved at each step. This can be accomplished by Newton iterations. In our numerical examples in Section 5, we employ the freely available SUNDIALS software package implementing adaptive variable order and variable step size BDF time stepping with a built-in Newton iterative solver. As an alternative to the fully implicit treatment of the penalty term, a linear-implicit scheme can be used, such as linear-implicit (or semi-implicit) extrapolation schemes for stiff systems (Hairer and Wanner, 1996; Deuflhard and Bornemann, 2002).

While the specific functional form of the penalty term in (3.10) is commonly used in the literature on the penalty approximation of variational inequalities and its applications in finance (e.g., Bensoussan and Lions, 1982, 1984; Forsyth and Vetzal, 2002; Friedman, 1976; Glowinski et al., 1981; Glowinski, 1984; Marcozzi, 2001; Sapariuc et al., 2004; Zvan et al., 1998a, 1998b), more general forms of the penalty term can be considered (see Glowinski, 1984 for general results about penalty approximations of variational inequalities). In fact, the penalty term $\frac{1}{\epsilon}(u_\epsilon)^-$ has a discontinuous first derivative with respect to u_ϵ. In the numerical solution, one needs to use Newton-type iterations to solve a nonlinear system of algebraic equations resulting from the discretization of the PDE. The discontinuity in the Jacobian of this system stemming from the discontinuity in the derivative of the penalty term with respect to u_ϵ

presents some computational challenges, as nonsmooth Newton-type iterative schemes for nonlinear systems with discontinuous Jacobians need to be used (see, e.g., Forsyth and Vetzal, 2002). An alternative is to consider more general penalty terms of the form $(\frac{1}{\epsilon}u^-)^p$ for some $p \geqslant 1$, which we call power-p penalty terms (Kovalov et al., 2007). Taking $p > 1$ restores the continuity of the derivative of the penalty term with respect to u, and standard Newton iterations with continuous Jacobian can be used. In the numerical experiments in this paper we take $p = 2$ and verify that the estimate (3.9) holds in this case as well. Thus, we consider a smoothed penalty term of the form

$$\pi_\epsilon(u_\epsilon) = -\left(\frac{1}{\epsilon}(u_\epsilon)^-\right)^p, \tag{3.12}$$

that leads to the penalty vector in the ODE system:

$$\pi_\epsilon(u_{\epsilon,i}) = -\left(\frac{1}{\epsilon}(u_{\epsilon,i})^-\right)^p, \tag{3.13}$$

where $p \geqslant 2$. This smoothed penalty term has $p-1$ continuous derivatives and has a continuous Jacobian in particular. This improves numerical performance of high-order time integrators.

4 General multi-dimensional jump-diffusion models

4.1 General formulations for European, barrier, and American derivatives

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$ be a complete filtered probability space with a filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$ satisfying the usual hypothesis that the filtration is right-continuous and every $\mathcal{F}_t$ contains all the $\mathbb{P}$-null sets of $\mathcal{F}$. Let $\{B_t,\ t \geqslant 0\}$ be an $\mathcal{F}_t$-adapted standard Brownian motion on $\mathbb{R}^n$ and $p(\mathrm{d}t\,\mathrm{d}z)$ an $\mathcal{F}_t$-adapted Poisson random measure on $[0,\infty) \times \mathbb{R}^n$ with intensity measure $q(\mathrm{d}t\,\mathrm{d}z) = \Lambda\,\mathrm{d}t\,F(\mathrm{d}z)$, such that $\Lambda \geqslant 0$ is a Poisson jump arrival intensity and F is a jump size (magnitude) probability measure on $\mathbb{R}^n$ for which $F\{0\} = 0$. That is, we consider that jumps arrive according to a Poisson process $\{N_t,\ t \geqslant 0\}$ with intensity Λ. On arrival, the jump size is an $\mathbb{R}^n$-valued random variable with probability distribution F (independent of the Poisson process N and Brownian motion B). For a time interval $[t_1, t_2]$, a Borel set $A \subset \mathbb{R}^n$ and $\omega \in \Omega$, the Poisson random measure $p(\omega; [t_1, t_2] \times A) = \#\{\text{jumps of size} \in A \text{ arriving during } [t_1, t_2]\}$. The Poisson random measure is a counting measure of a compound Poisson process with jump arrival intensity Λ and jump size distribution F.

We model the underlying economic uncertainty affecting financial variables such as asset prices, interest rates, foreign exchange rates, etc. in the risk-neutral economy as an $\mathcal{F}_t$-adapted Markov jump-diffusion process $\{X_t,\ t \geqslant 0\}$ in $\mathbb{R}^n$ with cádlág (right continuous with left limits) sample paths solving the

stochastic differential equation (SDE) with jumps

$$\mathrm{d}X_t = b(t, X_{t-})\,\mathrm{d}t + \sigma(t, X_{t-})\,\mathrm{d}B_t + \mathrm{d}J_t, \tag{4.1a}$$

with the deterministic initial condition $X_0 = x \in \mathbb{R}^n$, and jump component

$$J_t = \int_0^t \int_{\mathbb{R}^n} \gamma(s, X_{s-}, z)\, p(\mathrm{d}s\,\mathrm{d}z). \tag{4.1b}$$

If a Poisson jump arrives at time τ and has size z, then the jump-diffusion process X experiences a jump of size $\Delta X_\tau = X_\tau - X_{\tau-} = \gamma(\tau, X_{\tau-}, z)$. In particular, (4.1b) has the form:

$$J_t = \sum_{\tau_n \leqslant t} \Delta X_{\tau_n} = \sum_{\tau_n \leqslant t} \gamma(\tau_n, X_{\tau_n-}, Z_n),$$

where τ_n are jump arrival times and Z_n are jump sizes (i.i.d. random variables with distribution F, independent of N and B). We note that it is possible that $\gamma(\tau, X_{\tau-}, z) = 0$ and X has a zero jump even though the Poisson process experiences a jump. The intensity measure of the jump process J_t is $M(t, x; \mathrm{d}z)\,\mathrm{d}t$, where

$$M(t, x; A) = \Lambda F\big\{z:\ \gamma(t, x, z) \in A,\ \gamma(t, x, z) \neq 0\big\},$$

for any Borel set $A \subset \mathbb{R}^n$, while the arrival rate of nonzero jumps is

$$\lambda(t, x) = M\big(t, x; \mathbb{R}^n\big) = \Lambda F\big\{z:\ \gamma(t, x, z) \neq 0\big\} \leqslant \Lambda.$$

On arrival of a nonzero jump at time t when $X_{t-} = x$, the jump size of the process X is an $\mathbb{R}^n$-random variable with distribution

$$\mu(t, x; A) = \frac{M(t, x; A)}{\lambda(t, x)}.$$

We assume that there exists a unique strong solution of the SDE (4.1) that is an $\mathcal{F}_t$-adapted cádlág Markov jump-diffusion process in $\mathbb{R}^n$. In particular, it is sufficient to suppose that the drift vector $b : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$, the dispersion (volatility) matrix $\sigma : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$, and the jump-size vector $\gamma : [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ are Borel-measurable and satisfy the respective linear growth and local Lipschitz conditions (T denotes the matrix transpose):

$$\operatorname{tr} \sigma \cdot \sigma^T(t, x) + \big|b(t, x)\big|^2 + \int_{\mathbb{R}^n} \big|\gamma(t, x, z)\big|^2 F(\mathrm{d}z) \leqslant K\big(1 + |x|^2\big),$$

and for each $N \in \mathbb{N}$ there exists a constant C_N such that

$$\begin{aligned}&\operatorname{tr}\big(\sigma(t, x) - \sigma(t, y)\big)\big(\sigma(t, x) - \sigma(t, y)\big)^T + \big|b(t, x) - b(t, y)\big|^2 \\ &\quad + \int_{\mathbb{R}^n} \big|\gamma(t, x, z) - \gamma(t, y, z)\big|^2 F(\mathrm{d}z) \leqslant C_N |x - y|^2,\end{aligned}$$

for all $|x|, |y| \leqslant N$ (see Bensoussan and Lions, 1984, Theorem 6.2, p. 276, and Jacod and Shiryaev, 2003, Theorem 2.32, p. 158).

A European-style derivative security has a promised payoff $\psi = \psi(x)$ at expiration $T > 0$. Its present value at time $t \in [0, T]$ when the process is in state x, $X_t = x$, is given by its expected discounted payoff (under the EMM):

$$V(t, x) = \mathbb{E}_{t,x}\big[e^{-\int_t^T r(s,X_s)\,\mathrm{d}s}\psi(X_T)\big], \tag{4.2}$$

where the discount rate (default-free or defaultable) is assumed to be a function of the underlying state variable, $r = r(t, x) \geqslant 0$ (this framework includes the possibility of default governed by some default intensity, in which case r is the default-free rate plus the default intensity; for simplicity here we assume no recovery in default).

A knock-out derivative security in addition specifies an open domain $D \subset \mathbb{R}^n$ in the state space, and the contract pays a rebate $R(\tau, X_\tau)$ if the underlying state variable exits D prior to and including expiration:

$$\begin{aligned} V(t, x) = {} & \mathbb{E}_{t,x}\big[e^{-\int_t^T r(s,X_s)\,\mathrm{d}s}\psi(X_T)\mathbf{1}_{\{\tau > T\}}\big] \\ & + \mathbb{E}_{t,x}\big[e^{-\int_t^\tau r(s,X_s)\,\mathrm{d}s}R(\tau, X_\tau)\mathbf{1}_{\{\tau \leqslant T\}}\big], \end{aligned} \tag{4.3}$$

where $\tau := \inf\{u \in [t, T]\colon X_u \notin D\}$ is the first exit time from D (we assume that $X_t = x \in D$), and the rebate $R(\tau, X_\tau)$ generally depends on the knock-out time τ and the state of the process $X_\tau \in D^c$ at τ. Note that, in contrast to the pure diffusion case, it is possible that the process jumps right through the boundary and into the interior of the complement D^c (the so-called *overshoot*).

An American-style derivative security has a payoff $\psi = \psi(t, x)$, and the option holder can exercise at any time between the contract inception at time $t = 0$ and expiration $T > 0$ and receive the payoff. In general, the payoff can depend both on the time of exercise and the state of the process at the time of exercise. Assuming the option holder follows a value-maximizing exercise strategy, the value function of the American option is given by the value function of the optimal stopping problem:

$$V(t, x) = \sup_{\theta \in \Theta_{t,T}} \mathbb{E}_{t,x}\big[e^{-\int_t^\theta r(s,X_s)\,\mathrm{d}s}\psi(\theta, X_\theta)\big], \tag{4.4}$$

where the supremum is taken over the set of all stopping times θ taking values in $[t, T]$, denoted by $\Theta_{t,T}$.

Let $a = \sigma \cdot \sigma^T$ denote the diffusion matrix, and define a differential operator $\mathcal{A}$:

$$\begin{aligned} \mathcal{A}u(t, x) := {} & \frac{1}{2}\sum_{i,j=1}^n a_{ij}(t, x)\frac{\partial^2 u}{\partial x_i \partial x_j}(t, x) \\ & + \sum_{i=1}^n b_i(t, x)\frac{\partial u}{\partial x_i}(t, x) - \big(r(t, x) + \lambda(t, x)\big)u(t, x) \end{aligned}$$

and an integral operator $\mathcal{B}$:

$$\mathcal{B}u(t,x) := \lambda(t,x)\int_{\mathbb{R}^n} u(t,x+y)\mu(t,x;\mathrm{d}y).$$

Then the infinitesimal generator $\mathcal{G}$ of the (generally time-inhomogeneous) Markov jump-diffusion process is given by:

$$\mathcal{G}u(t,x) + r(t,x)u(t,x) = \mathcal{A}u(t,x) + \mathcal{B}u(t,x).$$

We now characterize the value functions (4.2)–(4.4) as solutions to PIDE problems. To simplify exposition and avoid dealing with technical complications, we also suppose that r, b_i, a_{ij}, and $\partial a_{ij}/\partial x_k$ are bounded on $\mathbb{R}^n$ and that the coercivity condition holds for the diffusion matrix; i.e., there exists a constant $\alpha > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(t,x)\xi_i\xi_j \geqslant \alpha|\xi|^2,$$

for all $\xi \in \mathbb{R}^n$. In particular, here we do not consider pure jump processes without the diffusion component, $a \equiv 0$, as this case requires special care (see Matache et al., 2004, 2005a). These conditions are not necessary and can be relaxed, but since our exposition here is informal, we do not go into this. Under these conditions, it follows that the value function $V(t,x)$ defined by (4.2) is the unique solution of the PIDE:

$$V_t + \mathcal{A}V + \mathcal{B}V = 0, \quad t \in [0,T),\ x \in \mathbb{R}^n$$

with the terminal condition

$$V(T,x) = \psi(x), \quad x \in \mathbb{R}^n.$$

The value function of the knock-out contract (4.3) is the unique solution of the same PIDE but subject to the additional *knock-out condition with rebate*:

$$V(t,x) = R(t,x), \quad x \in D^c,\ t \in [0,T).$$

An important observation is that in the presence of jumps the rebate must be specified everywhere in D^c and not only on the boundary ∂D, as in the pure diffusion case. This is due to the possibility of overshoot. Hence, the knock-out condition also must be imposed everywhere in D^c and not just on the boundary.

The value function of the American-style problem is the unique solution of the variational inequality:

$$V_t + \mathcal{A}V + \mathcal{B}V \leqslant 0, \qquad t \in [0,T),\ x \in \mathbb{R}^n, \tag{4.5a}$$

$$V \geqslant \psi, \qquad t \in [0,T),\ x \in \mathbb{R}^n, \tag{4.5b}$$

$$(V_t + \mathcal{A}V + \mathcal{B}V)\cdot(V-\psi) = 0, \quad t \in [0,T),\ x \in \mathbb{R}^n, \tag{4.5c}$$

with the terminal condition

$$V(T, x) = \psi(T, x), \quad x \in \mathbb{R}^n. \tag{4.5d}$$

The fundamental reference for jump-diffusion processes, optimal stopping problems, and variational inequalities is Bensoussan and Lions (1984) (see Theorems 3.3, 4.4, 9.3 in particular). The variational formulation of the American option problem in one-dimensional jump-diffusion models was developed by Zhang (1997), who extended the results in Jaillet et al. (1990) to jump-diffusions.

4.2 Localization and variational formulation

To simplify exposition, in what follows we assume that the process coefficients (a, b, r, λ, μ) and the contract specification (payoff ψ and rebate R) are time-independent. First consider European and American options (without contractual knock-out provisions). We start with localizing the original problem on $\mathbb{R}^n$ by considering an increasing exhausting sequence of bounded open domains $\{\Omega_k\}$ such that $\Omega_k \subset \Omega_{k+1}$ and $\bigcup_k \Omega_k = \mathbb{R}^n$. Then the value function $U = V(T - t, x)$ of the original problem is realized as the limit of a sequence of functions U_k which solve the localized PIDEs on bounded domains:

$$U_{k,t} - \mathcal{A}U_k - \mathcal{B}U_k = 0, \quad t \in (0, T], \ x \in \Omega_k$$

with the initial condition:

$$U_k(0, x) = \psi(x), \quad x \in \Omega_k.$$

An *artificial knock-out condition* is imposed everywhere in $(\Omega_k)^c = \mathbb{R}^n \backslash \Omega_k$:

$$U_k(t, x) = \psi(x), \quad x \in (\Omega_k)^c,$$

where we set the artificial rebate equal to the payoff for European and American options. In other words, we approximate the original option contract with an artificial option contract that knocks out when the process exits a bounded domain Ω_k and pays a rebate $\psi(X_\tau)$ at the first exit time τ. In the presence of jumps, we must impose the artificial knock-out condition *everywhere* in $(\Omega_k)^c$ due to the possibility of an overshoot. If the original option contract already has a knock-out provision with some bounded domain D, then we simply set the computational domain equal to D. If D is unbounded (e.g., single-barrier options), then we also need to localize the problem to a bounded domain $\Omega \subset D$.

For the localized problem, we have (see Bensoussan and Lions, 1984 for general results on localization to bounded domains):

$$\max_{t \in [0,T]} \left\| U(t, \cdot) - U_k(t, \cdot) \right\|_{L^\infty(G)} \to 0 \quad \text{as } k \to \infty$$

for any fixed compact set $G \in \Omega_k$. As discussed previously, the localization errors can be shown to decay exponentially in the size of the computational domain Ω_k. In the remainder of this section, we take a bounded computational

domain Ω_k as given and solve the PIDE on Ω_k with the artificial knock-out condition in $(\Omega_k)^c$. We drop the index k to lighten notation.

We consider a variational formulation on a given bounded domain Ω. We first consider European options. The PIDE has a nonhomogeneous artificial knock-out condition with artificial rebate $\psi(x)$. We homogenize it by introducing $u := U - \psi$. A variational formulation of the PIDE problem is obtained by considering a space of test functions square-integrable on Ω, with their (weak) first derivatives square-integrable on Ω, and vanishing in Ω^c. Multiplying the PIDE with a test function $v = v(x)$, integrating over Ω, and integrating by parts, we arrive at the variational (weak) formulation of the PIDE: find $u = u(t, x)$ such that for every test function $v = v(x)$

$$(u_t, v) + a(u, v) - b(u, v) + a(\psi, v) - b(\psi, v) = 0, \quad t \in (0, T], \tag{4.6a}$$

and

$$u(0, x) = 0, \tag{4.6b}$$

where $(u, v) = \int_\Omega u(x)v(x)\,\mathrm{d}x$ is the inner product in $L^2(\Omega)$, and the two bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are defined by

$$\begin{aligned} a(u, v) := {} & \frac{1}{2}\sum_{i,j=1}^{n} \int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_i}(x)\frac{\partial v}{\partial x_j}(x)\,\mathrm{d}x \\ & - \sum_{i=1}^{n} \int_\Omega a_i(x)\frac{\partial u}{\partial x_i}(x)v(x)\,\mathrm{d}x \\ & + \int_\Omega \big(r(x) + \lambda(x)\big)u(x)v(x)\,\mathrm{d}x, \end{aligned} \tag{4.7}$$

where

$$a_i(x) = b_i(x) - \frac{1}{2}\sum_{j=1}^{n} \frac{\partial a_{ij}}{\partial x_j}(x),$$

and

$$b(u, v) := (\mathcal{B}u, v) = \int_\Omega \int_{\mathbb{R}^n} u(x + y)v(x)\lambda(x)\mu(x; \mathrm{d}y)\,\mathrm{d}x. \tag{4.8}$$

The terms with the bilinear form b result from the jump dynamics. To solve the variational formulation, we seek a function u in an appropriate function space such that (4.6) hold for any test function v. The solution u vanishes outside of Ω. The value function U is then given by $U = \psi + u$. The general framework for variational formulations of PIDEs associated with jump-diffusion processes can be found in Bensoussan and Lions (1984). A variational formulation of

Merton's jump-diffusion model can be found in Zhang (1997). Matache et al. (2004, 2005a) develop a variational formulation for Lévy process-based models.

Similarly to the Black–Scholes–Merton case, for American options we obtain a variational inequality (that now acquires additional terms due to jumps):

$$(u_t, v-u) + a(u, v-u) - b(u, v) + a(\psi, v-u) - b(\psi, v) \geqslant 0, \quad \forall v \geqslant 0. \tag{4.9}$$

We seek such u that (4.9) is satisfied for all nonnegative test functions vanishing on Ω^c and such that u vanishes on Ω^c, with the vanishing initial condition (4.6b) and such that the nonnegativity constraint $u \geqslant 0$ is satisfied on $\overline{\Omega}$. We can approximate this variational inequality with nonlinear PDE in the penalized formulation. In the penalized formulation a penalty term approximates the action of the early exercise constraint:

$$(u_{\epsilon,t}, v) + a(u_\epsilon, v) - b(u_\epsilon, v) + a(\psi, v) - b(\psi, v) + \big(\pi_\epsilon(u_\epsilon), v\big) = 0 \tag{4.10}$$

with the vanishing initial condition (4.6b) and vanishing artificial knock-out condition in Ω^c.

4.3 Finite element approximation and integration of the ODE system

In $d \geqslant 1$ dimensions, the state space domain is partitioned into nonoverlapping *elements* (intervals in one dimension, triangles or rectangles in two dimensions, etc.), covering the state space domain with a mesh called a *triangulation*. In this chapter for simplicity we only consider rectangular elements. Consider a two-dimensional example with a bounded computational domain $\Omega = (\underline{x}, \bar{x}) \times (\underline{y}, \bar{y})$. We divide $[\underline{x}, \bar{x}]$ into $m_1 + 1$ equal intervals of length $h_x = (\underline{x} - \bar{x})/(m_1 + 1)$ and $[\underline{y}, \bar{y}]$ into $m_2 + 1$ equal intervals of length $h_y = (\underline{y} - \bar{y})/(m_2 + 1)$. The nodes are $(x_i, y_j) = (\underline{x} + ih_x, \underline{y} + jh_y)$, $i = 0, \ldots, m_1 + 1$, $j = 0, \ldots, m_2 + 1$. The rectangular two-dimensional finite element basis functions are defined for any $i = 1, \ldots, m_1$ and $j = 1, \ldots, m_2$ as the product of the one-dimensional basis functions:

$$\phi_{ij}(x, y) = \phi_i(x)\phi_j(y) = \phi\big((x - x_i)/h\big)\phi\big((y - y_j)/h\big),$$

where $\phi_i(x)$, $\phi_j(y)$, and $\phi(\cdot)$ are as previously defined for the one-dimensional case. The two-dimensional pyramid function ϕ_{ij} is equal to one at the node (x_i, y_j) and zero at all other nodes. There are $m_1 \times m_2$ nodes in the interior $(\underline{x}, \bar{x}) \times (\underline{y}, \bar{y})$. We order the nodes as follows: (x_1, y_1), (x_1, y_2), $\ldots$, (x_1, y_{m_2}), (x_2, y_1), (x_2, y_2), $\ldots$, (x_{m_1}, y_{m_2}), and number the nodes and the basis functions accordingly. We will use the same notation $\phi_i(x)$ for the basis functions in d dimensions, where $i = 1, \ldots, m$, and m is the total number of interior nodes in the triangulation (e.g., $m = m_1 \times \cdots \times m_d$ for rectangular elements

in d dimensions, where m_k is the number of intervals in discretizing kth coordinate). Much more general triangulations can be employed in the finite element method, covering the domain with an essentially unstructured mesh consisting of triangles or rectangles in two dimensions, or tetrahedra, prisms, or hexahedra in three dimensions, etc. Details can be found in Quarteroni and Valli (1997) and Achdou and Pironneau (2005), as well as other finite element references cited above. Associated basis functions are taken to be piecewise polynomials of order p. Here we only consider piecewise affine basis functions. For derivatives pricing problems, the approximation to the value function solving the variational formulation (4.6) (or (4.9) or (4.10) for American-style derivatives) is written in the form (2.7) with time-dependent coefficients $u_i(t)$ to be determined by integrating an ODE system similar to Eq. (2.8). In the general case of a d-dimensional jump-diffusion process and European-style options, the ODE system takes the form:

$$\mathbb{M}\mathbf{u}'(t) + \mathbb{A}\mathbf{u}(t) - \mathbb{B}\mathbf{u}(t) + \mathbf{F} = 0, \quad t \in (0, T], \tag{4.11}$$

with the initial condition (2.9). Here the mass matrix, the stiffness matrix, the load vector, and the initial condition vector are defined as before with the $L^2(\Omega)$ inner product and the bilinear form (4.7). In the presence of jumps we also have a *jump matrix* $\mathbb{B} = (b_{ij})_{i,j=1}^m$,

$$b_{ij} = b(\phi_j, \phi_i) = (\mathcal{B}\phi_j, \phi_i),$$

defined by the bilinear form (4.8) associated with jumps. In the one-dimensional Black–Scholes–Merton model, we were able to calculate the integrals in the expressions for elements of the mass and stiffness matrices in closed form. In general, elements of the mass, stiffness and jump matrices, as well as the load and the initial condition vector, are computed by numerical quadrature. In one dimension, the one-point Gaussian quadrature rule that evaluates the integrand at the center of the element (interval) of length h has error of the order $O(h^2)$ and is, thus, sufficient for the finite element approximation since the error of the finite element discretization of the variational formulation of the PIDE is itself $O(h^2)$ (see, e.g., Ciarlet, 1978, Section 4.1 on the use of numerical integration in the finite element method). In d dimensions, d-dimensional tensor product of the one-point Gaussian quadrature rule evaluates the integrand at the center of the d-dimensional element (here we are only considering rectangular elements). Higher order Gaussian quadrature may be used to improve the constant in the estimate (see Feng and Linetsky, 2006b).

We have reduced the option pricing problem to the solution of the ODE system (4.11). For a d-dimensional jump-diffusion model, we observe that $\mathbb{A} \sim O(h^{-d})$ and the elements of the matrix $\mathbb{A}$ increase as h decreases. In contrast, $\mathbb{B}$ and $\mathbb{M}$ decreases as h decreases. Hence, the term $\mathbb{A}\mathbf{u}$ resulting from discretization of the diffusion part of the PIDE generates stiffness, while the term $\mathbb{B}\mathbf{u}$ resulting from discretization of the integral operator does not generally generate stiffness. For stiff systems, fully explicit schemes are only

conditionally stable and may require prohibitively small time steps when h is small. Therefore, we treat the term $\mathbb{A}\mathbf{u}$ implicitly for stability reasons. Recall that $\mathbb{B}$ is a dense matrix (as opposed to the stiffness matrix $\mathbb{A}$ which is tri-diagonal in the one-dimensional case and has 3^d nonzero diagonals in d dimensions). Since it does not generate stiffness, we may treat the term $\mathbb{B}\mathbf{u}$ explicitly to avoid inverting the dense matrix. This is an example of an *implicit–explicit* (IMEX) time stepping scheme (see Hundsdorfer and Verwer, 2003, Section IV.4 for a survey of IMEX methods for ODEs and PDEs, where some terms in the ODE or PDE are treated implicitly, while other terms are treated explicitly).

In particular, the IMEX Euler time stepping scheme starts with the initial condition $\mathbb{M}\mathbf{u}^0 = \mathbf{C}$ and marches forward according to:

$$(\mathbb{M} + k\mathbb{A})\mathbf{u}^i = (\mathbb{M} + k\mathbb{B})\mathbf{u}^{i-1} - k\mathbf{F}, \quad i = 1, \ldots, N. \tag{4.12}$$

At each step, the linear system (4.12) is solved to determine the m-dimensional vector $\mathbf{u}^i$. This scheme is unconditionally stable,[1] first-order accurate in time, and its error is known to have an asymptotic expansion in the powers of the time step k of the form (2.11). This asymptotic error expansion suggests applying extrapolation to cancel lower order terms in the error expansion and to increase the order of the scheme, as we did for the pure diffusion Black–Scholes–Merton model. Some computational experiments with the extrapolation scheme based on the IMEX Euler are provided in Section 5 (see Feng and Linetsky, 2006b for more details). IMEX versions of other high order schemes are also available, such as Crank–Nicolson–Adams–Bashforth, IMEX BDF, etc. (e.g., Ascher et al., 1995, 1997).

Remark 1. In this chapter we have only considered the basic formulation of the Galerkin finite element method with piecewise affine basis functions. For the finite element method with more general piecewise polynomial basis functions see the finite element references given in the Introduction, as well as the monograph Solin et al. (2003). As an alternative to piecewise polynomial finite element basis functions, wavelet basis functions may also be used (Matache et al., 2004, 2005a, 2005b, 2006; Hilber et al., 2005). In particular, interesting recent research on sparse wavelet tensor products that effectively compress the dimension of the problem can be found in von Petersdorff and Schwab (2004), where some prototype high-dimensional problems (up to *twenty* dimensions) have been considered.

[1] Note that we include the term λU in the PIDE in the definition of the operator $\mathcal{A}$ to be treated implicitly, so only the integral is included in the definition of the operator $\mathcal{B}$ to be treated explicitly. The resulting IMEX scheme is proved to be unconditionally stable by d'Halluin et al. (2005). This is in contrast with the IMEX scheme in Zhang (1997), who treats the reaction term explicitly as well, resulting in a stability condition of the form $\Delta t \leqslant C$ with some constant C.

Remark 2. Within the framework of the finite element method, further computational efficiency is achieved by adaptively refining the finite element mesh (triangulation) based on local errors in the computed numerical solution. An application of adaptive mesh refinement to the pricing of American-style options can be found in Achdou and Pironneau (2005). General references on adaptive mesh refinement are Adjerid et al. (1999a, 1999b), Bergam et al. (2005), Babuška and Suri (1994), Eriksson and Johnson (1991, 1995).

5 Examples and applications

5.1 One-dimensional jump-diffusion models

Consider the setup of Section 4.1, where now $\{X_t,\ t \geqslant 0\}$ is a one-dimensional jump-diffusion process:

$$\mathrm{d}X_t = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}B_t + \mathrm{d}J_t, \quad \mu = r - q - \sigma^2/2 + \lambda\big(1 - \mathbb{E}\big[e^Z\big]\big),$$

where $\{J_t,\ t \geqslant 0\}$ is a jump process, a compound Poisson process with intensity $\lambda > 0$ and a given jump size (magnitude) distribution, i.e., $J_t = \sum_{n=1}^{N_t} Z_n$, where N_t is a Poisson process with intensity λ and $\{Z_n\}$ are i.i.d. jump magnitudes. It is also assumed that the Brownian motion, the Poisson process, and the jump magnitudes are all independent. The drift μ is adjusted so that the discounted total gains process, including price changes and dividends, is a martingale under the EMM, i.e., $\mathbb{E}[S_t] = e^{(r-q)t}S_0$ for each $t > 0$ for the price process.

In Merton's (1976) model the jump magnitude distribution is normal with mean m and standard deviation s with the probability density:

$$p(z) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(z-m)^2}{2s^2}\right).$$

In this model, the drift parameter is $\mu = r - q - \sigma^2/2 + \lambda[1 - \exp(m + s^2/2)]$. In Kou's (2002) model (see also Kou and Wang, 2004) the jump magnitude distribution is double exponential with the density:

$$p(z) = p\eta_1 e^{-\eta_1 z} 1_{\{z \geqslant 0\}} + (1-p)\eta_2 e^{\eta_2 z} 1_{\{z<0\}}$$

and $\mu = r - q - \sigma^2/2 + \lambda[(1-p)(\eta_2 + 1)^{-1} - p(\eta_1 - 1)^{-1}]$. In this model positive jumps occur with probability p and are exponentially distributed with mean $1/\eta_1$ with $\eta_1 > 1$, and negative jumps occur with probability $1 - p$ and are exponentially distributed with mean $1/\eta_2$ with $\eta_2 > 0$.

To price European options, we need to solve the PIDE

$$U_t - \mathcal{A}U - \mathcal{B}U = 0, \quad t \in (0, T],$$

with the initial condition $U(0, x) = \psi(x)$. For one-dimensional jump-diffusion models the operators are

$$\mathcal{A}U = \frac{1}{2}\sigma^2 U_{xx} + \mu U_x - (r + \lambda)U,$$
$$\mathcal{B}U(t, x) = \lambda \int_{\mathbb{R}} U(t, x + z)\, p(z)\, \mathrm{d}z.$$

As discussed in Section 4.1, for knock-out options due to the possibility of overshoot the knock-out condition needs to be imposed everywhere in Ω^c and not only on the boundary $\partial\Omega$.

After localization to a bounded computational domain $(\underline{x}, \bar{x})$, the bilinear forms in the variational formulation of the PIDE are

$$a(u, v) = \frac{1}{2}\sigma^2 \int_{\underline{x}}^{\bar{x}} u_x v_x \,\mathrm{d}x - \mu \int_{\underline{x}}^{\bar{x}} u_x v \,\mathrm{d}x + (r + \lambda) \int_{\underline{x}}^{\bar{x}} uv \,\mathrm{d}x$$

and

$$b(u, v) = (\mathcal{B}\, u, v) = \lambda \int_{\underline{x}}^{\bar{x}} \int_{-\infty}^{\infty} u(x + z)v(x)\, p(z)\, \mathrm{d}z\, \mathrm{d}x.$$

In the finite element formulation of this jump-diffusion model the mass and stiffness matrices $\mathbb{M}$ and $\mathbb{A}$ are the same as in the Black–Scholes–Merton model in Section 2.4, but we also have a jump matrix $\mathbb{B}$. Its elements $b_{ij} = b(\phi_j, \phi_i) = (\mathcal{B}\phi_j, \phi_i)$ have the form:

$$\begin{aligned} b_{ij} &= \lambda \int_{\underline{x}}^{\bar{x}} \int_{-\infty}^{\infty} \phi_j(x + z)\phi_i(x)\, p(z)\, \mathrm{d}z\, \mathrm{d}x \\ &= \lambda \int_{x_{i-1}}^{x_{i+1}} \int_{x_{j-1}}^{x_{j+1}} \phi_j(y)\phi_i(x)\, p(y - x)\, \mathrm{d}y\, \mathrm{d}x \\ &= \lambda h^2 \int_{-1}^{1} \int_{-1}^{1} \phi(u)\phi(w)\, p\big((w - u + j - i)h\big)\, \mathrm{d}w\, \mathrm{d}u \end{aligned}$$

and depend only on the difference $j - i$. Therefore, $\mathbb{B}$ is a *Toeplitz matrix* with the same elements along each diagonal. Hence, we only need to compute $2m - 1$ values. The double integral can be computed directly by the two-dimensional Gaussian quadrature. Alternatively, in our implementation we compute $b_{ij} = (\mathcal{B}\phi_j, \phi_i)$ as follows. We approximate the function $(\mathcal{B}\phi_j)(x)$

by its *finite element interpolant* $I_h\mathcal{B}\phi_j(x)$:

$$\mathcal{B}\phi_j(x) \approx I_h\mathcal{B}\phi_j(x) = \sum_{l\in\mathbb{Z}} \big(\mathcal{B}\phi_j(x_l)\big)\phi_l(x),$$

where I_h is the finite element interpolation operator. The finite element interpolant $I_h f(x) = \sum_l f(x_l)\phi_l(x)$ of a function $f(x)$ is equal to the value of the function at the nodes x_l and interpolates between the nodes with the piecewise-linear finite element basis functions $\phi_l(x)$. The error of the finite element interpolation is $O(h^2)$ and is, thus, of the same order as the spatial discretization error in our semi-discrete formulation of the PIDE. We then have the following approximation:

$$\begin{aligned}(\mathcal{B}\phi_j, \phi_i) &\approx \sum_{l=i-1}^{i+1} \mathcal{B}\phi_j(x_l)\cdot(\phi_l, \phi_i)\\ &= \frac{1}{6}h\mathcal{B}\phi_j(x_{i-1}) + \frac{2}{3}h\mathcal{B}\phi_j(x_i) + \frac{1}{6}h\mathcal{B}\phi_j(x_{i+1})\end{aligned}$$

for any $i, j \in \mathbb{Z}$, where

$$\begin{aligned}\mathcal{B}\phi_j(x_l) &= \lambda\int_{\mathbb{R}} \phi_j(x_l+z)\,p(z)\,\mathrm{d}z = \lambda h\int_{-1}^{1} \phi(x)\,p\big((x+j-l)h\big)\,\mathrm{d}x\\ &= \lambda h\int_0^1 \big[xp\big((x-1+j-l)h\big) + (1-x)\,p\big((x+j-l)h\big)\big]\,\mathrm{d}x.\end{aligned} \tag{5.1}$$

Note that $\mathcal{B}\phi_j(x_l)$ depends only on the difference $j-l$. Hence, to compute the jump matrix $\mathbb{B}$, we need to compute $2m+1$ values $\mathcal{B}\phi_j(x_l)$. In Kou's and Merton's models, the integral in (5.1) can be calculated analytically. For a general jump magnitude distribution, the integral (5.1) is computed by numerical quadrature. The one-point Gaussian quadrature rule that evaluates the integrand at the center of the integration interval of length h has errors of the order $O(h^2)$ and is, thus, sufficient for the finite element approximation (see, e.g., Ciarlet, 1978, Section 4.1):

$$\mathcal{B}\phi_j(x_l) \approx \frac{1}{2}h\lambda\big[p\big((j-l-1/2)h\big) + p\big((j-l+1/2)h\big)\big].$$

We integrate the resulting ODE system (4.11) by the extrapolation scheme based on the IMEX Euler scheme as described in Section 4.3. We note that the jump matrix $\mathbb{B}$ is Toeplitz. In the numerical solution of the system (4.12), we need to perform the jump matrix–vector multiplication at each time step. The Toeplitz matrix–vector multiplication can be accomplished efficiently in $O(m\log_2(m))$ floating point operations using the fast Fourier transform (FFT) (see Feng and Linetsky, 2006a, 2006b).

Table 1.
Parameter values used in numerical experiments

Kou	$\sigma = 0.1$, $\lambda = 3$, $p = 0.3$, $\eta_1 = 40$, $\eta_2 = 12$, $T = 1$ year
Merton	$\sigma = 0.1$, $\lambda = 3$, $m = -0.05$, $s = 0.086$, $T = 1$ year
SVCJ	$\lambda = 4$, $\nu = 0.01$, $m = -0.01$, $s = 0.01$, $\rho_D = -0.5$, $\rho_J = -1$ $\xi = 0.1$, $\kappa = 2$, $\theta = 0.04$, $T = 3$ months
Other parameters	$K = 100$, $L = 80$, $U = 120$, $r = 5\%$, $q = 2\%$

We now present some numerical examples for European, single- and double-barrier options. Parameters used in our numerical experiments are given in Table 1. We investigate pricing errors due to localization and spatial and temporal discretizations. Errors are in the maximum norm on the approximation domain $G = [\log(0.8), \log(1.2)]$ in the x-variable (corresponding to the approximation domain $G = [80, 120]$ in the underlying stock price variable S). For barrier options, benchmark prices are computed with small enough space steps and time steps. For European options in Kou's and Merton's models, accurate benchmark prices can be computed using available analytical solutions.

Figure 2 illustrates convergence of the finite element spatial discretization, temporal discretization, and localization in Kou's and Merton's models. The two plots at the top show spatial convergence and localization error decay for the down-and-out put (DOP) in Kou's model. The two plots in the middle row show spatial convergence and localization error decay for the up-and-out call (UOC) in Merton's model. The two bottom plots show temporal convergence of the IMEX extrapolation scheme for double-barrier put (DBP) options in Kou's model. The first and second plots in the first column plot the maximum norm pricing error as a function of the spatial discretization step size h. The error plots are in log–log scale and clearly demonstrate the Ch^2 convergence of the finite element approximation. While the double-barrier problem has a bounded domain and no localization is needed, for single-barrier and European options, we localize to bounded computational domains. In particular, for down-and-out options, the domain is $[\underline{x}, \bar{x}_k]$, where the lower barrier $\underline{x}$ is fixed contractually, $\underline{x} = \ln(L/K)$, and a sequence of increasing artificial upper barriers $\bar{x}_k$ is considered (denoted by xmax in the plots). Similarly, for up-and-out options, the upper barrier $\bar{x}$ is fixed contractually, $\bar{x} = \ln(U/K)$, and a sequence of decreasing artificial lower barriers $\underline{x}_k$ is considered (denoted by xmin in the plots). When a computational domain is fixed and h is refined, there is a minimum error beyond which no further error reduction can be obtained by refining h. This is a localization error corresponding to this computational domain. The computation domain needs to be enlarged to obtain further error reduction. This can be clearly seen in the first and second plots in the first column for dow-and-out puts and up-and-out calls, respectively. For each fixed computational domain, these plots exhibit the Ch^2

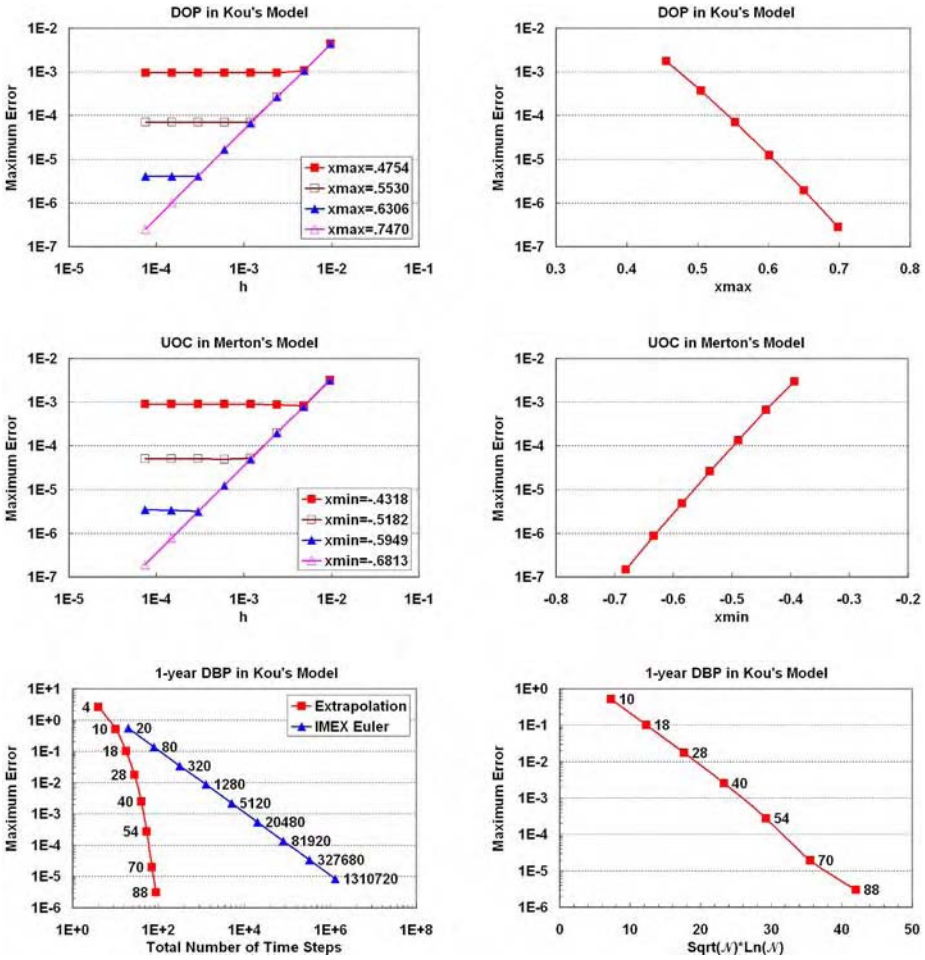


Figure 2. Maximum norm pricing errors in Kou's and Merton's models: spatial discretization, localization, and temporal discretization. DOP = down-and-out put, UOC = up-and-out call, DBP = double-barrier put. Parameters are given in Table 1. The approximation domain is [80, 120] in the stock price.

convergence up until the localization error starts to dominate. The localization error itself decays exponentially in the size of the computational domain, as shown in the first and second plots in the second column. The two bottom plots illustrate convergence of the temporal discretization for one-year double-barrier puts in Kou's model. The spatial discretization step size h was taken small enough to guarantee spatial discretization errors less than 10^{-5}, our target accuracy. For one-year options, two basic steps of six months each were taken. Time discretization errors are plotted for the IMEX Euler scheme and the extrapolation scheme. The plots illustrate that the extrapolation scheme is

remarkably fast and accurate. For the one-year double-barrier put option, our extrapolation scheme took a total of 88 time steps (in 0.07 seconds on a Dell Xeon 3.06 GHz PC) to achieve our target accuracy of 10^{-5}, while the IMEX Euler scheme took more than 1.3 million steps (in 833 seconds) to achieve the same accuracy. The second plot in the last row plots the maximum norm error as a function of $\mathcal{N}^{1/2}\ln\mathcal{N}$. This plot illustrates that our scheme exhibits error decay $O(\exp(-c\mathcal{N}^{1/2}\ln\mathcal{N}))$ in the number of time steps $\mathcal{N}$, consistent with Eq. (2.14).

5.2 *Duffie–Pan–Singleton model with stochastic volatility and contemporaneous and correlated jumps in asset return and volatility*

In the one-dimensional model, the volatility σ is constant. In contrast, in the model with stochastic volatility and contemporaneous and correlated jumps in the asset price and volatility (SVCJ), the instantaneous variance $V_t = \sigma_t^2$ is assumed to follow a CIR diffusion punctuated by positive jumps. The corresponding two-dimensional stochastic differential equation (SDE) is (Duffie et al., 2000)

$$\begin{aligned}
\mathrm{d}X_t &= (\mu - V_{t-}/2)\,\mathrm{d}t + \sqrt{V_{t-}}\Big[\sqrt{1-\rho_D^2}\,\mathrm{d}B_t^1 + \rho_D\,\mathrm{d}B_t^2\Big] + \mathrm{d}J_t^X,\\
\mathrm{d}V_t &= \kappa(\theta - V_{t-})\,\mathrm{d}t + \xi\sqrt{V_{t-}}\,\mathrm{d}B_t^2 + \mathrm{d}J_t^V,
\end{aligned}$$

where θ is the long-run variance level, κ is the rate of mean reversion, ξ is the volatility-of-volatility parameter, B^1 and B^2 are two independent standard Brownian motions, ρ_D is the correlation coefficient correlating Brownian shocks in the return and variance processes, and (J_t^X, J_t^V) is a two-dimensional jump process, a $\mathbb{R}\times\mathbb{R}^+$-valued compound Poisson process with intensity $\lambda > 0$ and a bi-variate jump magnitude distribution in $\mathbb{R}\times\mathbb{R}^+$. The process starts at $X_0 := x = \ln(S_0/K)$ and $V_0 = v > 0$ at time zero. The jump magnitudes (Z_n^X, Z_n^V) are i.i.d. with a joint bi-variate probability density $p(z^x, z^v)$. The marginal distribution of the jump size in variance is assumed to be exponential with mean ν. Conditional on a jump of size z^v in the variance process, the jump size in the return process X_t is assumed to be normally distributed with mean $m + \rho_J z^v$ (where ρ_J defines correlation between jumps in return and variance) and standard deviation s. The bi-variate density is

$$\begin{aligned}
p(z^x, z^v) &= \frac{1}{\nu\sqrt{2\pi s^2}}\exp\left(-\frac{z^v}{\nu} - \frac{(z^x - m - \rho_J z^v)^2}{2s^2}\right),\\
&z^x \in \mathbb{R},\ z^v > 0.
\end{aligned}$$

The drift parameter is: $\mu = r - q + \lambda[1 - (1-\nu\rho_J)^{-1}\exp(m + s^2/2)]$. The infinitesimal generator of the two-dimensional Markov process (X_t, V_t) is given by

$$\mathcal{G}f(x,v) = \frac{1}{2}vf_{xx} + \rho\xi v f_{vx} + \frac{1}{2}\xi^2 v f_{vv} + \left(\mu - \frac{1}{2}v\right)f_x + \kappa(\theta - v)f_v$$

$$+\lambda \int_{-\infty}^{\infty}\int_{0}^{\infty}\big[f\big(x+z^x, v+z^v\big)-f(x,v)\big]p\big(z^x,z^v\big)\,\mathrm{d}z^v\,\mathrm{d}z^x.$$

For future convenience, we introduce a scaled and centered dimensionless variance process $Y_t = (V_t - \theta)/\theta$. A jump of size $\triangle V$ in V_t corresponds to a jump of size $\triangle Y = \triangle V/\theta$ in Y_t. Hence, the joint distribution of jumps in the state variables (X_t, Y_t) has a density:

$$p\big(z^x, z^y\big) = \frac{\theta}{\nu\sqrt{2\pi s^2}}\exp\!\left(-\frac{\theta z^y}{\nu} - \frac{(z^x - m - \rho_J\theta z^y)^2}{2s^2}\right),$$
$$z^x \in \mathbb{R},\ z^y > 0.$$

In the PDE formulation of the two-dimensional SVJJ model, the value function also depends on the initial variance at time t, represented by the scaled and centered state variable $Y_t = y$, $U = U(t, x, y)$. The differential and integral operators are

$$\begin{aligned}\mathcal{A}U &= \frac{1}{2}\theta(y+1)U_{xx} + \rho_D\xi(y+1)U_{xy} + \frac{\xi^2}{2\theta}(y+1)U_{yy}\\ &\quad + \left(\mu - \frac{1}{2}\theta(y+1)\right)U_x - \kappa y U_y - (r+\lambda)U,\\ \mathcal{B}U(t,x,y) &= \lambda\int_{-\infty}^{\infty}\int_{0}^{\infty} U\big(t, x+z^x, y+z^y\big)\,p\big(z^x,z^y\big)\,\mathrm{d}z^y\,\mathrm{d}z^x.\end{aligned}$$

The initial condition is $U(0, x, y) = \psi(x)$, where $\psi(x)$ is the payoff (which for call and put options depends only on the x variable). Finally, for knock-out options the appropriate knock-out conditions are imposed.

In the SVJJ model, after localization to a bounded computational domain $(\underline{x}, \bar{x}) \times (\underline{y}, \bar{y})$, the bilinear forms in the variational formulation are

$$\begin{aligned}a(u,v) = \int_{\underline{x}}^{\bar{x}}\int_{\underline{y}}^{\bar{y}}\Bigg(&\frac{1}{2}(y+1)\left(\theta u_x v_x + \rho\xi u_y v_x + \rho\xi u_x v_y + \frac{1}{\theta}\xi^2 u_y v_y\right)\\ &+\left(ky + \frac{\xi^2}{2\theta}\right)u_y v - \left(\mu - \frac{1}{2}\rho\xi - \frac{1}{2}\theta - \frac{1}{2}\theta y\right)u_x v\\ &+(r+\lambda)uv\Bigg)\,\mathrm{d}y\,\mathrm{d}x\end{aligned}$$

and

$$b(u,v) = \lambda\int_{\underline{x}}^{\bar{x}}\int_{\underline{y}}^{\bar{y}}\int_{-\infty}^{\infty}\int_{0}^{\infty} u\big(x+z^x, y+z^y\big)v(x,y)\,p\big(z^x,z^y\big)\,\mathrm{d}z^y\,\mathrm{d}z^x\,\mathrm{d}y\,\mathrm{d}x.$$

The finite element formulation is constructed as described in Section 4.3. In our examples we consider rectangular finite elements. Explicit expressions for the mass $\mathbb{M}$, stiffness $\mathbb{A}$, and jump matrices $\mathbb{B}$ and the load vector are rather cumbersome and can be found in Feng and Linetsky (2006b).

We now present some numerical examples. Parameters used in our numerical examples are given in Table 1. For European options in the SVCJ model, the benchmark prices are computed using the Fourier transform method for affine jump-diffusions (Duffie et al., 2000). Figure 3 shows double-barrier put option pricing in the SVCJ model. The first plot shows the value function for a double-barrier put option as a function of the underlying stock price and volatility. The second plot shows convergence of the finite element spatial discretization. The Ch^2 error estimate is clearly verified. The two bottom plots show convergence of the extrapolation time stepping scheme based on the IMEX Euler scheme. Maximum norm errors of the IMEX Euler scheme and the extrapolation scheme with the basic step sizes $H = T = 0.25$ year are plotted. The extrapolation scheme is remarkably fast and accurate for this two-dimensional application with bi-variate jumps. The double-barrier put option is priced with the accuracy of nearly 10^{-4} in 35 time steps (in 6.68 seconds on a Dell Xeon 3.06 GHz PC). The IMEX Euler scheme takes about 10,000 time steps to attain 10^{-3} accuracy (in 484 seconds). The last plot illustrates the $O(\exp(-c\mathcal{N}^{1/2}\ln\mathcal{N}))$ error decay in the number of time steps $\mathcal{N}$.

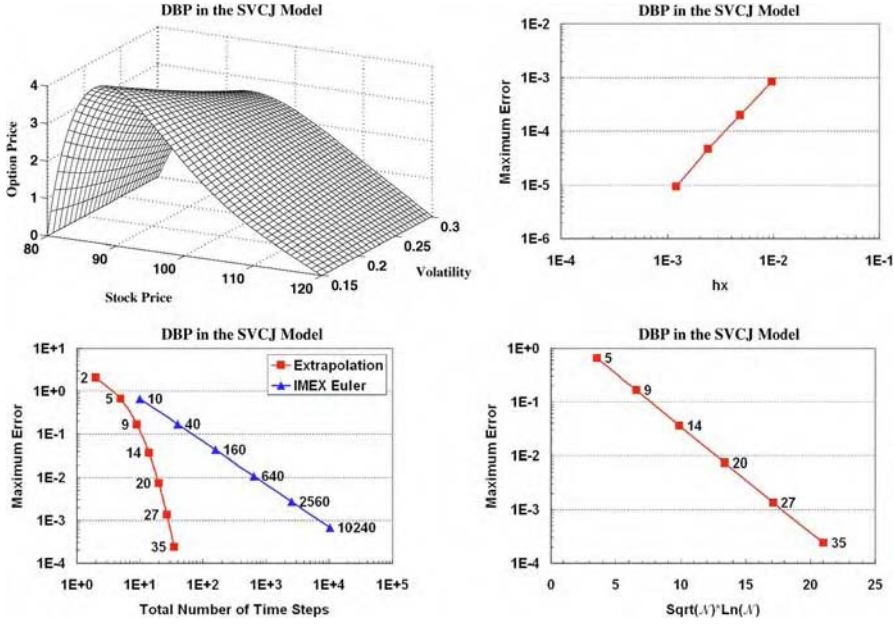


Figure 3. Double-barrier put option in the SVCJ model. Parameters are given in Table 1.

5.3 Multi-asset American-style options

In this section we consider an example of a multi-asset American-style option. Consider a Black–Scholes–Merton model with n assets with the risk-neutral dynamics:

$$\mathrm{d}S_i(t) = (r - q_i)S_i(t)\,\mathrm{d}t + \sigma_i S_i(t)\,\mathrm{d}W_i(t), \quad S_i(0) = S_i,$$

where W_i are n correlated Brownian motions with the correlation matrix ρ_{ij}, $\mathrm{d}W_i(t)\,\mathrm{d}W_j(t) = \rho_{ij}\,\mathrm{d}t$. We will consider an American-style put option on the geometric average of the prices of n assets. Introducing the geometric average process $I_t := (\prod_{i=1}^n S_i(t))^{1/n}$, the put payoff is $(K - I_t)^+$ if the option is exercised at time $t \in [0, T]$ between the contract inception and expiration T. We choose this geometric average option for our example due to the fact that the geometric average of n geometric Brownian motions is itself a geometric Brownian motion with volatility σ_I and risk-neutral drift $r - q_I$, where volatility and the "effective dividend yield" are

$$\sigma_I^2 = \frac{1}{n^2}\sum_{i,j=1}^n \rho_{ij}\sigma_i\sigma_j, \qquad q_I = \frac{1}{n}\sum_{i=1}^n\left(q_i + \frac{1}{2}\sigma_i^2\right) - \frac{1}{2}\sigma_I^2.$$

Therefore, the problem of valuing options on the geometric average of n assets can be reduced to the one of valuing options on the one-dimensional geometric Brownian motion I. This allows us to produce accurate benchmark prices for American options on geometric average by solving this one-dimensional option pricing problem. Specifically, we use a one-dimensional trinomial tree with two million time steps to produce benchmarks that are accurate to at least 10^{-6}. In the examples that follow, we assume that $K = 100$, all assets have initial values $S_i(0) = 100$, $\sigma_i = 20\%$, $q_i = 0$, and pairwise correlations $\rho_{ij} = 0.5$ for $i \neq j$.

It will be convenient to do the log transform to the dimensionless variables $X_i(t) = \ln(S_i(t)/K)$:

$$X_i(t) = x_i + \big(r - q_i - \sigma_i^2/2\big)t + \sigma_i W_i(t), \quad x_i = \ln(S_i/K).$$

The payoff function is then:

$$\psi(x_1, x_2, \ldots, x_n) = K\big(1 - e^{\frac{1}{n}\sum_{i=1}^n x_i}\big)^+.$$

We solve the nonlinear penalized PDE formulation of Section 3.2. The finite element spatial discretization leads to the ODE system of the form (3.11). We consider the penalty with $p = 2$ in our examples. To discretize spatially, we use rectangular n-dimensional finite elements (tensor products of one-dimensional finite elements) as discussed in Section 4.3. The ODE system is integrated with the adaptive step size and variable order BDF-based package SUNDIALS discussed in Section 2.5.

The nonlinear ODE system (3.11) approximates the original continuous valuation problem for the n-asset American put. It contains the following approximation parameters: the spatial step size h_i for each of the variables x_i

(we select $h_i = h$ in our example since all assets have the same volatility), the radius $R = \max_i R_i$ of the computational domain $[-R_1, R_1] \times \cdots \times [-R_n, R_n]$ (we select $R_i = R$ in our example since all assets have the same volatility), and the penalization parameter ϵ in (3.12), (3.13). In the numerical examples we used the penalty term (3.12)–(3.13) with $p = 2$. We select the approximation domain $G = [-0.25, 0.25] \times \cdots \times [-0.25, 0.25]$ in the x_i variables, which corresponds to the price interval [77.9, 128.4] for each underlying asset S_i. We integrate the nonlinear ODE system with the SUNDIALS solver discussed in Section 2.5 and compute the approximation error between the computed solution and the benchmark (obtained by the one-dimensional trinomial tree with two million time steps) in the maximum norm over the approximation domain G. From theory we expect to observe maximum norm errors to be $O(h^2)$ in the spatial discretization step size and $O(\epsilon)$ in the penalization parameter. We also expect exponential error decay in the radius of the computational domain $O(e^{-cR})$. Figure 4 presents our results for three-month American-style put options on the geometric average of four assets. The first plot shows the maximum norm error as a function of h. We clearly observe the quadrative convergence rate. The second plot shows the error as a function of the computational domain radius $R = \text{xmax}$. The exponential error decay with increasing xmax is evident. The third plot shows the error as a function of the penal-

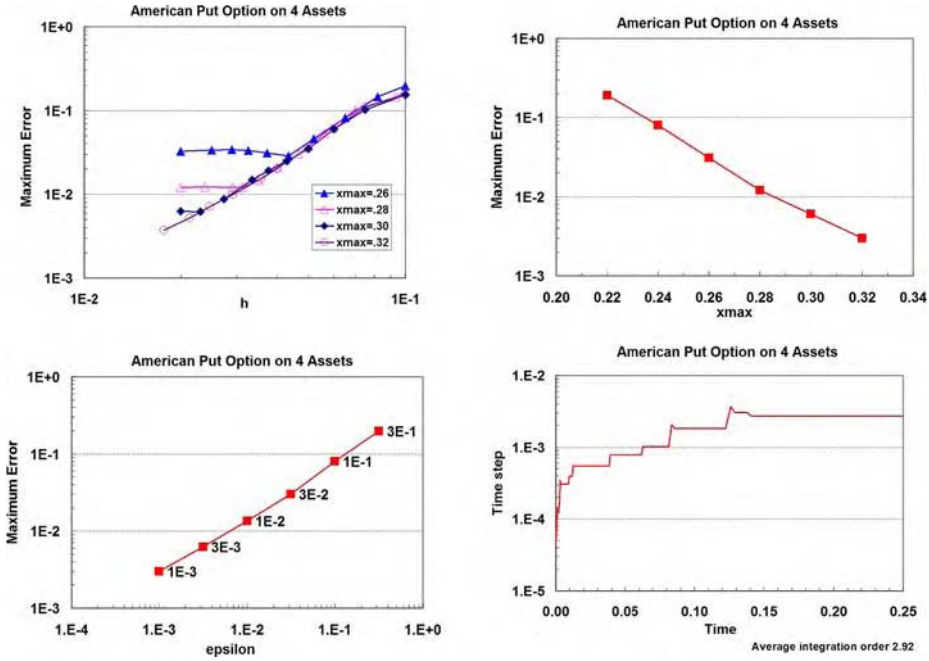


Figure 4. Convergence studies for the American-style put option on the geometric average of four assets.

ization parameter ϵ. The linear convergence in ϵ is verified. The fourth plot shows the time step history of the SUNDIALS solver in temporally integrating the ODE system. We observe that the adaptive solver starts with a small initial step size, and then rapidly increases the step size as the temporal integration progresses. The reason for this is as follows. The option payoff function is only C^0 (it has a kink at the strike). For any time prior to maturity, the American option value function is C^1 (it is continuous and once continuously differentiable across the optimal exercise boundary). The optimal exercise boundary itself changes rapidly near maturity. In the penalized formulation, this results in large gradients of the PDE solution near maturity. Thus, near maturity small time steps are needed to insure desired accuracy. As temporal integration progresses, the optimal exercise boundary of the American option flattens out and, correspondingly, the gradients in the solution of the nonlinear PDE decrease, allowing the solver to take progressively larger time steps to achieve the same error tolerance. The SUNDIALS solver also varies the integration order between the first and fifth order, based on the smoothness of the solution. High-order time stepping with larger time steps generally works best for smooth solutions, while low-order time stepping with small time steps is needed in the regions where the solution is not smooth. The average integration order realized by the solver in this example is 2.92.

6 Summary

In this chapter we briefly surveyed a powerful computational method for the valuation of options in multi-dimensional jump-diffusion models based on: (1) converting the PIDE or PIDVI to a variational (weak) form; (2) discretizing the weak formulation spatially by the Galerkin finite element method to obtain a system of ODEs; and (3) integrating the resulting system of ODEs in time. This finite element method-of-lines framework to solve PDEs and PIDEs is quite general theoretically and flexible and powerful computationally. The method is applicable to general Markov processes with diffusion and jump components with time- and state-dependent coefficients, as well as a large class of derivative contracts, including European and American-style exercise, barriers, averages, etc. The growing literature on financial engineering applications of variational methods includes:

- Multi-asset American options: Marcozzi (2001), Sapariuc et al. (2004), Achdou and Pironneau (2005);
- Foreign currency options with stochastic foreign and domestic interest rates: Choi and Marcozzi (2001, 2003), Kovalov et al. (2007);
- American-style Asian options: Marcozzi (2003);
- Stochastic volatility models: Zvan et al. (1998b, 1999), Hilber et al. (2005), Achdou and Tchou (2002);
- Jump-diffusion models: Zhang (1997), Feng and Linetsky (2006b);

- European and American options in Lévy process models: Matache et al. (2004, 2005a, 2005b, 2006);
- Convertible bonds: Kovalov and Linetsky (2007);
- Inverse problems arising in option calibration: Achdou et al. (2004), Achdou (2005);
- Bond options: Allegretto et al. (2003).

References

Achdou, Y. (2005). An inverse problem for a parabolic variational inequality arising in volatility calibration with American options. *SIAM Journal on Control and Optimization* 43, 1583–1615.

Achdou, Y., Pironneau, O. (2005). *Computational Methods for Option Pricing*. *SIAM Frontiers in Applied Mathematics*. SIAM, Philadelphia.

Achdou, Y., Tchou, N. (2002). Variational analysis for the Black and Scholes equation with stochastic volatility. *M2AN Mathematical Modelling and Numerical Analysis* 36, 373–395.

Achdou, Y., Indragoby, G., Pironneau, O. (2004). Volatility calibration with American options. *Methods and Applications of Analysis* 11 (3), 1–24.

Adjerid, S., Babuška, I., Flaherty, J.E. (1999a). A posteriori error estimation with finite element method of lines solution of parabolic systems. *Mathematical Models and Methods Applied in Science* 9, 261–286.

Adjerid, S., Belguendouz, B., Flaherty, J.E. (1999b). A posteriori finite element error estimation for diffusion problems. *SIAM Journal on Scientific Computing* 21, 728–746.

Allegretto, W., Lin, Y.P., Yang, H.T. (2003). Numerical pricing of American put options on zero-coupon bonds. *Applied Numerical Mathematics* 46 (2), 113–134.

Ascher, U.M., Ruuth, S.J., Wetton, B.T.R. (1995). Implicit–explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis* 32, 797–823.

Ascher, U.M., Ruuth, S.J., Spiteri, R.J. (1997). Implicit–explicit Runge–Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics* 25, 151–167.

Babuška, I., Suri, M. (1994). The p and hp versions of the finite element method, basic principles and properties. *SIAM Review* 36 (4), 578–632.

Bensoussan, A. (1984). On the theory of option pricing. *Acta Applicandae Mathematicae* 2 (2), 139–158.

Bensoussan, A., Lions, J.L. (1982). *Applications of Variational Inequalities in Stochastic Control*. Elsevier, Amsterdam.

Bensoussan, A., Lions, J.L. (1984). *Impulse Control and Quasi-Variational Inequalities*. Gauthier–Villars, Paris.

Bergam, A., Bernardi, C., Mghazli, Z. (2005). A posteriori analysis of the finite element discretization of a non-linear parabolic equation. *Mathematics of Computation* 74, 1097–1116.

Boman, M., (2001). A posteriori error analysis in the maximum norm for a penalty finite element method for the time-dependent obstacle problem. *Preprint*. Chalmers Finite Element Center, Goteborg, Sweden.

Choi, S., Marcozzi, M.D. (2001). A numerical approach to American currency option valuation. *Journal of Derivatives* 9 (2), 19–29.

Choi, S., Marcozzi, M.D. (2003). The valuation of foreign currency options under stochastic interest rates. *Computers and Mathematics with Applications* 45, 741–749.

Ciarlet, P.G. (1978). *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam (revised ed., *SIAM*, 2002).

Cottle, R., Pang, J.-S., Stone, R.E. (1992). *The Linear Complementary Problem*. Academic Press.

Cryer, C.W. (1971). The solution of a quadratic programming problem using systematic overrelaxation. *SIAM Journal on Control* 9, 385–392.

Deuflhard, P. (1985). Recent progress in extrapolation methods for ordinary differential equations. *SIAM Review* 27 (4), 505–535.

Deuflhard, P., Bornemann, F. (2002). *Scientific Computing with Ordinary Differential Equations*. Springer, Berlin.

d'Halluin, Y., Forsyth, P.A., Labahn, G. (2004). A penalty method for American options with jump-diffusion processes. *Numerische Mathematik* 97 (2), 321–352.

d'Halluin, Y., Forsyth, P.A., Vetzal, K.R. (2005). Robust numerical methods for contingent claims under jump-diffusion processes. *IMA Journal of Numerical Analysis* 25, 87–112.

Duffie, D., Pan, J., Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 68 (6), 1343–1376.

Eriksson, E., Johnson, C. (1991). Adaptive finite element methods for parabolic problems. I. A linear model problem. *SIAM Journal on Numerical Analysis* 28, 43–77.

Eriksson, E., Johnson, C. (1995). Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$. *SIAM Journal on Numerical Analysis* 32, 706–740.

Feng, L., Linetsky, V., (2006a). Pricing discretely monitored barrier options and defaultable bonds in Lévy process models: A Hilbert transform approach. *Mathematical Finance*, in press.

Feng, L., Linetsky, V., (2006b). Pricing options in jump-diffusion models: An extrapolation approach. *Operations Research*, in press.

Friedman, A. (1976). *Stochastic Differential Equations and Applications, vol. II*. Academic Press, New York.

Forsyth, P.A., Vetzal, K.R. (2002). Quadratic convergence for valuing American options using a penalty method. *SIAM Journal on Scientific Computing* 23 (6), 2095–2122.

Glowinski, R. (1984). *Numerical Methods for Nonlinear Variational Problems*. Springer, Berlin.

Glowinski, R., Lions, J.L., Tremolieris (1981). *Numerical Analysis of Variational Inequalities*. North-Holland, Amsterdam.

Hairer, E., Wanner, G. (1996). *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, second ed. Springer, Berlin.

Hilber, H., Matache, A.-M., Schwab, C. (2005). Sparse wavelet methods for option pricing under stochastic volatility. *Journal of Computational Finance* 8 (4).

Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S. (2005). SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software* 31 (3), 363–396. Also available as LLNL technical report UCRL-JP-200037.

Hindmarsh, A.C., Serban, R., Collier, A. (2006). User documentation for IDA, a differential-algebraic equation solver for sequential and parallel computers. Technical Report UCRL-MA-136910, December, Lawrence Livermore National Laboratory.

Hull, J. (2006). *Options, Futures and Other Derivatives*, sixth ed. Prentice Hall, NJ.

Hundsdorfer, W., Verwer, J.G. (2003). *Numerical Solution of Time-dependent Advection-Diffusion–Reaction Equations*. Springer, Berlin.

Jacod, J., Shiryaev, A.N. (2003). *Limit Theorems for Stochastic Processes*. Springer, Berlin.

Jaillet, P., Lamberton, D., Lapeyre, B. (1990). Variational inequalities and the pricing of American options. *Acta Applicandae Mathematicae* 21, 263–289.

Johnson, C. (1987). *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge Univ. Press, Cambridge.

Kangro, R., Nicolaides, R. (2000). Far field boundary conditions for Black–Scholes equations. *SIAM Journal on Numerical Analysis* 38 (4), 1357–1368.

Karatzas, I. (1988). On the pricing of American options. *Applied Mathematics and Optimization* 17, 37–60.

Kou, S.G. (2002). A jump-diffusion model for option pricing. *Management Science* 48 (8), 1086–1101.

Kou, S.G., Wang, H. (2004). Option pricing under a double exponential jump-diffusion model. *Management Science* 50, 1178–1192.

Kovalov, P., Linetsky, V. (2007). Valuing convertible bonds with stock price, volatility, interest rate, and default risk. *Working paper*. Northwestern University.

Kovalov, P., Linetsky, V., Marcozzi, M. (2007). Pricing multi-asset American options: A finite element method-of-lines with smooth penalty. *Journal of Scientific Computing*, in press.

Lamberton, D., Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*. Chapman & Hall.

Larsson, S., Thomee, V. (2003). *Partial Differential Equations with Numerical Methods*. Springer, Berlin.

Marcozzi, M. (2001). On the approximation of optimal stopping problems with application to financial mathematics. *SIAM Journal on Scientific Computing* 22 (5), 1865–1884.

Marcozzi, M.D. (2003). On the valuation of Asian options by variational methods. *SIAM Journal on Scientific Computing* 24 (4), 1124–1140.

Matache, A.-M., von Petersdorff, T., Schwab, C. (2004). Fast deterministic pricing of options on Lévy driven assets. *M2NA Mathematical Modelling and Numerical Analysis* 38 (1), 37–72.

Matache, A.-M., Nitsche, P.-A., Schwab, C. (2005a). Wavelet Galerkin pricing of American options on Lévy driven assets. *Quantitative Finance* 5 (4), 403–424.

Matache, A.-M., Schwab, C., Wihler, T. (2005b). Fast numerical solution of parabolic integro-differential equations with applications in finance. *SIAM Journal on Scientific Computing* 27 (2), 369–393.

Matache, A.-M., Schwab, C., Wihler, T. (2006). Linear complexity solution of parabolic integro-differential equations. *Numerische Mathematik*, in press.

Merton, R. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.

Pooley, D.M., Forsyth, P.A., Vetzal, K. (2003). Remedies for non-smooth payoffs in option pricing. *Journal of Computational Finance* 6 (4), 25–40.

Sapariuc, I., Marcozzi, M.D., Flaherty, J.E. (2004). A numerical analysis of variational valuation techniques for derivative securities. *Applied Mathematics and Computation* 159 (1), 171–198.

Solin, P., Segeth, K., Dolezel, I. (2003). *Higher Order Finite Element Methods*. Chapman & Hall/CRC.

Quarteroni, A., Valli, A. (1997). *Numerical Approximation of Partial Differential Equations*. Springer, Berlin.

Tavella, D., Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*. Wiley.

Thomee, V. (1997). *Galerkin Finite Element Methods for Parabolic Problems*. Springer, Berlin.

Thomee, V. (2001). From finite differences to finite elements. A short history of numerical analysis. *Journal of Computational and Applied Mathematics* 128, 1–54.

von Petersdorff, T., Schwab, C. (2004). Numerical solution of parabolic equations in high dimensions. *M2NA Mathematical Modelling and Numerical Analysis* 38, 93–128.

Wilmott, P., Dewynne, J., Howison, S. (1993). *Option Pricing: Mathematical Models and Computations*. Oxford Financial Press, Oxford, UK.

Zhang, X.-L. (1997). Numerical analysis of American option pricing in a jump-diffusion model. *Mathematics of Operations Research* 22 (3), 668–690.

Zvan, R., Forsyth, P.A., Vetzal, K.R. (1998a). Swing low, swing high. *RISK* 11, 71–75.

Zvan, R., Forsyth, P.A., Vetzal, K.R. (1998b). Penalty methods for American options with stochastic volatility. *Journal of Computational and Applied Mathematics* 91 (2), 199–218.

Zvan, R., Forsyth, P.A., Vetzal, K.R. (1999). A finite element approach to the pricing of discrete lookbacks with stochastic volatility. *Applied Mathematical Finance* 6, 87–106.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15008-8

Chapter 8

Discrete Barrier and Lookback Options

S.G. Kou

Department of Industrial Engineering and Operations Research, Columbia University
E-mail: sk75@columbia.edu

Abstract

Discrete barrier and lookback options are among the most popular path-dependent options in markets. However, due to the discrete monitoring policy almost no analytical solutions are available for them. We shall focus on the following methods for discrete barrier and lookback option prices: (1) Broadie–Yamamoto method based on fast Gaussian transforms. (2) Feng–Linetsky method based on Hilbert transforms. (3) A continuity correction approximation. (4) Howison–Steinberg approximation based on the perturbation method. (5) A Laplace inversion method based on Spitzer's identity. This survey also contains a new (more direct) derivation of a constant related to the continuity correction approximation.

1 Introduction

Discrete path-dependent options are the options whose payoffs are determined by underlying prices at a finite set of times, whereas the payoff of a continuous path-dependent option depends on the underlying price throughout the life of the option. Due to regulatory and practical issues, most of path-dependent options traded in markets are discrete path-dependent options.

Among the most popular discrete path-dependent options are discrete Asian options or average options, discrete American options or Bermuda options, discrete lookback options and discrete barrier options. The payoff of a discrete Asian option depends on a discrete average of the asset price. For example, a standard discrete (arithmetic European) Asian call option has a payoff $(\frac{1}{n}\sum_{i=1}^{n} S(t_i) - K)^+$ at maturity $T = t_n$, where $t_1, t_2, \ldots, t_n$ are monitoring points, K is the strike price of the call option, and $S(t)$ is the asset price at time t; see Zhang (1998), Hull (2005). A discrete American option is an American option with exercise dates being restricted to a discrete set of monitoring points; see Glasserman (2004).

In this survey we shall focus on discrete barrier and lookback options, because very often they can be studied in similar ways, as their payoffs all depend on the extrema of the underlying stochastic processes. The study of discrete Asian options is of separate interest, and requires totally different techniques. Discrete American options are closely related to numerical pricing of American options; there is a separate survey in this handbook on them. Due to the similarity between discrete barrier options and discrete lookback options, we shall focus on discrete barrier options, although most of the techniques discussed here can be easily adapted to study discrete lookback options.

1.1 Barrier and lookback options

A standard (also called floating) lookback call (put) gives the option holder the right to buy (sell) an asset at its lowest (highest) price during the life of the option. In other words, the payoffs of the floating lookback call and put options are $S(T) - m_{0,T}$ and $M_{0,T} - S(T)$, respectively, where $m_{0,T}$ and $M_{0,T}$ are minimum and maximum of the asset price between 0 and T. In a discrete time setting the minimum (maximum) of the asset price will be determined at discrete monitoring instants. In the same way, the payoffs of the fixed strike put and call are $(K - m_{0,T})^+$ and $(M_{0,T} - K)^+$. Other types of lookback options include percentage lookback options in which the extreme values are multiplied by a constant, and partial lookback options in which the monitoring interval for the extremum is a subinterval of $[0, T]$. We shall refer the interested reader to Andreasen (1998) for a detailed description.

A barrier option is a financial derivative contract that is activated (knocked in) or extinguished (knocked out) when the price of the underlying asset (which could be a stock, an index, an exchange rate, an interest rate, etc.) crosses a certain level (called a barrier). For example, an up-and-out call option gives the option holder the payoff of a European call option if the price of the underlying asset does not reach a higher barrier level before the expiration date. More complicated barrier options may have two barriers (double barrier options), and may have the final payoff determined by one asset and the barrier level determined by another asset (two-dimensional barrier options); see Zhang (1998) and Hull (2005).

Taken together, discrete lookback and barrier options are among the most popular path-dependent options traded in exchanges worldwide and also in over-the-counter markets. Lookback and barrier options are also useful outside the context of literal options. For example, Longstaff (1995) approximates the values of marketability of a security over a fixed horizon with a type of continuous-time lookback option and gives a closed-form expression for the value; the discrete version of lookback options will be relevant in his setting. Merton (1974), Black and Cox (1976), and more recently Leland and Toft (1996), Rich (1996), and Chen and Kou (2005) among others, have used barrier models for study credit risk and pricing contingent claims with endogenous

default. For tractability, this line of work typically assumes continuous monitoring of a reorganization boundary. But to the extent that the default can be modeled as a barrier crossing, it is arguably one that can be triggered only at the specific dates – e.g. coupon payment dates.

An important issue of pricing barrier options is whether the barrier crossing is monitored in continuous time or in discrete time. Most models assume the continuous time version mainly because this leads to analytical solutions; see, for example, Gatto et al. (1979), Goldman et al. (1979), and Conze and Viswanathan (1991), Heynen and Kat (1995) for continuous lookback options; and see, for example, Merton (1973), Heynen and Kat (1994a, 1994b), Rubinstein and Reiner (1991), Chance (1994), and Kunitomo and Ikeda (1992) for various formulae for continuously monitored barrier options under the classical Brownian motion framework. Recently, Boyle and Tian (1999) and Davydov and Linetsky (2001) have priced continuously monitored barrier and lookback options under the CEV model using lattice and Laplace transform methods, respectively; see Kou and Wang (2003, 2004) for continuously monitored barrier options under a jump-diffusion framework.

However in practice most, if not all, barrier options traded in markets are discretely monitored. In other words, they specify fixed times for monitoring of the barrier (typically daily closings). Besides practical implementation issues, there are some legal and financial reasons why discretely monitored barrier options are preferred to continuously monitored barrier options. For example, some discussions in trader's literature ("*Derivatives Week*", May 29th, 1995) voice concern that, when the monitoring is continuous, extraneous barrier breach may occur in less liquid markets while the major western markets are closed, and may lead to certain arbitrage opportunities.

Although discretely monitored barrier and lookback options are popular and important, pricing them is not as easy as that of their continuous counterparts for several reasons:

(1) There are essentially no closed solutions, except using m-dimensional normal distribution functions (m is the number of monitoring points), which cannot be computed easily if, for example, $m > 5$; see Section 3.
(2) Direct Monte Carlo simulation or standard binomial trees may be difficult, and can take hours to produce accurate results; see Broadie et al. (1999).
(3) Although the central limit theorem asserts that as $m \to \infty$ the difference between the discretely and continuously monitored barrier options should be small, it is well known that the numerical differences can be surprisingly large, even for large m; see, e.g., the table in Section 4.

Because of these difficulties, many numerical methods have been proposed for pricing discrete barrier and lookback options.

1.2 Overview of different methods

First of all, by using the change of numeraire argument the pricing of barrier and lookback options can be reduced to studying either the marginal distribution of the first passage time, or the joint probability of the first passage time and the terminal value of a discrete random walk; see Section 2. Although there are many representations available for these two classical problems, there is little literature on how to compute the joint probability explicitly before the recent interest in discrete barrier and lookback options. Many numerical methods have been developed in the last decade for discrete barrier and lookback options. Popular ones are:

(1) Methods based on convolution, e.g. the fast Gaussian transform method developed in Broadie and Yamamoto (2003) and the Hilbert transform method in Feng and Linetsky (2005). This is basically due to the fact that the joint probability of the first passage time and the terminal value of a discrete random walk can be written as m-dimensional probability distribution (hence a m-dimensional integral or convolution.) We will review these results in Section 3.

(2) Methods based on the asymptotic expansion of discrete barrier options in terms of continuous barrier options, assuming $m \to \infty$. Of course, as we mentioned, the straightforward result from the central limit theorem, which has error $o(1)$, does not give a good approximation. An approximation based on the results from sequential analysis (see, e.g., Siegmund, 1985 with the error order $o(1/\sqrt{m})$ is given in Broadie et al. (1999), whose proof is simplified in Kou (2003) and Hörfelt (2003). We will review these results in Section 4.

(3) Methods based on the perturbation analysis of differential equations, leading to a higher order expansion with the error order $o(1/m)$. This is investigated in Howison and Steinberg (2005) and Howison (2005). We will review these results in Section 5.

(4) Methods based on transforms. Petrella and Kou (2004) use Laplace transforms to numerically invert the Spitzer's identity associated with the first passage times. We will review these transform-based methods in Section 6.

Besides these specialized methods, there are also many "general methods," such as lattice methods, Monte Carlo simulation, etc. We call them general methods because in principle these methods can be applied to broader contexts, e.g. American options and other path-dependent options, not just for discrete barrier and lookback options. Broadly speaking, general methods will be less efficient than the methods which take advantage of the special structures of discrete barrier and lookback options. However, general methods are attractive if one wants to develop a unified numerical framework to price different types of options, not just discrete barrier and lookback options. Because of their generality, many methods could potentially belong to this category, and it is very difficult to give a comprehensive review for them. Below is only a short list of some of general methods.

(a) Lattice methods are among the most popular methods in option pricing. It is well known that the straightforward binomial tree is not efficient in pricing discrete and lookback barrier options, due to the inefficiencies in computing discrete extreme values of the sample paths involved in the payoffs. Broadie et al. (1999) proposed an enhanced trinomial tree method which explicitly uses the continuity correction in Broadie et al. (1997) and a shift node. A Dirichlet lattice method based on the conditional distribution via Brownian bridge is given in Kuan and Webber (2003). Duan et al. (2003) proposed a method based on Markov chain, in combination of lattice, simulation, and the quadrature method. Other lattice methods include adjusting the position of nodes (Ritchken, 1995; Cheuk and Vorst, 1997; Tian, 1999) and refining branching near the barrier (Figlewski and Gao, 1999; Ahn et al., 1999). See also Babbs (1992), Boyle and Lau (1994), Hull and White (1993), Kat (1995).

(b) Another popular general method is Monte Carlo simulation. Because the barrier options may involve events (e.g. barrier crossing) with very small probabilities, the straightforward simulation may have large variances. Variance reduction techniques, notably importance sampling and conditional sampling methods using Brownian bridge, can be used to achieve significant variance reduction. Instead of giving a long list of related papers, we refer the reader to an excellent book by Glasserman (2004).

(c) Since the price of a discrete barrier option can be formulated as a solution of partial differential equation, one can use various finite difference methods; see Boyle and Tian (1998) and Zvan et al. (2000).

(d) Because the prices of a discrete barrier price can be written in terms of m-dimensional integrals, one can also use numerical integration methods. See Ait-Sahalia and Lai (1997, 1998), Sullivan (2000), Tse et al. (2001), Andricopoulos et al. (2003), and Fusai and Recchioni (2003).

1.3 Outline of the survey

Due to the page limit, this survey focuses on methods that takes into account of special structures of the discrete barrier and lookback options, resulting in more efficient algorithms but with narrower scopes. In particular, we shall survey the following methods

(1) Broadie–Yamamoto method based on the fast Gaussian transform; see Section 3.

(2) Feng–Linetsky method based on Hilbert transform; see Section 3.

(3) A continuity correction approximation; see Section 4.

(4) Howison–Steinberg approximation based on the perturbation method; see Section 5.

(5) A Laplace inversion method based on Spitzer's identity; see Section 6.

This survey also contains a new (more direct) derivation of the constant related to the continuity correction; see Appendix B.

Because this is a survey article we shall focus on giving intuition and comparing different methods, rather than giving detailed proofs which can be found in individual papers. For example, when we discuss the continuity correction we use a picture to illustrate the idea, rather than giving a proof. When we present the Howison–Steinberg approximation we spend considerable time on the basic background of the perturbation method (so that people with only probabilistic background can understand the intuition behind the idea), rather than giving the mathematical details, which involve both the Spitzer function for Wiener–Hopf equations and can be found in the original paper by Howison and Steinberg (2005).

2 A representation of barrier options via the change of numeraire argument

We assume the price of the underlying asset $S(t)$, $t \geqslant 0$, satisfies $S(t) = S(0)\exp\{\mu t + \sigma B(t)\}$, where under the risk-neutral probability P^*, the drift is $\mu = r - \sigma^2/2$, r is the risk-free interest rate and $B(t)$ is a standard Brownian motion under P^*. In the continuously monitored case, the standard finance theory implies that the price of a barrier option will be the expectation, taken with respect to the risk-neutral measure P^*, of the discounted (with the discount factor being e^{-rT} with T the expiration date of the option) payoff of the option. For example, the price of a continuous up-and-out call option is given by

$$V(H) = \mathrm{E}^*\big(e^{-rT}\big(S(T) - K\big)^+ I\big(\tau(H, S) > T\big)\big),$$

where $K \geqslant 0$ is the strike price, $H > S(0)$ is the barrier and, for any process $Y(t)$, the notation $\tau(x, Y)$ means that $\tau(x, Y) := \inf\{t \geqslant 0\colon\ Y(t) \geqslant x\}$. The other seven types of the barrier options can be priced similarly. In the Brownian motion framework, all eight types of the barrier options can be priced in closed forms; see Merton (1973).

In the discretely monitoring case, under the risk neutral measure P^*, at the nth monitoring point, $n\Delta t$, with $\Delta t = T/m$, the asset price is given by

$$S_n = S(0)\exp\left\{\mu n\Delta t + \sigma\sqrt{\Delta t}\sum_{i=1}^{n} Z_i\right\} = S(0)\exp\big(W_n\sigma\sqrt{\Delta t}\big),$$
$$n = 1, 2, \ldots, m,$$

where the random walk W_n is defined by

$$W_n := \sum_{i=1}^{n}\left(Z_i + \frac{\mu}{\sigma}\sqrt{\Delta t}\right),$$

the drift is given by $\mu = r - \sigma^2/2$, and the Z_i's are independent standard normal random variables. By analogy, the price of the discrete up-and-out-call

option is given by

$$V_m(H) = \mathrm{E}^*\big(e^{-rT}(S_m - K)^+ I\big(\tau'(H,S) > m\big)\big)$$
$$= \mathrm{E}^*\big\{e^{-rT}(S_m - K)^+ I\big\{\tau'\big(a/\big(\sigma\sqrt{T}\big), W\big) > m\big\}\big\},$$

where $a := \log(H/S(0)) > 0$, $\tau'(H,S) = \inf\{n \geqslant 1\colon S_n \geqslant H\}$, $\tau'(x,W) = \inf\{n \geqslant 1\colon W_n \geqslant x\sqrt{m}\}$.

For any probability measure P, let $\hat{P}$ be defined by

$$\frac{\mathrm{d}\hat{P}}{\mathrm{d}P} = \exp\left\{\sum_{i=1}^m a_i Z_i - \frac{1}{2}\sum_{i=1}^m a_i^2\right\},$$

where the a_i, $i = 1, \ldots, n$, are arbitrary constants, and the Z_i's are standard normal random variables under the probability measure P. Then a discrete Girsanov theorem (Karatzas and Shreve, 1991, p. 190) implies that under the probability measure $\hat{P}$, for every $1 \leqslant i \leqslant m$, $\hat{Z}_i := Z_i - a_i$ is a standard normal random variable.

By using the discrete Girsanov theorem, we can represent the price of a discrete barrier options as a difference of two probabilities under different measures. This is called the change of numeraire argument; for a survey, see Schroder (1999). It is applied to the case of discrete barrier options by Kou (2003) and Hörfelt (2003) independently. However, the methods in Kou (2003) and Hörfelt (2003) lead to slightly different barrier correction formulae. To illustrate the change of numeraire argument for the discrete barrier options. Let us consider the case of the discrete up-and-out call option, as the other seven options can be treated similarly; see e.g. Haug (1999).

First note that

$$\begin{aligned}
&\mathrm{E}^*\big(e^{-rT}(S_m - K)^+ I\big(\tau'(H,S) > m\big)\big)\\
&\quad = \mathrm{E}^*\big(e^{-rT}(S_m - K) I\big(S_m \geqslant K, \tau'(H,S) > m\big)\big)\\
&\quad = \mathrm{E}^*\big(e^{-rT} S_m I\big(S_m \geqslant K, \tau'(H,S) > m\big)\big)\\
&\qquad - K e^{-rT} P^*\big(S_m \geqslant K, \tau'(H,S) > m\big).
\end{aligned}$$

Using the discrete Girsanov theorem with $a_i = \sigma\sqrt{\Delta t}$, we have that the first term in the above equation is given by

$$\begin{aligned}
&\mathrm{E}^*\left[e^{-rT} S(0) \exp\left\{\mu m \Delta t + \sigma\sqrt{\Delta t}\sum_{i=1}^m Z_i\right\} I\big(S_m \geqslant K, \tau'(H,S) > m\big)\right]\\
&\quad = S(0)\mathrm{E}^*\left[\exp\left\{-\frac{1}{2}\sigma^2 T + \sigma\sqrt{\Delta t}\sum_{i=1}^m Z_i\right\} I\big(S_m \geqslant K, \tau'(H,S) > m\big)\right]\\
&\quad = S(0)\hat{\mathrm{E}}\big(I\big(S_m \geqslant K, \tau'(H,S) > m\big)\big)\\
&\quad = S(0)\hat{P}\big(S_m \geqslant K, \tau'(H,S) > m\big).
\end{aligned}$$

Under $\hat{P}$, $\log S_m$ has a mean $\mu m\Delta t + \sigma\sqrt{\Delta t}\cdot m\sigma\sqrt{\Delta t} = (\mu+\sigma^2)T$ instead of μT under the measure P^*. Therefore, the price of a discrete up-and-out-call option is given by

$$V_m(H) = S(0)\hat{P}\left(W_m \geqslant \frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}, \tau'\left(a\big/\left(\sigma\sqrt{T}\right), W\right) > m\right) - Ke^{-rT}P^*\left(W_m \geqslant \frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}, \tau'\left(a\big/\left(\sigma\sqrt{T}\right), W\right) > m\right),$$

where

$$\text{under } \hat{P}, \quad W_m = \sum_{i=1}^{m}\left(\hat{Z}_i + \left\{(\mu+\sigma^2)/\sigma\right\}\sqrt{\frac{T}{m}}\right) = \sum_{i=1}^{m}\left(\hat{Z}_i + \left\{\left(r+\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\frac{T}{m}}\right)$$

and

$$\text{under } P^*, \quad W_m = \sum_{i=1}^{m}\left(Z_i + (\mu/\sigma)\sqrt{\frac{T}{m}}\right) = \sum_{i=1}^{m}\left(Z_i + \left\{\left(r-\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\frac{T}{m}}\right)$$

with $\hat{Z}_i$ and Z_i being standard normal random variables under $\hat{P}$ and P^*, respectively.

Therefore, the problem of pricing discrete barrier options is reduced to studying the joint probability of the first passage time (τ') and the terminal values (W_m) of a discrete random walk. Note that we have a first passage problem for the random walk W_n with a small drift ($\frac{\mu}{\sigma}\sqrt{\Delta t} \to 0$, as $m \to \infty$) to cross a high boundary ($a\sqrt{m}/(\sigma\sqrt{T}) \to \infty$, as $m \to \infty$).

3 Convolution, Broadie–Yamamoto method via the fast Gaussian transform, and Feng–Linetsky method via Hilbert transform

As we have seen in the last section, under the geometric Brownian motion model the prices of discrete barrier options can be represented as probabilities of random walk with increments having normal distributions. Thus, in principle analytical solutions of discrete barrier options can be derived using multivariate normal distributions; see, e.g., Heynan and Kat (1995) and Reiner (2000).

To give an illustration of the idea, consider a discrete up-and-in call option with two monitoring points, $t_1 = T/3$, $t_2 = 2T/3$, and $H < K$. Note that the

maturity T is not a monitoring point. We have

$$\begin{aligned} V_3(H) = {} & S(0)N_2\big(\hat{a}_{1,H,-}, \hat{a}_K; \sqrt{t_1/T}\big) \\ & - Ke^{-rT}N_2\big(a^*_{1,H,-}, a^*_K; \sqrt{t_1/T}\big) \\ & + S(0)N_3\big(\hat{a}_{1,H,+}, \hat{a}_{2,H,-}, \hat{a}_K; -\sqrt{t_1/t_2}, -\sqrt{t_1/T}, \sqrt{t_2/T}\big) \\ & - Ke^{-rT}N_3\big(a^*_{1,H,+}, a^*_{2,H,-}, a^*_K; -\sqrt{t_1/t_2}, -\sqrt{t_1/T}, \sqrt{t_2/T}\big), \end{aligned} \tag{1}$$

where the constants are

$$\begin{aligned} \hat{a}_{1,H,\pm} &\equiv \frac{\pm\log(H/S(0)) - \{(r + \frac{1}{2}\sigma^2)\}t_1}{\sigma\sqrt{t_1}}, \\ a^*_{1,H,\pm} &\equiv \frac{\pm\log(H/S(0)) - \{(r - \frac{1}{2}\sigma^2)\}t_1}{\sigma\sqrt{t_1}} \\ \hat{a}_{2,H,\pm} &\equiv \frac{\pm\log(H/S(0)) - \{(r + \frac{1}{2}\sigma^2)\}t_2}{\sigma\sqrt{t_2}}, \\ a^*_{2,H,\pm} &\equiv \frac{\pm\log(H/S(0)) - \{(r - \frac{1}{2}\sigma^2)\}t_2}{\sigma\sqrt{t_2}}, \\ \hat{a}_K &\equiv \frac{-\log(K/S(0)) - \{(r + \frac{1}{2}\sigma^2)\}T}{\sigma\sqrt{T}}, \\ a^*_K &\equiv \frac{-\log(K/S(0)) - \{(r - \frac{1}{2}\sigma^2)\}T}{\sigma\sqrt{T}}. \end{aligned}$$

The proof of (1) is given ir Appendix A. Here N_2 and N_3 denote the standard bivariate and trivariate normal distributions:

$$N_2(z_1, z_2; \varrho) = P(Z_1 \leqslant z_1, Z_2 \leqslant z_2),$$

where Z_1 and Z_2 are standard bivariate normal random variables with correlation ϱ, and

$$N_3(z_1, z_2, z_3; \varrho_{12}, \varrho_{13}, \varrho_{23}) = P(Z_1 \leqslant z_1, Z_2 \leqslant z_2, Z_3 \leqslant z_3),$$

with correlations $\varrho_{12}, \varrho_{13}, \varrho_{23}$. The pricing formula in (1) can be easily generalized to the case of m (not necessarily equally spaced) monitoring points, so that the price of a discrete barrier option with m monitoring points can be written involving the sum of multivariate normal distribution functions, with the highest dimension in the multivariate normal distributions being m.

However, m-dimensional normal distribution functions can hardly be computed easily if, for example, $m > 5$. Reiner (2000) proposed to use the fast Fourier transform to compute the convolution in the multivariate normal distribution. Recently there are two powerful ways to evaluate the convolution. One is the fast Gaussian transform in Broadie and Yamamoto (2003) in which

the convolution is computed very fast under the Gaussian assumption. The second method is the Hilbert transform method in Feng and Linetsky (2005), in which they recognize an interesting linking between Fourier transform of indicator functions and Hilbert transform. Feng and Linetsky method is more general, as it works as long as the asset returns follow a Lévy process. Below we give a brief summary of the two methods.

3.1 Broadie–Yamamoto method via the fast Gaussian transform

One of the key idea of Broadie–Yamamoto method is to recognize that one can compute integrals in convolution very fast if the integrals only involves normal density. For example, consider the discrete sum of Gaussian densities.

$$A(x_m) = \sum_{n=1}^{N} w_n \exp\left\{-\frac{(x_m - y_n)^2}{\delta}\right\}, \quad i = 1, \ldots, M.$$

The direct computation of the above sums will need $O(NM)$ operations. However, by using the Hermite functions to approximate the Gaussian densities, one can perform the above sum in $O(N) + O(1) + O(M) = O(\max(N, M))$ operations.

More precisely, the Hermite expansion yields

$$\exp\left\{-\frac{(x_m - y_n)^2}{\delta}\right\} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \frac{1}{i!j!}\left(\frac{y_n - y_0}{\sqrt{\delta}}\right)^j \left(\frac{x_m - x_0}{\sqrt{\delta}}\right)^i \times H_{i+j}\left(\frac{x_0 - y_0}{\sqrt{\delta}}\right),$$

where $H_{i+j}(\cdot)$ is the Hermite function. The expansion converges quite fast, typically eight terms may be enough. In other words, we have an approximation

$$\exp\left\{-\frac{(x_m - y_n)^2}{\delta}\right\} \approx \sum_{i=1}^{\alpha_{\max}}\sum_{j=1}^{\alpha_{\max}} \frac{1}{i!j!}\left(\frac{y_n - y_0}{\sqrt{\delta}}\right)^j \left(\frac{x_m - x_0}{\sqrt{\delta}}\right)^i \times H_{i+j}\left(\frac{x_0 - y_0}{\sqrt{\delta}}\right),$$

where $\alpha_{\max}$ is a small number, say no more than 8. Using this approximation, we have the Gaussian sum is given by

$$\begin{aligned} A(x_m) &\approx \sum_{n=1}^{N} w_n \sum_{i=1}^{\alpha_{\max}}\sum_{j=1}^{\alpha_{\max}} \frac{1}{i!j!}\left(\frac{y_n - y_0}{\sqrt{\delta}}\right)^j \left(\frac{x_m - x_0}{\sqrt{\delta}}\right)^i H_{i+j}\left(\frac{x_0 - y_0}{\sqrt{\delta}}\right) \\ &= \sum_{i=1}^{\alpha_{\max}} \frac{1}{i!}\left[\sum_{j=1}^{\alpha_{\max}} \frac{1}{j!}\left\{\sum_{n=1}^{N} w_n \left(\frac{y_n - y_0}{\sqrt{\delta}}\right)^j\right\} H_{i+j}\left(\frac{x_0 - y_0}{\sqrt{\delta}}\right)\right] \end{aligned}$$

$$\times \left(\frac{x_m - x_0}{\sqrt{\delta}} \right)^i.$$

Now the algorithm becomes

1. Compute $B_j = \sum_{n=1}^{N} w_n (\frac{y_n - y_0}{\sqrt{\delta}})^j$ for $j = 1, \ldots, \alpha_{\max}$.
2. Compute $C_i = \sum_{j=1}^{\alpha_{\max}} \frac{1}{j!} B_j H_{i+j} (\frac{x_0 - y_0}{\sqrt{\delta}})$ for $i = 1, \ldots, \alpha_{\max}$.
3. Approximate $A(x_m)$ as $\sum_{i=1}^{\alpha_{\max}} \frac{1}{i!} C_i (\frac{x_m - x_0}{\sqrt{\delta}})^i$ for $m = 1, \ldots, M$.

When $\alpha_{\max}$ is fixed, the total number of operations is therefore $O(N) + O(1) + O(M) = O(\max(N, M))$. Broadie and Yamamoto (2003) show that the above fast Gaussian transform is very fast. In fact, it is perhaps the fastest algorithm we can get so far under the Gaussian assumption. Of course, the algorithm relies on the special structure of the Gaussian distribution. For other distributions, similar algorithms might be available if some fast and accurate expansions of the density functions are available.

3.2 Feng–Linetsky method via Hilbert transform

Feng and Linetsky (2005) proposed a novel method to compute the convolution related to discrete barrier options via Hilbert transform. The key idea is that multiplying a function with the indicator function in the state space corresponds to Hilbert transform in the Fourier space. The method computes a sequence of Hilbert transforms at the discrete monitoring points, and then conducts one final Fourier inversion to get the option price. Feng–Linetsky method is quite general, as it works in principle for any Lévy process and for both single and double barrier options. The method also works very fast, as the number of operations is $O(MN \log_2 N)$, where M is the number of monitoring points and N is the number of sample points needed to compute the Hilbert transform.

To get an intuition of the idea, we shall illustrate a basic version of the method in terms of computing the probability $p(x)$ for a standard Brownian motion $B(t)$

$$\begin{aligned} p(x) &= P\big(\min\{B_\Delta, B_{2\Delta}, \ldots, B_{M\Delta}\} > 0 \mid B_0 = x\big) \\ &= \mathrm{E}\left\{ \prod_{i=1}^{M} I[B_{i\Delta} > 0] \;\middle|\; B_0 = x \right\}. \end{aligned}$$

We can compute $p(x)$ by the backward induction

$$\begin{aligned} v^M(x) &= I(x > 0), \\ v^{M-1}(x) &= I(x > 0) \cdot \mathrm{E}\big\{ I(B_\Delta > 0) \mid B_0 = x \big\} \\ &= I(x > 0) \cdot \mathrm{E}\big\{ v^M(B_\Delta) \mid B_0 = x \big\}, \end{aligned}$$

$$
\begin{aligned}
v^{M-2}(x) &= I(x>0)\cdot \mathrm{E}\big\{I(B_\Delta>0)I(B_{2\Delta}>0)\mid B_0=x\big\}\\
&= I(x>0)\cdot \mathrm{E}\big\{v^{M-1}(B_\Delta)\mid B_0=x\big\},\\
&\cdots\\
p(x) &= \mathrm{E}\big\{v^1(B_\Delta)\mid B_0=x\big\}.
\end{aligned}
$$

To take a Fourier transform we introduce a rescaling factor $e^{\alpha x}$,

$$
v_\alpha^j(x) = e^{\alpha x}v^j(x),\quad \alpha<0,
$$

because the indicator function $I(x>0)$ is not a L^1 function. This is equivalent to perform a Laplace transform. The backward induction becomes

$$
v_\alpha^M(x) = e^{\alpha x}I(x>0),
$$

$$
\begin{aligned}
v_\alpha^{M-1}(x) &= e^{\alpha x}\cdot I(x>0)\cdot \mathrm{E}\big\{I(B_\Delta>0)\mid B_0=x\big\}\\
&= I(x>0)\cdot \mathrm{E}\big\{e^{-\alpha(B_\Delta-x)}e^{\alpha B_\Delta}I(B_\Delta>0)\mid B_0=x\big\}\\
&= I(x>0)\cdot \mathrm{E}\big\{e^{-\alpha(B_\Delta-x)}v_\alpha^M(B_\Delta)\mid B_0=x\big\}\\
&= e^{\Delta\alpha^2/2}\cdot I(x>0)\cdot \mathrm{E}\big\{e^{-\Delta\alpha^2/2}e^{-\alpha(B_\Delta-x)}v_\alpha^M(B_\Delta)\mid B_0=x\big\}\\
&= e^{\Delta\alpha^2/2}\cdot I(x>0)\cdot \mathrm{E}_{-\alpha}\big\{v_\alpha^M(B_\Delta)\mid B_0=x\big\},
\end{aligned}
$$

where $E_{-\alpha}$ means Brownian motion with drift $-\alpha$ and the last equality follows from Girsanov theorem. Similarly,

$$
\begin{aligned}
v_\alpha^{M-2}(x) &= e^{\alpha x}\cdot I(x>0)\cdot \mathrm{E}\big\{I(B_\Delta>0)I(B_{2\Delta}>0)\mid B_0=x\big\}\\
&= I(x>0)\cdot \mathrm{E}\big\{e^{-\alpha(B_\Delta-x)}I(B_\Delta>0)\\
&\qquad\cdot \mathrm{E}\big\{e^{-\alpha(B_{2\Delta}-B_\Delta)}v^M(B_{2\Delta})\mid B_\Delta\big\}\mid B_0=x\big\}\\
&= I(x>0)\cdot \mathrm{E}\big\{e^{-\alpha(B_\Delta-x)}v^{M-1}(B_\Delta)\mid B_0=x\big\}\\
&= e^{\Delta\alpha^2/2}I(x>0)\cdot \mathrm{E}\big\{e^{-\Delta\alpha^2/2}e^{-\alpha(B_\Delta-x)}v^{M-1}(B_\Delta)\mid B_0=x\big\}\\
&= e^{\Delta\alpha^2/2}I(x>0)\cdot \mathrm{E}_{-\alpha}\big\{v_\alpha^{M-1}(B_\Delta)\mid B_0=x\big\}.
\end{aligned}
$$

In general, we have a backward induction

$$
\begin{aligned}
&v_\alpha^M(x) = e^{\alpha x}I(x>0),\\
&v_\alpha^{j-1}(x) = e^{\Delta\alpha^2/2}I(x>0)\cdot \mathrm{E}_{-\alpha}\big\{v^j(B_\Delta)\mid B_0=x\big\},\quad j=M,\ldots,2,\\
&p(x) = e^{-\alpha x}e^{\Delta\alpha^2/2}\cdot \mathrm{E}_{-\alpha}\big\{v^1(B_\Delta)\mid B_0=x\big\}.
\end{aligned}
$$

Denote $\hat{v}_\alpha^j(x)$ to be the Fourier transform of $v_\alpha^j(x)$, which is possible as $e^{\alpha x}I(x>0)$ is a L^1 function. Now the Fourier transform in the backward induction will involve Fourier transform of product of the indicator function and another function. The key observation in Feng and Linetsky (2005) is that Fourier transform of the product of the indicator function and a function can

be written in terms of Hilbert transform. More precisely,

$$\mathcal{F}(I_{(0,\infty)} \cdot f)(\xi) = \frac{1}{2}(\mathcal{F}f)(\xi) + \frac{i}{2}(\mathcal{H}f)(\xi),$$

where $\mathcal{F}$ denotes Fourier transform and $\mathcal{H}$ denotes Hilbert transform defined by the Cauchy principle value integral, i.e.

$$(\mathcal{H}f)(\xi) = \frac{1}{\pi} P.V. \int_{-\infty}^{\infty} \frac{f(\eta)}{\xi - \eta}\, \mathrm{d}\eta.$$

To compute $p(x)$ one needs to compute $M - 1$ Hilbert transforms and then conducts one final Fourier inversion. As shown in Feng and Linetsky (2005), a Hilbert transform can be computed efficiently by using approximation theory in Hardy spaces which leads to a simple trapezoidal-like quadrature sum.

In general Feng–Linetsky method is slower than Broadie–Yamamoto method, if the underlying model is Gaussian (e.g. under Black–Scholes model or Merton (1976) normal jump diffusion model). For example, as it is pointed out in Feng and Linetsky (2005) it may take 0.01 seconds for Broadie–Yamamoto to achieve accuracy of 10^{-12} under the Black–Scholes model, while it may take 0.04 seconds for Feng–Linetsky method to achieve accuracy of 10^{-8}. The beauty of Feng–Linetsky method is that it works for general Lévy processes with very reasonable computational time.

4 Continuity corrections

4.1 The approximation

Broadie et al. (1997) proposed a continuity correction for the discretely monitored barrier option, and justified the correction both theoretically and numerically (Chuang, 1996 independently suggested the approximation in a heuristic way). The resulting approximation, which only relies on a simple correction to the Merton (1973) formula (thus trivial to implement), is nevertheless quite accurate and has been used in practice; see, for example, the textbook by Hull (2005).

More precisely, let $V(H)$ be the price of a continuous barrier option, and $V_m(H)$ be the price of an otherwise identical barrier option with m monitoring points. Then for any of the eight discrete monitored regular barrier options the approximation is

$$V_m(H) = V\left(He^{\pm\beta\sigma\sqrt{T/m}}\right) + o(1/\sqrt{m}), \tag{2}$$

with $+$ for an up option and $-$ for a down option, where the constant $\beta = -\frac{\zeta(1/2)}{\sqrt{2\pi}} \approx 0.5826$, ζ the Riemann zeta function. The approximation (2) was

Table 1.
Up-and-Out-Call Option Price Results, $m = 50$ (daily monitoring).

Barrier	Continuous barrier	Corrected barrier, Eq. (2)	True	Relative error of Eq. (2) (in percent)
155	12.775	12.905	12.894	0.1
150	12.240	12.448	12.431	0.1
145	11.395	11.707	11.684	0.2
140	10.144	10.581	10.551	0.3
135	8.433	8.994	8.959	0.4
130	6.314	6.959	6.922	0.5
125	4.012	4.649	4.616	0.7
120	1.938	2.442	2.418	1.0
115	0.545	0.819	0.807	1.5

This table is taken from Broadie et al. (1997, Table 2.6). The option parameters are $S(0) = 110$, $K = 100$, $\sigma = 0.30$ per year, $r = 0.1$, and $T = 0.2$ year, which represents roughly 50 trading days.

proposed in Broadie et al. (1997), where it is proved for four cases: down-and-in call, down-and-out call, up-and-in put, and up-and-out put. Kou (2003) covered all eight cases with a simpler proof (see also Hörfelt, 2003). The continuity corrections for discrete lookback options are given in Broadie et al. (1999).

To get a feel of the accuracy of the approximation, Table 1 is taken from Broadie et al. (1997). The numerical results suggest that, even for daily monitored discrete barrier options, there can still be big differences between the discrete prices and the continuous prices. The improvement from using the approximation, which shifts the barrier from H to $He^{\pm\beta\sigma\sqrt{T/m}}$ in the continuous time formulae, is significant.

Cao and Kou (2007) derived some barrier correction formulae for two-dimensional barrier options and partial barrier options, which have some complications. For example, for a partial barrier option one cannot simply shifts the barrier up or down uniformly by a fixed constant, and one has to study carefully the different roles that the barrier plays in a partial barrier option; more precisely, the same barrier can sometimes be a terminal value, sometimes as a upcrossing barrier, and sometimes as a downcrossing barrier, all depending on what happens along the sample paths.

4.2 Continuity correction for random walk

The idea of continuity correction goes back to a classical technique in "sequential analysis," in which corrections to normal approximation are made to adjust for the "overshoot" effects when a discrete random walk crosses a barrier; see, for example, Chernoff (1965), Siegmund (1985), and Woodroofe (1982).

For a standard Brownian motion $B(t)$ under any probability space P, define the stopping times for discrete random walk and for continuous-time Brownian motion as

$$\tau'(b, U) := \inf\{n \geqslant 1\colon U_n \geqslant b\sqrt{m}\},$$
$$\tilde{\tau}'(b, U) := \inf\{n \geqslant 1\colon U_n \leqslant b\sqrt{m}\},$$
$$\tau(b, U) := \inf\{t \geqslant 0\colon U(T) \geqslant b\},$$
$$\tilde{\tau}(b, U) := \inf\{t \geqslant 0\colon U(T) \leqslant b\}.$$

Here $U(t) := vt + B(t)$ and U_n is a random walk with a small drift (as $m \to \infty$), $U_n := \sum_{i=1}^{n}(Z_i + \frac{v}{\sqrt{m}})$, where the Z_i's are independent standard normal random variables under P. Note that for general Lévy processes, we have the discrete random increments Z_i's are independent standard random variables, not necessarily normally distributed under P. In the case of Brownian motion, the approximation comes from a celebrated result in sequential analysis (Siegmund and Yuh, 1982; Siegmund, 1985, pp. 220–224): For any constants $b \geqslant y$ and $b > 0$, as $m \to \infty$,

$$P\bigl(U_m < y\sqrt{m}, \tau'(b, U) \leqslant m\bigr) = P\bigl(U(1) \leqslant y, \tau(b + \beta/\sqrt{m}, U) \leqslant 1\bigr) + o\bigl(1/\sqrt{m}\bigr). \tag{3}$$

Here the constant β is the limiting expectation of the overshoot,

$$\beta = \frac{\mathrm{E}(A_N^2)}{2\mathrm{E}(A_N)},$$

where the mean zero random walk A_n is defined as $A_n := \sum_{i=1}^{n} Z_i$, and N is the first ladder height associated with A_n, $N = \min\{n \geqslant 1\colon A_n > 0\}$. For general Lévy processes, there will be some extra terms in addition to the constant β.

4.2.1 An intuition via the reflection principle

To get an intuition of (3), we consider the reflection principle for the standard Brownian motions when the drift $v = 0$. The general case with a nonzero drift can be handled by using the likelihood ratio method. The reflection principle (see, e.g., Karatzas and Shreve, 1991) for the standard Brownian motion yields that

$$P\bigl(U(1) \leqslant y, \tau(b, U) \leqslant 1\bigr) = P\bigl(U(1) \geqslant 2b - y\bigr).$$

Intuitively, due to the random overshoot $R_m := U_{\tau'} - b\sqrt{m}$, the reflection principle for random walk should be

$$P\bigl(U_m < y\sqrt{m}, \tau'(b, U) \leqslant m\bigr) = P\bigl(U_m \geqslant 2(b\sqrt{m} + R_m) - y\sqrt{m}\bigr).$$

See Fig. 1 for an illustration.

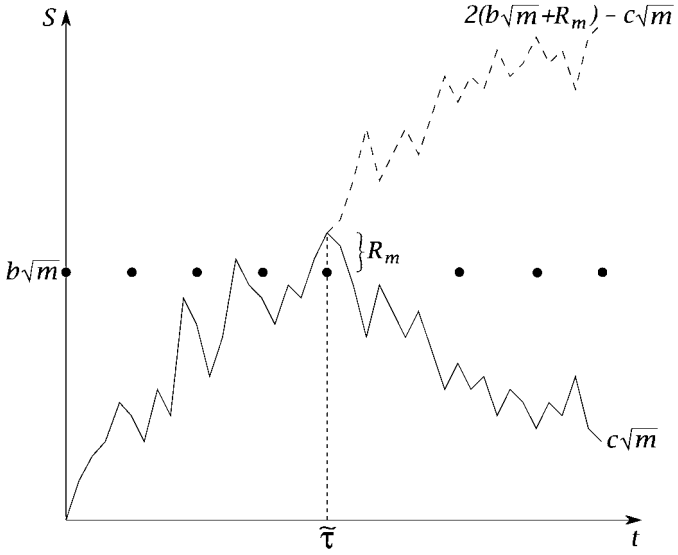


Fig. 1. An Illustration of the Heuristic Reflection Principle.

Replacing the random variable R_m by its is expectation $\mathrm{E}(R_m)$ and using the fact from the renewal theory that

$$\mathrm{E}(R_m) \to \frac{\mathrm{E}(A_N^2)}{2\mathrm{E}(A_N)} = \beta, \tag{4}$$

we have

$$\begin{aligned} &P\big(U_m < y\sqrt{m}, \tau'(b,U) \leqslant m\big) \\ &\quad \approx P\left(U_m \geqslant \left\{2\left(b + \frac{\beta}{\sqrt{m}}\right)\right\}\sqrt{m} - y\sqrt{m}\right) \\ &\quad \approx P\left(U(1) \geqslant 2\left(b + \frac{\beta}{\sqrt{m}}\right) - y\right) \\ &\quad = P\big(U(1) \leqslant y, \tau\big(b + \beta/\sqrt{m}, U\big) \leqslant 1\big), \end{aligned}$$

thus providing an intuition for (3).

4.2.2 Calculating the constant β

For any independent identically distributed random variables Z_i with mean zero and variance one there are two ways to compute β, one by infinite series and the other a one-dimensional integral.

In the first approach, we have the following result from Spitzer (1960) about $\mathrm{E}(A_N)$:

$$\mathrm{E}(A_N) = \frac{1}{\sqrt{2}} e^{\omega_0},$$

and from Lai (1976) about $\mathrm{E}(A_N^2)$:

$$\mathrm{E}(A_N^2) = \left\{\omega_2 + \frac{\mathrm{E}(Z_1^3)}{3\sqrt{2}} - \sqrt{2}\omega_1\right\} e^{\omega_0},$$

where

$$\omega_0 = \sum_{n=1}^{\infty} \frac{1}{n}\left(P\{A_n \leqslant 0\} - \frac{1}{2}\right),$$

$$\omega_2 = 1 - \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-\frac{1}{2}}{n}(-1)^n\right\},$$

$$\binom{x}{n} = x(x-1)\cdots(x-n+1)/n!,$$

$$\omega_1 = \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}\left(\mathrm{E}\left[\left(A_n/\sqrt{n}\right)^-\right] - \frac{1}{\sqrt{2\pi}}\right).$$

In the special of normal random variables, an explicit calculation of β is available. Indeed in this case we have $\omega_0 = 0$, $\omega_1 = 0$, $E(Z_1^3) = 0$, whence

$$\beta = \frac{\mathrm{E}(A_N^2)}{2\mathrm{E}(A_N)} = \frac{\{\omega_2 + \frac{\mathrm{E}(Z_1^3)}{3\sqrt{2}} - \sqrt{2}\omega_1\}e^{\omega_0}}{2\frac{1}{\sqrt{2}}e^{\omega_0}} = \frac{\omega_2}{\sqrt{2}}.$$

In Appendix B, we shall prove in the case of normal random variables, i.e. the Brownian model,

$$\beta = \frac{\mathrm{E}(A_N^2)}{2\mathrm{E}(A_N)} = \frac{\omega_2}{\sqrt{2}} = -\frac{\zeta(1/2)}{\sqrt{2\pi}} \tag{5}$$

with ζ being Riemann zeta function. Comparing to the existing proofs, the proof in Appendix B appears to be more direct and is new.

The link between β and the Riemann zeta function as in (5) has been noted by Chernoff (1965) in an optimal stopping problem via Wiener–Hopf integral equations. The links between Wiener–Hopf integral equations and the Riemann zeta function are advanced further by Howison and Steinberg (2005), who provide a very elegant second order expansion via the perturbation method and the Spitzer function. The proof that the constant calculated in Chernoff (1965) from Wiener–Hopf equations and the constant in Siegmund (1979) for the continuity correction are the same is given in Hogan (1986). Later Chang and Peres (1997) who give a much more general result regarding connections between ladder heights and Riemann zeta function in the case of normal random variables, which covers (5) as a special case. See also a related expansion in Blanchet and Glynn (2006), Asmussen et al. (1995). In Appendix B we shall prove (5) in the case of normal random variables directly without

using the general results in Chang and Peres (1997) or appealing to the argument in Hogan (1986).

There is also another integral representation (Siegmund, 1985, p. 225) for β, if Z_1 is a continuous random variable:

$$\beta = \frac{\mathrm{E}(Z_1^3)}{6} - \frac{1}{\pi}\int_0^\infty \frac{1}{\lambda^2}\,\mathrm{Re}\left\{\log\left(\frac{2(1-E(\exp\{i\lambda Z_1\}))}{\lambda^2}\right)\right\}\mathrm{d}\lambda. \tag{6}$$

In the case of normal random variables we have $\mathrm{E}(\exp\{i\lambda Z_1\}) = e^{-\lambda^2/2}$, and

$$\begin{aligned}\beta &= -\frac{1}{\pi}\int_0^\infty \frac{1}{\lambda^2}\log\left(\frac{1-e^{-\lambda^2/2}}{\lambda^2/2}\right)\mathrm{d}\lambda \\ &= -\frac{1}{\pi\sqrt{2}}\int_0^\infty \frac{1}{x^2}\log\left(\frac{1-e^{-x^2}}{x^2}\right)\mathrm{d}x.\end{aligned}$$

It is shown by Comtet and Majumdar (2005) that

$$\frac{1}{\pi}\int_0^\infty \frac{1}{x^2}\log\left(\frac{1-e^{-x^\alpha}}{x^\alpha}\right)\mathrm{d}x = \frac{\zeta(1/\alpha)}{(2\pi)^{1/\alpha}\sin(\frac{\pi}{2\alpha})}, \quad 1 < \alpha \leqslant 2.$$

In particular, letting $\alpha = 2$ yields

$$\frac{1}{\pi}\int_0^\infty \frac{1}{x^2}\log\left(\frac{1-e^{-x^2}}{x^2}\right)\mathrm{d}x = -\frac{\zeta(1/2)}{(2\pi)^{1/2}\sin(\frac{\pi}{4})} = \frac{\zeta(1/2)}{(\pi)^{1/2}},$$

and

$$\beta = -\frac{1}{\pi\sqrt{2}}\int_0^\infty \frac{1}{x^2}\log\left(\frac{1-e^{-x^2}}{x^2}\right)\mathrm{d}x = -\frac{\zeta(1/2)}{\sqrt{2\pi}}.$$

Comtet and Majumdar (2005) also evaluated (6) for other symmetric distributions, such as symmetric Laplace and uniform distributions.

4.2.3 A difficulty in generalization

The above theory of continuity correction depends crucially on the idea of asymptotic analysis of a random walk (in our case $\sum_{i=1}^n Z_i$) indexed by a single exponential family of random variables. In our case the exponential family has a base of $N(0,1)$ related to Z_i and the members in the family being $Z_i + v/\sqrt{m}$ with a distribution $N(v/\sqrt{m}, 1)$. In the general case, such as jump diffusion models, it is not clear how to write down a formula for the continuity correction for an exponential family of distribution indexed by a single parameter, because there could be several sources of randomness (Brownian parts, jump parts, etc.) and several parameters involved.

5 Perturbation method

Since the price of a discrete barrier option can be written as a solution to a partial differential equation (PDE) with piecewise linear boundary conditions, numerical techniques from PDEs are also useful. One particular numerical technique is the perturbation method, which formally matches various asymptotic expansions to get approximations. This has been used by Howison and Steinberg (2005), Howison (2005) to get very accurate approximation for discrete barrier and American options.

5.1 Basic concept of the perturbation method

The perturbation method first identifies a parameter to be small so that approximations can be made around the zero value of the parameter. In fact the perturbation method will identify two solutions, inner and outer solutions, to match boundary conditions. The final approximation is a sum of both solutions minus a matching constant. To illustrate the basic idea of the perturbation method, let us consider an ordinary differential equation

$$\varepsilon y'' + y' = t, \quad 0 < t < 1; \quad y(0) = y(1) = 1,$$

where the parameter ε is a small number. If we let $\varepsilon = 0$, then we get a solution $y = t^2/2 + C$. However this solution cannot satisfy both boundary conditions $y(0) = y(1) = 1$. To get around with this difficulty, we shall have two solutions, one near 0 (inner solution) and one near 1 (outer solution) so that the final approximation can properly combine the two (called "matching").

The outer solution is given by setting $\varepsilon = 0$ and matching the value at the right boundary,

$$y_1(t) = \frac{t^2}{2} + \frac{1}{2}, \quad y_1(1) = 1.$$

For the inner solution we can rescale the time, as we are more interested in what happen around $t = 0$. Using $s = t/\varepsilon$ and $A(s) = y(t)$, we have a rescaled equation

$$\frac{\varepsilon}{\varepsilon^2}\frac{\mathrm{d}^2 A}{\mathrm{d}s^2} + \frac{1}{\varepsilon}\frac{\mathrm{d}A}{\mathrm{d}s} = \varepsilon s, \quad \text{or} \quad \frac{\mathrm{d}^2 A}{\mathrm{d}s^2} + \frac{\mathrm{d}A}{\mathrm{d}s} = \varepsilon^2 s.$$

Letting $\varepsilon = 0$ yields a linear ordinary differential equation,

$$\frac{\mathrm{d}^2 A}{\mathrm{d}s^2} + \frac{\mathrm{d}A}{\mathrm{d}s} = 0,$$

which has a solution $A(s) = a + be^{-s}$. Changing it back to t we have the inner solution $y_2(t) = a + be^{-t/\varepsilon}$. Matching the boundary at 0, we have

$$y_2(t) = (1 - b) + be^{-t/\varepsilon}, \quad y_2(0) = 1.$$

Next we need to choose b to match the inner and outer solution at some immediate region after $t = 0$. To do this we try $u = t/\sqrt{\varepsilon}$. Then

$$y_1\big(u\sqrt{\varepsilon}\big) = \frac{u^2\varepsilon}{2} + \frac{1}{2} \to \frac{1}{2},$$
$$y_2\big(u\sqrt{\varepsilon}\big) = (1-b) + be^{-u\sqrt{\varepsilon}/\varepsilon} \to 1 - b,$$

yielding that $1 - b = 1/2$ or $b = 1/2$. In summary the outer and inner solutions are

$$y_1(t) = \frac{t^2}{2} + \frac{1}{2}, \qquad y_2(t) = \frac{1}{2} + \frac{1}{2}e^{-t/\varepsilon}.$$

Finally the perturbation approximation is given by the sum of the (matched) inner and outer solutions minus the common limiting value around time 0:

$$y_1(t) + y_2(t) - \lim_{\varepsilon\to 0} y_1\big(u\sqrt{\varepsilon}\big) = \left(\frac{t^2}{2} + \frac{1}{2}\right) + \left(\frac{1}{2} + \frac{1}{2}e^{-t/\varepsilon}\right) - \frac{1}{2}$$
$$= \frac{t^2}{2} + \frac{1}{2} + \frac{1}{2}e^{-t/\varepsilon}.$$

5.2 Howison–Steinberg approximation

Howison and Steinberg (2005) and Howison (2005) use both inner and outer solutions to get very accurate approximation for discrete barrier options and Bermudan (discrete American) options. Indeed, the approximation not only gives the first order correction as in Broadie et al. (1997), it also leads to a second order correction.

The outer expansion is carried out assuming that the underlying asset price is away from the barrier; in this case a barrier option price can be approximated by the price for the corresponding standard call and put options. The inner solution corresponds to the case when the asset price is close to the barrier.

Since the barrier crossing is only monitored at some discrete time points, the resulting inner solution will be a periodic heat equation. Howison and Steinberg (2005) present an elegant asymptotic analysis of the periodic heat equation by using the result of the Spitzer function (Spitzer, 1957, 1960) for the Wiener–Hopf equation. Afterwards, they matched the inner and outer solutions to get expansions. We shall not give the mathematical details here, and ask the interested reader to read the inspiring papers by Howison and Steinberg (2005) and Howison (2005).

In fact the approximation in Howison and Steinberg (2005) is so good that it can formally give the second order approximation with the order $o(1/m)$ for discrete barrier options, which is more accurate than the order $o(1/\sqrt{m})$ in the continuity correction in Broadie et al. (1997). The only drawback seems to be that perturbation methods generally lack rigorous proofs.

6 A Laplace transform method via Spitzer's identity

Building on the result in Ohgren (2001) and the Laplace transform (with respect to log-strike prices) introduced in Carr and Madan (1999), Petrella and Kou (2004) developed a method based on Laplace transform that easily allows us to compute the price and hedging parameters (the Greeks) of discretely monitored lookback and barrier options at *any* point in time, even if the previous achieved maximum (minimum) cannot be ignored. The method in Petrella and Kou (2004) can be implemented not only under the classical Brownian model, but also under more *general* models (e.g. jump-diffusion models) with stationary independent increments. A similar method using Fourier transforms in the case of pricing discrete lookback options at the monitoring points (but not at any time points hence with no discussion of the hedging parameters) was independently suggested in Borovkov and Novikov (2002, pp. 893–894). The method proposed in Petrella and Kou (2004) is more general, as it is applicable to price both discrete lookback and barrier options at any time points (hence to compute the hedging parameters).

6.1 Spitzer's identity and a related recursion

Consider the asset value $S(t)$, monitored in the interval $[0, T]$ at a sequence of equally spaced monitoring points, $0 \equiv t(0) < t(1) < \cdots < t(m) \equiv T$. Let $X_i := \log\{S(t(i))/S(t(i-1))\}$, where X_i is the return between $t(i-1)$ and $t(i)$. Denote $t(l)$ to be a monitoring point such that time t is between the $(l-1)$th and lth monitoring points, i.e., $t(l-1) \leqslant t < t(l)$. Define the maxima of the return process between the monitoring points to be $\widetilde{M}_{l,k} := \max_{l \leqslant j \leqslant k} \sum_{i=l+1}^{j} X_i$, $l = 0, \ldots, k$, where we have used the convention that the sum is zero if the index set is empty. Assume that $X_1, X_2, \ldots,$ are independent identically distributed (i.i.d.) random variables. With $X_{s,t} := \log\{S(t)/S(s)\}$ being the return between time s and time t, $t \geqslant s$, define

$$C(u, v; t) := \mathrm{E}^*\big[e^{uX_{t,t(l)}}\big]\mathrm{E}^*\big[e^{u\widetilde{M}_{l,m}+vX_{t,T}}\big] = \hat{x}_{l,m}\mathrm{E}^*\big[e^{(u+v)X_{t,t(l)}}\big], \quad (7)$$

where

$$\hat{x}_{l,k} := \mathrm{E}^*\big[e^{u\widetilde{M}_{l,k}+vB_{l,k}}\big], \quad l \leqslant k; \quad B_{l,k} := \sum_{i=l+1}^{k} X_i. \quad (8)$$

Define for $0 \leqslant l \leqslant k$,

$$\hat{a}_{l,k} := \mathrm{E}^*\big[e^{(u+v)B^+_{l,k}}\big] + \mathrm{E}^*\big[e^{-vB^-_{l,k}}\big] - 1, \quad u, v \in \mathbb{C}. \quad (9)$$

Spitzer (1956) proved that, for $s < 1$ and $u, v \in C$, with $\mathrm{Im}(u) \geqslant 0$ and $\mathrm{Im}(v) \geqslant 0$:

$$\sum_{k=0}^{\infty} s^k \hat{x}_{l,k} = \exp\left(\sum_{k=1}^{\infty} \frac{s^k}{k} \hat{a}_{l,k}\right), \quad (10)$$

where $B^+_{l,k}$ and $B^-_{l,k}$ denote the positive and negative part of the $B_{l,k}$, respectively. We can easily extend (10) to any $u, v \in C$, by limiting $s \leqslant s'_0$ for some s'_0 small enough.

To get $\hat{x}_{l,k}$ from $\hat{a}_{l,k}$ we can invert the Spitzer's identity by using Leibniz's formula at $s = 0$, as in Ohgren (2001). In fact, Petrella and Kou (2004) show that for any given l, we have

$$\hat{x}_{l,k+1} = \frac{1}{k-l+1} \sum_{j=0}^{k-l} \hat{a}_{l,k+1-j} \hat{x}_{l,l+j}. \tag{11}$$

To compute $\hat{a}_{l,k}$, when u and v are real numbers, Petrella and Kou (2004) also show that

$$\mathrm{E}^*[e^{uB^+_{l,k}}] = \begin{cases} 1 + \mathrm{E}^*\big[(e^{uB_{l,k}} - 1)\mathbf{1}_{\{uB_{l,k}>0\}}\big] = 1 + C_1(u,k), \\ \quad \text{if } u \geqslant 0, \\ 1 - \mathrm{E}^*\big[(1 - e^{uB_{l,k}})\mathbf{1}_{\{uB_{l,k}<0\}}\big] = 1 - P_1(u,k), \\ \quad \text{if } u < 0, \end{cases} \tag{12}$$

$$\mathrm{E}^*\big[e^{-vB^-_{l,k}}\big] = \begin{cases} 1 + \mathrm{E}^*\big[(e^{-vB_{l,k}} - 1)\mathbf{1}_{\{vB_{l,k}<0\}}\big] = 1 + C_1(-v,k), \\ \quad \text{if } v \geqslant 0, \\ 1 - \mathrm{E}^*\big[(1 - e^{-vB_{l,k}})\mathbf{1}_{\{vB_{l,k}>0\}}\big] = 1 - P_1(-v,k), \\ \quad \text{if } v < 0, \end{cases} \tag{13}$$

where $C_1(u,k)$ is the value of a European call option with strike $K = 1$ on the asset $\overline{S}_t$ with $\overline{S}_0 = 1$ and return $u \cdot X_{t(l),t(k)}$ (ignoring the discount factor), and $P_1(u,k)$ is the value of a European put option with strike $K = 1$ on the asset $\overline{S}_t$ with $\overline{S}_0 = 1$ and return $u \cdot X_{t(l),t(k)}$. In other words, we can easily compute $\hat{a}_{l,k}$ via analytical solutions of the standard call and put options.

6.2 *Laplace transform for discrete barrier options*

To save the space, we shall only discuss the case of barrier options, as the case of lookback options can be treated similarly; see Petrella and Kou (2004). Let $\xi > 1$ and $\zeta > 0$ and assume that $C(-\zeta, 1-\xi; t) < \infty$. At any time $t \in [t(l-1), t(l))$, $m \geqslant l \geqslant 1$, Petrellla and Kou (2004) show that the double Laplace transform of $f(\kappa, h; S(t)) = \mathrm{E}^*[(e^\kappa - S(T))^+ \mathbf{1}_{\{M_{0,T}<e^h\}} \mid \mathcal{F}_t]$ is given by

$$\begin{aligned} \hat{f}(\xi,\zeta) &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\xi\kappa-\zeta h} f\big(\kappa, h; S(t)\big)\,\mathrm{d}\kappa\,\mathrm{d}h \\ &= \big(S(t)\big)^{-(\xi+\zeta-1)} \cdot \frac{C(-\zeta, 1-\xi; t)}{\xi(\xi-1)\zeta}, \end{aligned} \tag{14}$$

with the function C defined in (7). The Greeks can also be computed similarly. For example, at any time $t \in [t(l-1), t(l))$, with $1 \leqslant l \leqslant m$, we have:

$$\frac{\partial}{\partial S(t)} UOP(t, T)$$
$$= -e^{-r(T-t)} \mathcal{L}^{-1}_{\xi,\zeta}\left(\frac{(\xi+\zeta-1)(S(t))^{-(\xi+\zeta)}}{\xi(\xi-1)\zeta} \times C(-\zeta, 1-\xi; t) \right)\Bigg|_{\log(K),\log(H)}.$$

To illustrate the algorithm, without loss of generality, we shall focus on computing the price and the hedging parameters (the Greeks) for an up-and-out put option.

The Algorithm:

Input: Analytical formulae of standard European call and put options.

Step 1: Use the European call and put formulae to calculate $\hat{a}_{i,k}$, via (9), (12) and (13).

Step 2: Use the recursion in Eq. (11) to compute $\hat{x}_{l,k}$.

Step 3: Compute $C(u, v; t)$ from Eq. (7).

Step 4: Numerically invert the Laplace transforms given in Eq. (14).

In Step 4 the Laplace transforms are inverted by using two-sided Euler inversion algorithms in Petrella (2004), which are extensions of one-sided Euler algorithms in Abate and Whitt (1992) and Choudhury et al. (1994).

The algorithm essentially only requires to input the standard European call and put prices, thanks to Spitzer's identity. The algorithm can also be extended to price other derivatives, whose values are a function of the joint distribution of the terminal asset value and its discretely monitored maximum (or minimum) throughout the lifetime of the option, such as partial lookback options. As demonstrated in the numerical examples in Petrella and Kou (2004), for a wide variety of parameters (including the cases where the barrier is very close to the initial asset price and there are many monitoring points), the algorithm is quite fast (typically only requires a few seconds), and is quite accurate (typically up to three decimal points). The total workload for both barrier and lookback options is of the order $O(NM^2)$, where N is the number terms needed for Laplace inversion, and M is the total number of monitoring points.

7 Which method to use

So far we have introduced four recent methods tailored to discrete barrier and lookback options. A natural question is which method is suitable for your particular needs. The answer really depends on four considerations: speed, accuracy, generality (e.g. whether a method is applicable to models beyond the standard Brownian model), and programming effort.

The consideration of programming effort is often ignored in the literature, although we think it is important. For example, the popularity of binomial trees and Monte Carlo methods in computational finance is a testimony that simple algorithms with little programming effort are better than faster but more complicated methods. This is also because the CPU time improves every year with increasing computer technology. Therefore, an algorithm not only compete with other algorithms but also with ever faster microprocessors. A tenfold increase in the computation speed of an algorithm is less important five years from now, but the simplicity of the algorithm will remain throughout time.

In terms of speed and programming effort, the fastest and easiest ones are the approximation methods, such as the continuity correction and Howison and Steinberg method, as they have analytical solutions. However, approximations will not yield exact results. More precisely, if you can tolerate about 5 to 10% pricing error (which is common in practice, as the bid–ask spreads for standard call/put options are in the range of 5 to 10% and the bid–ask spreads for barrier and lookback options are even more), then you should choose the approximation methods. A drawback for the approximation methods is that it is not clear how to generalize the approximations outside the classical Brownian model.

If accuracy is of great concern, e.g. when you need to set up some numerical benchmarks, then the "exact" methods will be needed. For example, if you use the standard Brownian model or models that only involves normal random variables (such as Merton's normal jump diffusion model), then Broadie–Yamamoto method via the fast Gaussian transform is perhaps the best choice, as it is very fast and accurate, and it is quite easy to implement.

However, if you want to price options under more general Levy processes for a broader class of return processes, including non-Gaussian distributions (e.g. the double exponential jump-diffusion model), which may not be easily written as a mixture of independent Gaussian random variables, then Feng–Linetsky or the Laplace transform via Spitzer's identity may be appropriate. Feng–Linetsky method is a powerful method that can produce very accurate answers in a fast way, and is faster than the Laplace method via Spitzer's identity; but it perhaps requires more programming effort (related to computing Hilbert transforms) than the Laplace transform method. Furthermore, it is very easy to compute, almost at no additional computational cost, the hedging parameters (the Greeks) using the Laplace transform method via Spitzer's identity.

Appendix A. Proof of (1)

By considering the events $\{\tau'(a/(\sigma\sqrt{T}), W) = 1\}$ and $\{\tau'(a/(\sigma\sqrt{T}), W) = 2\}$, we have

$$
\begin{aligned}
V_3(H) &= S(0)\sum_{i=1}^{2}\hat{P}\Big(W_3 \geqslant \frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}, \tau'\big(a\big/\big(\sigma\sqrt{T}\big), W\big) = i\Big) \\
&\quad - Ke^{-rT}\sum_{i=1}^{2}P^*\Big(W_3 \geqslant \frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}, \tau'\big(a\big/\big(\sigma\sqrt{T}\big), W\big) = i\Big) \\
&= S(0)\hat{P}\Big(-W_1 \leqslant -\frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, -W_3 \leqslant -\frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}\Big) \\
&\quad + S(0)\hat{P}\Big(W_1 < \frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, \\
&\qquad -W_2 \leqslant -\frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, \\
&\qquad -W_3 \leqslant -\frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}\Big) \\
&\quad - Ke^{-rT}P^*\Big(-W_1 \leqslant -\frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, \\
&\qquad -W_3 \leqslant -\frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}\Big) \\
&\quad - Ke^{-rT}P^*\Big(W_1 < \frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, \\
&\qquad -W_2 \leqslant -\frac{\log(H/S(0))}{\sigma\sqrt{T}}\sqrt{3}, \\
&\qquad -W_3 \leqslant \frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}}\Big).
\end{aligned}
$$

Observe the correlations

$$
\begin{aligned}
&\varrho(W_1, W_2) = \varrho(Z_1, Z_1 + Z_2) = \sqrt{1/2} = \sqrt{t_1/t_2}, \\
&\varrho(W_1, -W_3) = \varrho(Z_1, -Z_1 - Z_2 - Z_3) = -\sqrt{t_1/T}, \\
&\varrho(W_2, -W_3) = \varrho(Z_1 + Z_2, -Z_1 - Z_2 - Z_3) = -\sqrt{t_2/T},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{Var}(W_k) = k, \qquad \hat{\mathrm{E}}(W_k) = k\frac{r + \frac{1}{2}\sigma^2}{\sigma}\sqrt{\Delta t}, \\
&\mathrm{E}^*(W_k) = k\frac{r - \frac{1}{2}\sigma^2}{\sigma}\sqrt{\Delta t}, \quad k = 1, 2, 3.
\end{aligned}
$$

Note some identities for calculation related to $\hat{P}$

$$
\pm\frac{\log(H/S(0))}{\sigma\sqrt{\Delta t}} - \Big\{\Big(r + \frac{1}{2}\sigma^2\Big)\Big/\sigma\Big\}\sqrt{\Delta t}
$$

$$= \frac{\pm\log(H/S(0)) - \{(r+\frac{1}{2}\sigma^2)\}t_1}{\sigma\sqrt{t_1}} \equiv \hat{a}_{1,H,\pm},$$

$$\frac{1}{\sqrt{2}}\left(\pm\frac{\log(H/S(0))}{\sigma\sqrt{\Delta t}} - 2\left\{\left(r+\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\Delta t}\right)$$

$$= \frac{\pm\log(H/S(0)) - \{(r+\frac{1}{2}\sigma^2)\}t_2}{\sigma\sqrt{t_2}} \equiv \hat{a}_{2,H,\pm},$$

$$\frac{1}{\sqrt{3}}\left(-\frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}} - 3\left\{\left(r+\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\Delta t}\right)$$

$$= \frac{-\log(K/S(0)) - \{(r+\frac{1}{2}\sigma^2)\}T}{\sigma\sqrt{T}} \equiv \hat{a}_K,$$

and some identities for calculation related to P^*

$$\pm\frac{\log(H/S(0))}{\sigma\sqrt{\Delta t}} - \left\{\left(r-\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\Delta t}$$

$$= \frac{\pm\log(H/S(0)) - \{(r-\frac{1}{2}\sigma^2)\}t_1}{\sigma\sqrt{t_1}} \equiv a^*_{1,H,\pm},$$

$$\frac{1}{\sqrt{2}}\left(\pm\frac{\log(H/S(0))}{\sigma\sqrt{\Delta t}} - 2\left\{\left(r-\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\Delta t}\right)$$

$$= \frac{\pm\log(H/S(0)) - \{(r-\frac{1}{2}\sigma^2)\}t_2}{\sigma\sqrt{t_2}} \equiv a^*_{2,H,\pm},$$

$$\frac{1}{\sqrt{3}}\left(-\frac{\log(K/S(0))}{\sigma\sqrt{\Delta t}} - 3\left\{\left(r-\frac{1}{2}\sigma^2\right)\Big/\sigma\right\}\sqrt{\Delta t}\right)$$

$$= \frac{-\log(K/S(0)) - \{(r-\frac{1}{2}\sigma^2)\}T}{\sigma\sqrt{T}} \equiv a^*_K,$$

from which the conclusion follows.

Appendix B. Calculation of the constant β

First of all, we show that the series in (5)

$$\sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-1/2}{n}(-1)^n\right\} \tag{15}$$

converges absolutely. Using Stirling's formula (e.g. Chow and Teicher, 1997)

$$n! = n^n e^{-n}\sqrt{2\pi n}\cdot\varepsilon_n,\ e^{\frac{1}{12n+1}} < \varepsilon_n < e^{\frac{1}{12n}},$$

we have

$$\begin{aligned}\sqrt{\pi}\binom{-1/2}{n}(-1)^n &= \sqrt{\pi}\frac{1}{2}\frac{3}{2}\cdots\frac{(2n-1)}{2}/n! = \sqrt{\pi}\frac{(2n)!}{2^{2n}}\frac{1}{n!n!}\\ &= \sqrt{\pi}\frac{(2n)^{2n}e^{-2n}\sqrt{2\pi\cdot 2n}}{2^{2n}}\\ &\quad\times\frac{1}{n^n e^{-n}\sqrt{2\pi n}\cdot n^n e^{-n}\sqrt{2\pi n}}\frac{\varepsilon_{2n}}{\varepsilon_n\varepsilon_n}\\ &= \frac{1}{\sqrt{n}}\frac{\varepsilon_{2n}}{\varepsilon_n\varepsilon_n}.\end{aligned}$$

Since $\frac{\varepsilon_{2n}}{\varepsilon_n\varepsilon_n} = 1 + O(1/n)$, we have the terms inside the series (15):

$$\begin{aligned}\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-1/2}{n}(-1)^n = \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n}}\frac{\varepsilon_{2n}}{\varepsilon_n\varepsilon_n} &= \frac{1}{\sqrt{n}}\left(1 - \frac{\varepsilon_{2n}}{\varepsilon_n\varepsilon_n}\right)\\ &= O\left(\frac{1}{n\sqrt{n}}\right),\end{aligned}$$

from which we know that the series (15) converges absolutely.

Next, in the case of the standard normal density we have

$$\beta = \frac{E(A_N^2)}{2E(A_N)} = \frac{\omega_2}{\sqrt{2}} = \frac{1}{\sqrt{2}}\left(1 - \frac{1}{\sqrt{\pi}}\sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-\frac{1}{2}}{n}(-1)^n\right\}\right).$$

It was shown in Hardy (1905) that

$$\lim_{x\uparrow 1}\sum_{n=1}^{\infty}\left(\frac{x^n}{n^s} - \Gamma(1-s)\left\{\log\left(\frac{1}{x}\right)\right\}^{s-1}\right) = \zeta(s).$$

Taking $s = 1/2$ and using the fact that $\Gamma(1/2) = \sqrt{\pi}$, we have

$$\lim_{x\uparrow 1}\sum_{n=1}^{\infty}\frac{x^n}{\sqrt{n}} - \sqrt{\pi}\left\{\log\left(\frac{1}{x}\right)\right\}^{-1/2} = \zeta(1/2). \tag{16}$$

Furthermore, letting $x = 1 - \varepsilon$ yields

$$\begin{aligned}&\frac{1}{\sqrt{1-x}} - \left\{\log\left(\frac{1}{x}\right)\right\}^{-1/2}\\ &\quad= \frac{\sqrt{\log(1/x)} - \sqrt{1-x}}{\sqrt{\log(1/x)}\sqrt{1-x}}\\ &\quad= \frac{\log(1/x) - (1-x)}{\sqrt{\log(1/x)}\sqrt{1-x}(\sqrt{\log(1/x)} + \sqrt{1-x})}\\ &\quad= \frac{-\log(1-\varepsilon) - \varepsilon}{\sqrt{-\log(1-\varepsilon)}\sqrt{\varepsilon}(\sqrt{-\log(1-\varepsilon)} + \sqrt{\varepsilon})}\end{aligned}$$

$$= \frac{O(\varepsilon^2)}{O(\sqrt{\varepsilon})O(\sqrt{\varepsilon})\{O(\sqrt{\varepsilon}) + O(\sqrt{\varepsilon})\}} = O(\sqrt{\varepsilon}) \to 0,$$

as $x \uparrow 1$. Therefore, we have

$$\lim_{x\uparrow 1}\left(\frac{1}{\sqrt{1-x}} - \left\{\log\left(\frac{1}{x}\right)\right\}^{-1/2}\right) = 0.$$

This is interesting, as the both terms $1/\sqrt{1-x}$ and $\{\log(\frac{1}{x})\}^{-1/2}$ go to infinity as $x \uparrow 1$ but the difference goes to zero.

The above limit, in conjunction with (16), yields

$$\lim_{x\uparrow 1}\sum_{n=1}^{\infty}\frac{x^n}{\sqrt{n}} - \sqrt{\pi}\frac{1}{\sqrt{1-x}} = \zeta(1/2).$$

Since $(1-x)^{-\alpha} = \sum_{n=0}^{\infty}\binom{-\alpha}{n}(-x)^n$, we have

$$\lim_{x\uparrow 1}\sum_{n=1}^{\infty}\frac{x^n}{\sqrt{n}} - \sqrt{\pi}\sum_{n=1}^{\infty}\binom{-1/2}{n}(-x)^n - \sqrt{\pi}\binom{-1/2}{0}(-x)^0 = \zeta(1/2).$$

In other words,

$$\lim_{x\uparrow 1}\sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-1/2}{n}(-1)^n\right\}x^n = \sqrt{\pi} + \zeta(1/2),$$

and

$$\sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-1/2}{n}(-1)^n\right\} = \sqrt{\pi} + \zeta(1/2),$$

because the series (15) converges absolutely so that we can interchange the limit and summation.

In summary we have

$$\begin{aligned}\beta &= \frac{1}{\sqrt{2}}\left(1 - \frac{1}{\sqrt{\pi}}\sum_{n=1}^{\infty}\left\{\frac{1}{\sqrt{n}} - \sqrt{\pi}\binom{-\frac{1}{2}}{n}(-1)^n\right\}\right)\\ &= \frac{1}{\sqrt{2}}\left(1 - \frac{1}{\sqrt{\pi}}\{\sqrt{\pi} + \zeta(1/2)\}\right) = -\frac{\zeta(1/2)}{\sqrt{2\pi}}.\end{aligned}$$

References

Abate, J., Whitt, W. (1992). The Fourier series method for inverting transforms of probability distributions. *Queueing Systems* 10, 5–88.

Ahn, D.G., Figlewski, S., Gao, B. (1999). Pricing discrete barrier options with an adaptive mesh. *Journal of Derivatives* 6, 33–44.

Ait-Sahalia, F., Lai, T.L. (1997). Valuation of discrete barrier and hindsight options. *Journal of Financial Engineering* 6, 169–177.

Ait-Sahalia, F., Lai, T.L. (1998). Random walk duality and valuation of discrete lookback options. *Applied Mathematical Finance* 5, 227–240.

Andricopoulos, A., Widdicks, M., Duck, P., Newton, D. (2003). Universal option valuation using quadrature methods. *Journal of Financial Economics* 67, 447–471.

Andreasen, J. (1998). The pricing of discretely sampled Asian and lookback options: A change of numeraire approach. *Journal of Computational Finance* 2, 5–30.

Asmussen, S., Glynn, P., Pitman, J. (1995). Discretization error in simulation of one-dimensional reflecting Brownian motion. *Annals of Applied Probability* 5, 875–896.

Babbs, S. (1992). Binomial valuation of lookback options. Working paper. Midland Global Markets, London.

Blanchet, J., Glynn, P. (2006). Complete corrected diffusion approximations for the maximum of random walk. *Annals of Applied Probability*, in press.

Black, F., Cox, J. (1976). Valuing corporate debts: Some effects of bond indenture provisions. *Journal of Finance* 31, 351–367.

Borovkov, K., Novikov, A. (2002). On a new approach to calculating expectations for option pricing. *Journal of Applied Probability* 39, 889–895.

Boyle, P.P., Lau, S.H. (1994). Bumping up against the barrier with the binomial method. *Journal of Derivatives* 2, 6–14.

Boyle, P.P., Tian, Y. (1998). An explicit finite difference approach to the pricing of barrier options. *Applied Mathematical Finance* 5, 17–43.

Boyle, P.P., Tian, Y. (1999). Pricing lookback and barrier options under the CEV process. *Journal of Financial and Quantitative Analysis* 34, 241–264.

Broadie, M., Yamamoto, Y. (2003). A double-exponential fast Gauss transform algorithm for pricing discrete path-dependent options. Working paper. Columbia University. *Operations Research* (in press).

Broadie, M., Glasserman, P., Kou, S.G. (1997). A continuity correction for discrete barrier options. *Mathematical Finance* 7, 325–349.

Broadie, M., Glasserman, P., Kou, S.G. (1999). Connecting discrete and continuous path-dependent options. *Finance and Stochastics* 3, 55–82.

Cao, M., Kou, S.G. (2007). Continuity correction for two dimensional and partial barrier options. Working paper. Columbia University.

Carr, P., Madan, D.B. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance* 2, 61–73.

Chance, D.M. (1994). The pricing and hedging of limited exercise of caps and spreads. *Journal of Financial Research* 17, 561–584.

Chang, J.T., Peres, Y. (1997). Ladder heights, Gaussian random walks and the Riemann zeta function. *Annals of Probability* 25, 787–802.

Chen, N., Kou, S.G. (2005). Credit spreads, optimal capital structure, and implied volatility with endogenous default and jump risk. Preprint. Columbia University. *Mathematical Finance* (in press).

Chernoff, H. (1965). Sequential Tests for the Mean of a Normal Distribution IV. *Annals of Mathematical Statistics* 36, 55–68.

Cheuk, T., Vorst, T. (1997). Currency lookback options and the observation frequency: A binomial approach. *Journal of International Money and Finance* 16, 173–187.

Choudhury, G.L., Lucantoni, D.M., Whitt, W. (1994). Multidimensional transform inversion with applications to the transient M/M/1 queue. *Annals of Applied Probability* 4, 719–740.

Chow, Y.S., Teicher, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*, third ed. Springer, New York.

Chuang, C.S. (1996). Joint distributions of Brownian motion and its maximum, with a generalization to corrected BM and applications to barrier options. *Statistics and Probability Letters* 28, 81–90.

Comtet, A., Majumdar, S.N. (2005). Precise asymptotic for a random walker's maximum. Working paper. Institut Henri Poincaré, Paris, France.

Conze, R., Viswanathan, R. (1991). Path-dependent options: The case of lookback options. *Journal of Finance* 46, 1893–1907.
Davydov, D., Linetsky, V. (2001). Pricing and hedging path-dependent options under the CEV process. *Management Science* 47, 949–965.
Duan, J.C., Dudley, E., Gauthier, G., Simonato, J.G. (2003). Pricing discrete monitored barrier options by a Markov chain. *Journal of Derivatives* 10, 9–32.
Feng, L., Linetsky, V. (2005). Pricing discretely monitored barrier options and defaultable bonds in Lévy process models: A Hilbert transform approach. Working paper. Northwestern University. *Mathematical Finance* (in press).
Figlewski, S., Gao, B. (1999). The adaptive mesh model: A new approach to efficient option pricing. *Journal of Financial Economics* 53, 313–351.
Fusai, G., Recchioni, M.C. (2003). Analysis of quadrature methods for the valuation of discrete barrier options. Working paper. Universita del Piemonte.
Gatto, M., Goldman, M.B., Sosin, H. (1979). Path-dependent options: "buy at the low, sell at the high". *Journal of Finance* 34, 1111–1127.
Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer, New York.
Goldman, M.B., Sosin, H., Shepp, L. (1979). On contingent claims that insure ex post optimal stock market timing. *Journal of Finance* 34, 401–414.
Hardy, G.H. (1905). A method for determining the behavior of certain classes of power series near a singular point on the circle of convergence. *Proceedings London Mathematical Society* 2, 381–389.
Haug, E.G. (1999). Barrier put-call transformations. Working paper. Tempus Financial Engineering.
Heynen, R.C., Kat, H.M. (1994a). Crossing barriers. *Risk* 7 (June), 46–49. Correction (1995), *Risk* 8 (March), 18. Reprinted in: Jarrow, R. (Ed.), *Over the Rainbow: Developments in Exotic Options and Complex Swaps*. RISK/FCMC, London, pp. 179–182.
Heynen, R.C., Kat, H.M. (1994b). Partial barrier options. *Journal of Financial Engineering* 3, 253–274.
Heynen, R.C., Kat, H.M. (1995). Lookback options with discrete and partial monitoring of the underlying price. *Applied Mathematical Finance* 2, 273–284.
Hörfelt, P. (2003). Extension of the corrected barrier approximation by Broadie, Glasserman, and Kou. *Finance and Stochastics* 7, 231–243.
Hogan, M. (1986). Comment on a problem of Chernoff and Petkau. *Annals of Probability* 14, 1058–1063.
Howison, S. (2005). A matched asymptotic expansions approach to continuity corrections for discretely sampled options. Part 2: Bermudan options. Preprint. Oxford University.
Howison, S., Steinberg, M. (2005). A matched asymptotic expansions approach to continuity corrections for discretely sampled options. Part 1: Barrier options. Preprint. Oxford University.
Hull, J.C. (2005). *Options, Futures, and Other Derivative Securities*, fourth ed. Prentice Hall, Englewood Cliffs, NJ.
Hull, J., White, A. (1993). Efficient procedures for valuing European and American path-dependent options. *Journal of Derivatives* 1, 21–31.
Karatzas, I., Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*, second ed. Springer, New York.
Kat, H. (1995). Pricing lookback options using binomial trees: an evaluation. *Journal of Financial Engineering* 4, 375–397.
Kou, S.G. (2003). On pricing of discrete barrier options. *Statistica Sinica* 13, 955–964.
Kou, S.G., Wang, H. (2003). First passage times of a jump diffusion process. *Advances in Applied Probability* 35, 504–531.
Kou, S.G., Wang, H. (2004). Option pricing under a double exponential jump diffusion model. *Management Science* 50, 1178–1192.
Kuan, G., Webber, N.J. (2003). Valuing discrete barrier options on a Dirichlet lattice. Working paper. University of Exester, UK.
Kunitomo, N., Ikeda, M. (1992). Pricing Options with Curved Boundaries. *Mathematical Finance* 2, 275–298.
Lai, T.L. (1976). Asymptotic moments of random walks with applications to ladder variables and renewal theory. *Annals of Probability* 11, 701–719.

Leland, H.E., Toft, K. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *Journal of Finance* 51, 987–1019.
Longstaff, F.A. (1995). How much can marketability affect security values? *Journal of Finance* 50, 1767–1774.
Merton, R.C. (1973). Theory of rational option pricing. *Bell Journal of Economic Management Science* 4, 141–183.
Merton, R.C. (1974). On the pricing of corporate debts: the risky structure of interest rates. *Journal of Finance* 29, 449–469.
Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.
Ohgren, A. (2001). A remark on the pricing of discrete lookback options. *Journal of Computational Finance* 4, 141–147.
Petrella, G. (2004). An extension of the Euler method for Laplace transform inversion algorithm with applications in option pricing. *Operations Research Letters* 32, 380–389.
Petrella, G., Kou, S.G. (2004). Numerical pricing of discrete barrier and lookback options via Laplace transforms. *Journal of Computational Finance* 8, 1–37.
Reiner, E. (2000). Convolution methods for path-dependent options. Preprint. UBS Warburg Dillon Read.
Rich, D. (1996). The mathematical foundations of barrier option pricing theory. *Advances in Futures Options Research* 7, 267–312.
Ritchken, P. (1995). On pricing barrier options. *Journal of Derivatives* 3, 19–28.
Rubinstein, M., Reiner, E. (1991). Breaking down the barriers. *Risk* 4 (September), 28–35.
Schroder, M. (1999). Changes of numeraire for pricing futures, forwards and options. *Review of Financial Studies* 12, 1143–1163.
Siegmund, D. (1979). Corrected diffusion approximation in certain random walk problems. *Advances in Applied Probability* 11, 701–719.
Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.
Siegmund, D., Yuh, Y.-S. (1982). Brownian approximations for first passage probabilities. *Zeitschrift für Wahrsch. verw. Gebiete* 59, 239–248.
Spitzer, F. (1956). A combinatorial lemma and its application to probability theory. *Transactions of the American Mathematical Society* 82, 323–339.
Spitzer, F. (1957). The Wiener–Hopf equation whose kernel is a probability density. *Duke Mathematical Journal* 24, 327–343.
Spitzer, F. (1960). The Wiener–Hopf equation whose kernel is a probability density (II). *Duke Mathematical Journal* 27, 363–372.
Sullivan, M.A. (2000). Pricing discretely monitored barrier options. *Journal of Computational Finance* 3, 35–52.
Tian, Y. (1999). Pricing complex barrier options under general diffusion processes. *Journal of Derivatives* 6 (Fall), 51–73.
Tse, W.M., Li, L.K., Ng, K.W. (2001). Pricing discrete barrier and hindsight options with the tridiagonal probability algorithm. *Management Science* 47, 383–393.
Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*. Society for Industrial and Applied Mathematics, Philadelphia.
Zhang, P.G. (1998). *Exotic Options*, second ed. World Scientific, Singapore.
Zvan, R., Vetzal, K.R., Forsyth, P.A. (2000). PDE methods for pricing barrier options. *Journal of Economic Dynamics and Control* 24, 1563–1590.

This page intentionally left blank

PART III

Interest Rate and Credit Risk Models and Derivatives

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15009-X

Chapter 9

Topics in Interest Rate Theory

Tomas Björk

Department of Finance, Stockholm School of Economics, PO Box 6501, S-113 83 Stockholm, Sweden
E-mail: tomas.bjork@hhs.se

Abstract

The purpose of this chapter is to give an overview of some recent aspects of interest rate theory. After having recapitulated some basic results, we discuss forward rate models in the Heath–Jarrow–Morton framework. We then go on to a more detailed investigation of the geometric properties of the forward rate equation, such as consistency problems and finite dimensional realizations. The LIBOR market model is given a separate section. We end by showing how stochastic potential theory can be used to construct and analyze positive interest rate models.

1 Introduction

The purpose of this essay is to give an overview of some recent aspects of interest rate theory. The reader is assumed to be familiar with arbitrage theory, and basic interest rate theory including martingale models for the short rate and affine term structures. We do not assume familiarity with Heath–Jarrow–Morton. The list of topics below is subjective, largely reflecting my personal interests. The mathematical level is informal in the sense that we often are content with giving a heuristic argument, and silently assume that all objects under study are "regular enough" or "integrable enough."

At the end of each section there are Notes with references to the literature. For general information on arbitrage theory, see the textbooks (Bingham and Kiesel, 2004; Björk, 2003, and Duffie, 2001) which also provide chapters on interest rate theory. For monographs on interest rate theory see Brigo and Mercurio (2001) and Cairns (2004).

2 Basics

We consider a financial market model living on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$ where $\mathbf{F} = \{\mathcal{F}_t\}_{t \geqslant 0}$. 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. At this point we do not make any particular assumptions about P – it can be interpreted as the objective measure or as an equivalent martingale measure. We also assume that the measure Q is a martingale measure. Our main object of study is the zero coupon bond market, and we need some formal definitions.

Definition 2.1. A **zero coupon bond** with **maturity date** T, also called a T-bond, is a contract which guarantees the holder \$1 to be paid on the date T. The price at time t of a bond with maturity date T is denoted by $p(t, T)$.

We now make an assumption to guarantee the existence of a sufficiently rich bond market.

Assumption 2.1. We assume that

1. There exists a (frictionless) market for T-bonds for every $T > 0$.
2. For every fixed T, the process $\{p(t, T);\ 0 \leqslant t \leqslant T\}$ is an optional stochastic process with $p(t, t) = 1$ for all t.
3. For every fixed t, $p(t, T)$ is P-a.s. continuously differentiable in the T-variable. This partial derivative is often denoted by

$$p_T(t, T) = \frac{\partial p(t, T)}{\partial T}.$$

Given the given bond market above, one can define a (surprisingly large) number of **riskless interest rates**, and the basic construction is as follows. Suppose that we are standing at time t, and let us fix two other points in time S and T with $t < S < T$. The immediate project is to write a contract at time t which allows us to make an investment of \$1 at time S, and to have a **deterministic** rate of return, determined at the contract time t, over the interval $[S, T]$. This can easily be achieved as follows.

1. At time t we sell one S-bond. This will give us $\$p(t, S)$.
2. For this money we can by exactly $p(t, S)/p(t, T)$ T-bonds. Thus our net investment at time t equals zero.
3. At time S the S-bond matures, so we are obliged to pay out \$1.
4. At time T the T-bonds mature at \$1 a piece, so we will receive the amount $\$p(t, S)/p(t, T) \cdot 1$.
5. The net effect of all this is that, based on a contract at t, an investment of \$1 at time S has yielded $\$p(t, S)/p(t, T)$ at time T.

We now go on to compute the relevant interest rates implied by the construction above. We will use two (out of many possible) ways of quoting forward rates, namely as continuously compounded rates or as simple rates.

The **simple** forward rate (or **LIBOR** rate) L, is the solution to the equation

$$1 + (T - S)L = \frac{p(t, S)}{p(t, T)},$$

whereas the **continuously compounded** forward rate R is the solution to the equation

$$e^{R(T-S)} = \frac{p(t, S)}{p(t, T)}.$$

The simple rate notation is the one used in the market, whereas the continuously compounded notation is used in theoretical contexts. They are of course logically equivalent, and the formal definitions are as follows.

Definition 2.2.

1. The simple **forward rate for** $[S, T]$ **contracted at** t, henceforth referred to as the **LIBOR** forward rate, is defined as
$$L(t; S, T) = -\frac{p(t, T) - p(t, S)}{(T - S)p(t, T)}.$$
2. The simple **spot rate for** $[S, T]$, henceforth referred to as the **LIBOR** spot rate, is defined as
$$L(S, T) = -\frac{p(S, T) - 1}{(T - S)p(S, T)}.$$
3. The continuously compounded **forward rate for** $[S, T]$ **contracted at** t is defined as
$$R(t; S, T) = -\frac{\log p(t, T) - \log p(t, S)}{T - S}.$$
4. The **continuously compounded spot rate**, $R(S, T)$, for the period $[S, T]$ is defined as
$$R(S, T) = -\frac{\log p(S, T)}{T - S}.$$
5. The **instantaneous forward rate with maturity** T**, contracted at** t**,** is defined by
$$f(t, T) = -\frac{\partial \log p(t, T)}{\partial T}.$$
6. The instantaneous **short rate at time** t is defined by
$$r(t) = f(t, t).$$

We note that spot rates are forward rates where the time of contracting coincides with the start of the interval over which the interest rate is effective, i.e. $t = S$. The instantaneous forward rate, which will be of great importance below, is the limit of the continuously compounded forward rate when $S \to T$. It can thus be interpreted as the riskless interest rate, contracted at t, over the infinitesimal interval $[T, T + \mathrm{d}T]$.

We now go on to define the money account process B.

Definition 2.3. The **money account** process is defined by

$$B_t = e^{\int_0^t r(s)\,\mathrm{d}s},$$

i.e.

$$\begin{cases} \mathrm{d}B(t) = r(t)B(t)\,\mathrm{d}t, \\ B(0) = 1. \end{cases}$$

The interpretation of the money account is that you may think of it as describing a bank with the stochastic short rate r. It can also be shown that investing in the money account is equivalent to investing in a self-financing "rolling over" trading strategy, which at each time t consists entirely of "just maturing" bonds, i.e. bonds which will mature at $t + \mathrm{d}t$.

We recall the following fundamental result in arbitrage theory.

Theorem 2.1. *Let $X \in \mathcal{F}_T$ be a T-claim, i.e. a contingent claim paid out at time T, and let Q be the "risk neutral" martingale measure with B as the numeraire. Then the arbitrage free price is given by*

$$\Pi(t; X) = E^Q\Big[e^{-\int_0^t r_s\,\mathrm{d}s} X \mid \mathcal{F}_t\Big]. \tag{1}$$

In particular we have

$$p(t, T) = E^Q\Big[e^{-\int_0^t r_s\,\mathrm{d}s} \mid \mathcal{F}_t\Big]. \tag{2}$$

As an immediate consequence of the definitions we have the following useful formulas.

Lemma 2.1. *For $t \leqslant s \leqslant T$ we have*

$$p(t, T) = p(t, s) \cdot \exp\left\{-\int_s^T f(t, u)\,\mathrm{d}u\right\},$$

and in particular

$$p(t, T) = \exp\left\{-\int_t^T f(t, s)\,\mathrm{d}s\right\}.$$

3 Forward rate models

In this section we give a brief recap of forward rate models a long the lines of Heath–Jarrow–Morton.

3.1 The HJM framework

We now turn to the specification of the Heath–Jarrow–Morton (HJM) framework (see Heath et al., 1992). We start by specifying everything under a given objective measure P.

Assumption 3.1. We assume that, for every fixed $T > 0$, the forward rate $f(\cdot, T)$ has a stochastic differential which under the objective measure P is given by

$$\mathrm{d}f(t,T) = \alpha(t,T)\,\mathrm{d}t + \sigma(t,T)\,\mathrm{d}\bar{W}(t), \tag{3}$$

$$f(0,T) = f^{\star}(0,T), \tag{4}$$

where $\bar{W}$ is a (d-dimensional) P-Wiener process whereas $\alpha(\cdot, T)$ and $\sigma(\cdot, T)$ are adapted processes.

Note that conceptually equation (3) is one stochastic differential in the t-variable for each fixed choice of T. The index T thus only serves as a "mark" or "parameter" in order to indicate which maturity we are looking at. Also note that we use the observed forward rated curve $\{f^{\star}(0,T);\ T \geqslant 0\}$ as the initial condition. This will automatically give us a perfect fit between observed and theoretical bond prices at $t = 0$, thus relieving us of the task of inverting the yield curve.

Remark 3.1. It is important to observe that the HJM approach to interest rates is not a proposal of a specific **model**, like, for example, the Vasiček model. It is instead a **framework** to be used for analyzing interest rate models. Every short rate model can be equivalently formulated in forward rate terms, and for every forward rate model, the arbitrage free price of a contingent T-claim $\mathcal{X}$ will still be given by the pricing formula

$$\Pi(0;\mathcal{X}) = E^{Q}\Big[e^{-\int_0^T r(s)\,\mathrm{d}s} \cdot \mathcal{X}\Big],$$

where the short rate as usual is given by $r(s) = f(s,s)$.

We now show how bond price dynamics are induced by a given specification of forward rate dynamics.

Proposition 3.1. *If the forward rate dynamics are given by (3) then the induced bond price dynamics are given by*

$$\begin{aligned} \mathrm{d}p(t,T) = p(t,T)\left\{r(t) + A(t,T) + \frac{1}{2}\left\|S(t,T)\right\|^2\right\}\mathrm{d}t \\ + p(t,T)S(t,T)\,\mathrm{d}W(t), \end{aligned}$$

where $\|\cdot\|$ *denotes the Euclidean norm, and*

$$\begin{cases} A(t,T) = -\int_t^T \alpha(t,s)\,\mathrm{d}s, \\ S(t,T) = -\int_t^T \sigma(t,s)\,\mathrm{d}s. \end{cases} \tag{5}$$

Proof. We give a slightly heuristic argument. The full formal proof, see Heath et al. (1992), is an integrated version of the proof given here, but the infinitesimal version below is (hopefully) easier to understand. Using the definition of the forward rates we may write

$$p(t,T) = e^{Y(t,T)}, \tag{6}$$

where Y is given by

$$Y(t,T) = -\int_t^T f(t,s)\,\mathrm{d}s. \tag{7}$$

From the Itô formula we then obtain the bond dynamics as

$$\mathrm{d}p(t,T) = p(t,T)\,\mathrm{d}Y(t,T) + \frac{1}{2}p(t,T)\big(\mathrm{d}Y(t,T)\big)^2, \tag{8}$$

and it remains to compute $\mathrm{d}Y(t,T)$. We have

$$\mathrm{d}Y(t,T) = -\mathrm{d}\left(\int_t^T f(t,s)\,\mathrm{d}s\right),$$

and the problem is that in the integral the t-variable occurs in two places: as the lower limit of integration, and in the integrand $f(t,s)$. This is a situation that is not covered by the standard Itô formula, but it is easy to guess the answer. The t appearing as the lower limit of integration should give rise to the term

$$\frac{\partial}{\partial t}\left(\int_t^T f(t,s)\,\mathrm{d}s\right)\mathrm{d}t.$$

Furthermore, since the stochastic differential is a linear operation, we should be allowed to move it inside the integral, thus providing us with the term

$$\left(\int_t^T \mathrm{d}f(t,s)\,\mathrm{d}s\right).$$

We have therefore arrived at

$$\mathrm{d}Y(t,T) = -\frac{\partial}{\partial t}\left(\int_t^T f(t,s)\,\mathrm{d}s\right)\mathrm{d}t - \int_t^T \mathrm{d}f(t,s)\,\mathrm{d}s,$$

which, using the fundamental theorem of integral calculus, as well as the forward rate dynamics, gives us

$$\mathrm{d}Y(t,T) = f(t,t)\,\mathrm{d}t - \int_t^T \alpha(t,s)\,\mathrm{d}t\,\mathrm{d}s - \int_t^T \sigma(t,s)\,\mathrm{d}W_t\,\mathrm{d}s.$$

We now use a stochastic Fubini Theorem, allowing us to exchange $\mathrm{d}t$ and $\mathrm{d}W_t$ with $\mathrm{d}s$. We also recognize $f(t,t)$ as the short rate $r(t)$, thus obtaining

$$\mathrm{d}Y(t,T) = r(t)\,\mathrm{d}t + A(t,T)\,\mathrm{d}t + S(t,T)\,\mathrm{d}W_t,$$

with A and S as above. We therefore have

$$\big(\mathrm{d}Y(t,T)\big)^2 = \big\|S(t,T)\big\|^2\,\mathrm{d}t,$$

and, substituting all this into (8), we obtain our desired result. □

3.2 *Absence of arbitrage*

Suppose now that we have specified α, σ and $\{f^\star(0,T);\ T \geqslant 0\}$. Then we have specified the entire forward rate structure and thus, by the relation

$$p(t,T) = \exp\left\{-\int_t^T f(t,s)\,\mathrm{d}s\right\}, \tag{9}$$

we have in fact specified the entire term structure $\{p(t,T);\ T > 0,\ 0 \leqslant t \leqslant T\}$. Since we have d sources of randomness (one for every Wiener process), and an infinite number of traded assets (one bond for each maturity T), we run a clear risk of having introduced arbitrage possibilities into the bond market. The first question we pose is thus very natural: How must the processes α and σ be related in order that the induced system of bond prices admits no arbitrage possibilities? The answer is given by the HJM drift condition below.

Theorem 3.1 *(HJM Drift Condition). Assume that the family of forward rates is given by (3) and that the induced bond market is arbitrage free. Then there exists a d-dimensional column-vector process*

$$\lambda(t) = \big[\lambda_1(t), \ldots, \lambda_d(t)\big]'$$

with the property that for all $T \geqslant 0$ and for all $t \leqslant T$, we have

$$\alpha(t,T) = \sigma(t,T)\int_t^T \sigma(t,s)'\,\mathrm{d}s - \sigma(t,T)\lambda(t). \tag{10}$$

In these formulas $'$ denotes transpose.

Proof. From Proposition 3.1 we have the bond dynamics

$$\begin{aligned} \mathrm{d}p(t,T) &= p(t,T)\left\{r(t)+A(t,T)+\frac{1}{2}\left\|S(t,T)\right\|^2\right\}\mathrm{d}t \\ &\quad + p(t,T)S(t,T)\,\mathrm{d}\bar{W}(t), \end{aligned} \tag{11}$$

where

$$\begin{cases} A(t,T) = -\int_t^T \alpha(t,s)\,\mathrm{d}s, \\ S(t,T) = -\int_t^T \sigma(t,s)\,\mathrm{d}s. \end{cases} \tag{12}$$

We now look for a candidate martingale measure Q so we perform a Girsanov transformation by specifying the dynamics of the likelihood process as

$$\mathrm{d}L(t) = L(t)\lambda'(t)\,\mathrm{d}\bar{W}(t), \tag{13}$$

$$L(0) = 1, \tag{14}$$

where

$$L(t) = \frac{\mathrm{d}Q}{\mathrm{d}P}, \quad \text{on } \mathcal{F}_t.$$

From the Girsanov Theorem we know that

$$\mathrm{d}\bar{W}(t) = \lambda(t)\,\mathrm{d}t + \mathrm{d}W(t),$$

where W is Q-Wiener so, substituting this into (11) gives us the bond price Q-dynamics as

$$\mathrm{d}p(t,T) = p(t,T)\left\{r(t)+A(t,T)+\frac{1}{2}\left\|S(t,T)\right\|^2+S(t,T)\lambda(t)\right\}\mathrm{d}t \tag{15}$$

$$+ p(t,T)S(t,T)\,\mathrm{d}W(t). \tag{16}$$

Furthermore, Q is a martingale measure with the money account B as the numeraire if and only if the local rate of return of every asset price under Q equals the short rate. We thus have

$$r(t)+A(t,T)+\frac{1}{2}\left\|S(t,T)\right\|^2+S(t,T)\lambda(t)=0.$$

Taking the T-derivative of this equation gives us Eq. (10). □

3.3 Martingale modeling

As a special case we now turn to the question of martingale modeling, and thus assume that the forward rates are specified directly under a martingale

measure Q as

$$\mathrm{d}f(t,T) = \alpha(t,T)\,\mathrm{d}t + \sigma(t,T)\,\mathrm{d}W(t), \tag{17}$$

$$f(0,T) = f^{\star}(0,T), \tag{18}$$

where W is a (d-dimensional) Q-Wiener process. Since a martingale measure automatically provides arbitrage free prices, we no longer have a problem of absence of arbitrage, but instead we have another problem. This is so because we now have the following two different formulas for bond prices

$$p(0,T) = \exp\left\{-\int_0^T f(0,s)\,\mathrm{d}s\right\},$$

$$p(0,T) = E^Q\left[\exp\left\{-\int_0^T r(s)\,\mathrm{d}s\right\}\right],$$

where the short rate r and the forward rates f are connected by $r(t) = f(t,t)$. In order for these formulas to hold simultaneously, we have to impose some sort of consistency relation between α and σ in the forward rate dynamics. The result is the famous Heath–Jarrow–Morton drift condition.

Proposition 3.2 *(HJM Drift Condition). Under the martingale measure Q, the processes α and σ must satisfy the following relation, for every t and every $T \geqslant t$.*

$$\alpha(t,T) = \sigma(t,T)\int_t^T \sigma(t,s)'\,\mathrm{d}s. \tag{19}$$

Proof. We only need to observe that if we start by modeling directly under the martingale measure, then we may apply Proposition 3.1 with $\lambda = 0$. □

The moral of Proposition 3.2 is that when we specify the forward rate dynamics (under Q) we may freely specify the volatility structure. The drift parameters are then uniquely determined.

To see at least how part of this machinery works we now study the simplest example conceivable, which occurs when the process σ is a deterministic constant. With a slight abuse of notation let us thus write $\sigma(t,T) \equiv \sigma$, where $\sigma > 0$. Equation (19) gives us the drift process as

$$\alpha(t,T) = \sigma\int_t^T \sigma\,\mathrm{d}s = \sigma^2(T-t), \tag{20}$$

so by integrating Eq. (3) we obtain

$$f(t,T) = f^{\star}(0,T) + \int_0^t \sigma^2(T-s)\,\mathrm{d}s + \int_0^t \sigma\,\mathrm{d}W(s), \tag{21}$$

i.e.

$$f(t,T) = f^{\star}(0,T) + \sigma^2 t\left(T - \frac{t}{2}\right) + \sigma W(t). \tag{22}$$

In particular we see that r is given as

$$r(t) = f(t,t) = f^{\star}(0,t) + \sigma^2\frac{t^2}{2} + \sigma W(t), \tag{23}$$

so the short rate dynamics are

$$\mathrm{d}r(t) = \left\{f_T(0,t) + \sigma^2 t\right\}\mathrm{d}t + \sigma\,\mathrm{d}W(t), \tag{24}$$

which is exactly the Ho–Lee model, fitted to the initial term structure. Observe in particular the ease with which we obtained a perfect fit to the initial term structure.

3.4 The Musiela parameterization

In many practical applications it is more natural to use time **to** maturity, rather than time **of** maturity, to parameterize bonds and forward rates. If we denote running time by t, time of maturity by T, and time to maturity by x, then we have $x = T - t$, and in terms of x the forward rates are defined as follows.

Definition 3.1. For all $x \geqslant 0$ the forward rates $r(t,x)$ are defined by the relation

$$r(t,x) = f(t,t+x). \tag{25}$$

Suppose now that we have the standard HJM-type model for the forward rates under a martingale measure Q

$$\mathrm{d}f(t,T) = \alpha(t,T)\,\mathrm{d}t + \sigma(t,T)\,\mathrm{d}W(t). \tag{26}$$

The question is to find the Q-dynamics for $r(t,x)$, and we have the following result, known as the Musiela equation.

Proposition 3.3 *(The Musiela Equation). Assume that the forward rate dynamics under Q are given by (26). Then*

$$\mathrm{d}r(t,x) = \left\{\mathbf{F}r(t,x) + D(t,x)\right\}\mathrm{d}t + \sigma_0(t,x)\,\mathrm{d}W(t), \tag{27}$$

where

$$\sigma_0(t,x) = \sigma(t,t+x),$$

$$D(t,x) = \sigma_0(t,x)\int_0^x \sigma_0(t,s)'\,\mathrm{d}s,$$

$$\mathbf{F} = \frac{\partial}{\partial x}.$$

Proof. We give a heuristic proof which can be made precise. Using a slight variation of the Itô formula we have

$$\mathrm{d}r(t,x) = \mathrm{d}f(t,t+x) + \frac{\partial f}{\partial T}(t,t+x)\,\mathrm{d}t,$$

where the differential in the term $\mathrm{d}f(t,t+x)$ only operates on the first t. We thus obtain

$$\mathrm{d}r(t,x) = \alpha(t,t+x)\,\mathrm{d}t + \sigma(t,t+x)\,\mathrm{d}W(t) + \frac{\partial}{\partial x}r(t,x)\,\mathrm{d}t,$$

and, using the HJM drift condition, we obtain our result. □

The point of the Musiela parameterization is that it highlights Eq. (27) as an infinite dimensional SDE. It has become an indispensable tool of modern interest rate theory and we will use it repeatedly below.

3.5 Notes

The forward rate methodology was introduced in the seminal paper Heath et al. (1992). The Musiela parameterization was developed in Brace and Musiela (1994), and Musiela (1993).

4 Change of numeraire

In this section we will give a very brief account of the change of numeraire technique. We will then use the results in Section 5. All the results are standard. See Björk (2003) for more details and bibliographic information.

4.1 Generalities

Consider as given a financial market (not necessarily a bond market) with the usual locally risk free asset B, and a risk neutral martingale measure Q. We recall from general theory that a measure is a martingale measure only relative to some chosen numeraire asset, and we recall that the risk neutral martingale measure, with the money account B as numeraire, has the property of martingalizing all processes of the form $S(t)/B(t)$ where S is the arbitrage free price process of any traded asset.

Assumption 4.1. Assume that Q is a fixed risk neutral martingale measure, and $S_0(t)$ is a strictly positive process with the property that the process $S_0(t)/B(t)$ is a Q-martingale.

The economic interpretation of this assumption is of course that $S_0(t)$ is the arbitrage free price process of a traded asset. We are now searching for a measure $Q^\star$ with the property that, for every arbitrage free price process $\Pi(t)$, the process $\Pi(t)/S_0(t)$ is a $Q^\star$-martingale.

In order to get an idea of what $Q^\star$ must look like, let us consider a fixed time T and a T-contract X. Assuming enough integrability we then know that the arbitrage free price of X at time $t = 0$ is given by

$$\Pi(0; X) = E^Q\left[\frac{X}{B(T)}\right]. \tag{28}$$

Assume, on the other hand, that the measure $Q^\star$ actually exists, with a Radon–Nikodym derivative process

$$L(t) = \frac{\mathrm{d}Q^\star}{\mathrm{d}Q}, \quad \text{on } \mathcal{F}_t.$$

Then we know that, because of the assumed $Q^\star$-martingale property of the process $\Pi(t; X)/S_0(t)$, we have

$$\frac{\Pi(0; X)}{S_0(0)} = E^\star\left[\frac{\Pi(T; X)}{S_0(T)}\right] = E^\star\left[\frac{X}{S_0(T)}\right] = E^Q\left[L(T)\frac{X}{S_0(T)}\right].$$

Thus we have

$$\Pi(0; X) = E^Q\left[L(T)\frac{X \cdot S_0(0)}{S_0(T)}\right], \tag{29}$$

and, comparing (28) with (29), we see that a natural candidate as likelihood process for the intended change of measure is given by $L(t) = S_0(t)/S_0(0) \times B(t)$.

We now go on to the formal definitions and results.

Definition 4.1. Under Assumption 4.1 define, for any fixed t, the measure $Q^\star$ on $\mathcal{F}_t$ by

$$\frac{\mathrm{d}Q^\star}{\mathrm{d}Q} = L(t), \tag{30}$$

where the likelihood process L is defined by

$$L(t) = \frac{S_0(t)}{S_0(0) \cdot B(t)}. \tag{31}$$

We note at once that L is a positive Q-martingale with $L(0) = 1$, so the measure $Q^\star$ is indeed a probability measure. We now want to prove that $Q^\star$ martingalizes every process of the form $\Pi(t)/S_0(t)$, where $\Pi(t)$ is any arbitrage free price process. The formalization of this idea is the following result.

Proposition 4.1. *Define $Q^\star$ as above. Assume that $\Pi(t)$ is a process such that $\Pi(t)/B(t)$ is a Q-martingale. Then the process $\Pi(t)/S_0(t)$ is a $Q^\star$-martingale.*

Proof. Denoting integration with respect to $Q^\star$ by $E^\star$, and using the abstract Bayes's formula, we obtain

$$E^\star\left[\frac{\Pi(t)}{S_0(t)} \mid \mathcal{F}_s\right] = \frac{E^Q\left[L(t)\frac{\Pi(t)}{S_0(t)} \mid \mathcal{F}_s\right]}{L(s)} = \frac{E^Q\left[\frac{\Pi(t)}{B(t)S_0(0)} \mid \mathcal{F}_s\right]}{L(s)}$$
$$= \frac{\Pi(s)}{B(s)S_0(0)L(s)} = \frac{\Pi(s)}{S_0(s)}. \qquad \square$$

As an immediate corollary we have the following.

Proposition 4.2. *Define $Q^\star$ as above and consider a T-claim X such that $X/B(T) \in L^1(Q)$. Then the price process, $\Pi(t; X)$ is given by*

$$\Pi(t; X) = S_0(t)E^\star\left[\frac{X}{S_0(T)} \mid \mathcal{F}_t\right]. \tag{32}$$

Remark 4.1. Note that it is easy to find the Girsanov transformation which carries Q into $Q^\star$. Since $Q^\star$ martingalizes the process $S_0(t)/B(t)$, the $Q^\star$-dynamics of S_0 must be of the form

$$\mathrm{d}S_0(t) = r(t)S_0(t)\,\mathrm{d}t + S_0(t)v(t)\,\mathrm{d}M(t) \tag{33}$$

where M is the driving Q-martingale of S_0 (typically M is a Wiener process), and v is the volatility for S_0. From (33) and (31) it now follows that the likelihood process L has the Q-dynamics

$$\mathrm{d}L(t) = L(t)v(t)\,\mathrm{d}M(t), \tag{34}$$

so we can easily read off the relevant Girsanov kernel directly from the volatility of the S_0-process.

4.2 Forward measures

In this section we specialize the theory developed in the previous section to the case when the new numeraire chosen is a bond maturing at time T. As can be expected, this choice of numeraire is particularly useful when dealing with interest rate derivatives.

Suppose therefore that we are given a specified bond market model with a fixed martingale measure Q. For a fixed time of maturity T we now choose the process $p(t, T)$ as our new numeraire.

Definition 4.2. The **T-forward measure** Q^T is defined by

$$\mathrm{d}Q^T = L^T(t)\,\mathrm{d}Q$$

on $\mathcal{F}_t$ for $0 \leqslant t \leqslant T$ where

$$L^T(t) = \frac{p(t, T)}{B(t)\,p(0, T)}. \tag{35}$$

Observing that $P(T, T) = 1$ we have the following useful pricing formula as an immediate corollary of Proposition 4.2.

Proposition 4.3. *Assume that the T-claim X has the property that* $X/B(T) \in L^1(Q)$. *Then*

$$\Pi(t; X) = p(t, T)E^T[X \mid \mathcal{F}_t], \tag{36}$$

where E^T *denotes integration w.r.t.* Q^T.

5 LIBOR market models

In the previous chapters we have concentrated on studying interest rate models based on *infinitesimal* interest rates like the instantaneous short rate and the instantaneous forward rates. While these objects are nice to handle from a mathematical point of view, they have two main disadvantages.

- The instantaneous short and forward rates can never be observed in real life.
- If you would like to calibrate your model to cap- or swaption data, then this is typically very complicated from a numerical point of view if you use one of the "instantaneous" models.

A further fact from real life, which has been somewhat disturbing from a theoretical point of view is the following.

- For a very long time, the market practice has been to value caps, floors, and swaptions by using a formal extension of the Black (1976) model. Such an extension is typically obtained by an approximation argument where the short rate at one point in the argument is assumed to be deterministic, while later on in the argument the LIBOR rate is assumed to be stochastic. This is of course logically inconsistent.
- Despite this, the market happily continues to use Black-76 for the pricing of caps, floors, and swaptions.

In a situation like this, where market practice seems to be at odds with academic work there are two possible attitudes for the theorist: you can join them (the market) or you can try to beat them, and since the fixed income market does not seem to collapse because of the use of Black-76, the more realistic alternative seems to be to join them.

Thus there has appeared a natural demand for constructing logically consistent (and arbitrage free!) models having the property that the theoretical prices for caps, floors and swaptions produced by the model are of the Black-76 form. This project has in fact been carried out very successfully, starting with Miltersen et al. (1997), Brace et al. (1997) and Jamshidian (1997). The basic structure of the models is as follows.

- In stead of modeling instantaneous interest rates, we model discrete **market rates** like LIBOR rates in the LIBOR market models, or forward swap rates in the swap market models.
- Under a suitable choice of numeraire(s), these market rates can in fact be modeled log normally.
- The market models will thus produce pricing formulas for caps and floors (the LIBOR models), and swaptions (the swap market models) which are of the Black-76 type and thus conforming with market practice.
- By construction the market models are thus very easy to calibrate to market data for caps/floors and swaptions respectively. They are then used to price more exotic products. For this later pricing part, however, we will typically have to resort to some numerical method, like Monte Carlo.

5.1 Caps: definition and market practice

In this section we discuss LIBOR caps and the market practice for pricing and quoting these instrument. To this end we consider a fixed set of increasing maturities $T_0, T_1, \ldots, T_N$ and we define α_i, by

$$\alpha_i = T_i - T_{i-1}, \quad i = 1, \ldots, N.$$

The number α_i is known as the **tenor**, and in a typical application we could for example have all α_i equal to a quarter of a year.

Definition 5.1. We let $p_i(t)$ denote the zero coupon bond price $p(t, T_i)$ and let $L_i(t)$ denote the LIBOR forward rate, contracted at t, for the period $[T_{i-1}, T_i]$, i.e.

$$L_i(t) = \frac{1}{\alpha_i} \cdot \frac{p_{i-1}(t) - p_i(t)}{p_i(t)}, \quad i = 1, \ldots, N. \tag{37}$$

We recall that a **cap** with **cap rate** R and **resettlement dates** $T_0, \ldots, T_N$ is a contract which at time T_i gives the holder of the cap the amount

$$X_i = \alpha_i \cdot \max\big[L_i(T_{i-1}) - R, 0\big], \tag{38}$$

for each $i = 1, \ldots, N$. The cap is thus a portfolio of the individual **caplets** $X_1, \ldots, X_N$. We note that the forward rate $L_i(T_{i-1})$ above is in fact the spot rate at time T_{i-1} for the period $[T_{i-1}, T_i]$, and determined already at time T_{i-1}. The amount X_i is thus determined at T_{i-1} but not payed out until at time T_i. We also note that, formally speaking, the caplet X_i is a call option on the underlying spot rate.

The market practice is to use the Black-76 formula for the pricing of caplets.

Definition 5.2 *(Black's Formula for Caplets)*. The Black-76 formula for the caplet

$$X_i = \alpha_i \cdot \max\big[L(T_{i-1}, T_i) - R, 0\big], \tag{39}$$

is given by the expression

$$\mathbf{Capl_i^B}(t) = \alpha_i \cdot p_i(t)\big\{L_i(t)N[d_1] - RN[d_2]\big\}, \quad i = 1, \ldots, N, \tag{40}$$

where

$$d_1 = \frac{1}{\sigma_i\sqrt{T_i - t}}\left[\ln\left(\frac{L_i(t)}{R}\right) + \frac{1}{2}\sigma_i^2(T - t)\right], \tag{41}$$

$$d_2 = d_1 - \sigma_i\sqrt{T_i - t}. \tag{42}$$

The constant σ_i is known as the **Black volatility** for caplet No. i. In order to make the dependence on the Black volatility σ_i explicit we will sometimes write the caplet price as $Capl_i^B(t; \sigma_i)$.

It is implicit in the Black formula that the forward rates are lognormal (under some probability measure), but until recently there was no firm theoretical base for the use of the Black-76 formula for caplets. One of the main goals of this chapter is precisely that of investigating whether it is possible to build an arbitrage free model object which produces formulas of the Black type for caplet prices.

In the market, cap prices are not quoted in monetary terms but instead in terms of **implied Black volatilities**, and these volatilities can furthermore be quoted as **flat volatilities** or as **spot volatilities** (also known as **forward volatilities**). They are defined as follows.

Let us consider a fixed date t, the fixed set of dates $T_0, T_1, \ldots, T_N$ where $t \leqslant T_0$, and a fixed cap rate R. We assume that, for each $i = 1, \ldots, N$, there is a traded cap with resettlement dates $T_0, T_1, \ldots, T_i$, and we denote the corresponding observed market price by $\mathbf{Cap_i^m}$. From this data we can easily compute the market prices for the corresponding caplets as

$$\mathbf{Capl_i^m}(t) = \mathbf{Cap_i^m}(t) - \mathbf{Cap_{i-1}^m}(t), \quad i = 1, \ldots, N \tag{43}$$

with the convention $\mathbf{Cap_0^m}(t) = 0$. Alternatively, given market data for caplets we can easily compute the corresponding market data for caps.

Definition 5.3. Given market price data as above, the implied Black volatilities are defined as follows.

- The implied **flat volatilities** $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$ are defined as the solutions of the equations

$$\mathbf{Cap_i^m}(t) = \sum_{k=1}^{i} \mathbf{Capl_k^B}(t; \bar{\sigma}_i), \quad i = 1, \ldots, N. \tag{44}$$

- The implied **forward** or **spot** volatilities $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$ are defined as solutions of the equations

$$\mathbf{Capl}_\mathbf{i}^\mathbf{m}(t) = \mathbf{Capl}_\mathbf{i}^\mathbf{B}(t; \bar{\sigma}_i), \quad i = 1, \ldots, N. \tag{45}$$

A sequence of implied volatilities $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$ (flat or spot) is called a volatility **term structure**. Note that we use the same notation $\bar{\sigma}_i$ for flat as well as for spot volatilities. In applications this will be made clear by the context.

Summarizing the formal definition above, the flat volatility $\bar{\sigma}_i$ is volatility implied by the Black formula if you use *the same* volatility for each caplet in the cap with maturity T_i. The spot volatility σ_i is just the implied volatility from caplet No. i. The difference between flat and forward volatilities is thus similar to the difference between yields and forward rates. A typical shape of the volatility term structure (flat or spot) for caps with, say, a three months tenor is that it has an upward hump for maturities around two–three years, but that the long end of the curve is downward sloping.

5.2 The LIBOR market model

We now turn from market practice to the construction of the so-called LIBOR market models. To motivate these models let us consider the theoretical arbitrage free pricing of caps. The price $c_i(t)$ of a caplet No. i of course one hand given by the standard risk neutral valuation formula

$$\mathbf{Capl}_\mathbf{i}(t) = \alpha_i E^Q\Big[e^{-\int_0^{T_i} r(s)\,\mathrm{d}s} \cdot \max\big[L_i(T_{i-1}) - R, 0\big] \mid \mathcal{F}_t\Big],$$
$$i = 1, \ldots, N,$$

but it is much more natural to use the T_i forward measure to obtain

$$\mathbf{Capl}_\mathbf{i}(t) = \alpha_i p_i(t) E^{T_i}\big[\max\big[L_i(T_{i-1}) - R, 0\big] \mid \mathcal{F}_t\big], \quad i = 1, \ldots, N, \tag{46}$$

where E^{T_i} denotes expectation under the Q^{T_i}. In order to have a more compact notation we will from now on denote Q^{T_i} by Q^i.

The focal point of the LIBOR models is the following simple result.

Lemma 5.1. *For every $i = 1, \ldots, N$, the LIBOR process L_i is a martingale under the corresponding forward measure Q^{T_i}, on the interval $[0, T_{i-1}]$.*

Proof. We have

$$\alpha_i \cdot L_i(t) = \frac{p_{i-1}(t)}{p_i(t)} - 1.$$

The process 1 is obviously a martingale under any measure. The process p_{i-1}/p_i is the price of the T_{i-1} bond normalized by the numeraire p_i. Since p_i is the numeraire for the martingale measure Q^{T_i}, the process p_{i-1}/p_i is thus

trivially a martingale on the interval $[0, T_{i-1}]$. Thus $\alpha_i L_i$ is a martingale and hence L_i is also a martingale. □

The basic idea is now to define the LIBOR rates such that, for each i, $L_i(T)$ will be lognormal under "its own" measure Q^i, since then all caplet prices in (46) will be given by a Black type formula. In order to do this we consider the following objects as given *a priori*.

- A set of resettlement dates $T_0, \ldots, T_N$.
- An arbitrage free market bonds with maturities $T_0, \ldots, T_N$.
- A k-dimensional Q^N-Wiener process W^N.
- For each $i = 1, \ldots, N$ a *deterministic* function of time $\sigma_i(t)$.
- An initial nonnegative forward rate term structure $L_1(0), \ldots, L_N(0)$.
- For each $i = 1, \ldots, N$, we define W^i as the k-dimensional Q^i-Wiener process generated by W^N under the Girsanov transformation $Q^N \to Q^i$.

Definition 5.4. If the LIBOR forward rates have the dynamics

$$\mathrm{d}L_i(t) = L_i(t)\sigma_i(t)\,\mathrm{d}W^i(t), \quad i = 1, \ldots, N, \tag{47}$$

where W^i is Q^i-Wiener as described above, then we say that we have a discrete tenor **LIBOR market model** with volatilities $\sigma_1, \ldots, \sigma_N$.

From the definition above it is not obvious that, given a specification of $\sigma_1, \ldots, \sigma_N$, there exists a corresponding LIBOR market model. In order to arrive at the basic pricing formulas a quickly as possible we will temporarily ignore the existence problem, but we will come back to it below and thus provide the missing link.

5.3 Pricing caps in the LIBOR model

Given a LIBOR market model, the pricing of a caplet, and hence also a cap, is trivial. Since L_i in (47) is just a GBM we obtain

$$L_i(T) = L_i(t) \cdot e^{\int_t^T \sigma_i(s)\,\mathrm{d}W^i(s) - \frac{1}{2}\int_t^T \|\sigma_i(s)\|^2\,\mathrm{d}s}.$$

Since σ_i is assumed to be deterministic this implies that, conditional on $\mathcal{F}_t$, $L_i(T)$ is lognormal, i.e. we can write

$$L_i(T) = L_i(t)e^{Y_i(t,T)},$$

where $Y_i(t, T)$ is normally distributed with expected value

$$m_i(t, T) = -\frac{1}{2}\int_t^T \big\|\sigma_i(s)\big\|^2\,\mathrm{d}s, \tag{48}$$

and variance

$$\Sigma_i^2(t,T) = \int_t^T \|\sigma_i(s)\|^2 \,\mathrm{d}s. \tag{49}$$

Using these results and (46), a simple calculation gives us the pricing formula for caps. Alternatively we see that the expectation E^i for the cap price in (46) is just the call price, within the Black–Scholes framework, in a world with zero short rate on an underlying traded asset with lognormal distribution as above.

Proposition 5.1. *In the LIBOR market model, the caplet prices are given by*

$$\mathbf{Capl_i}(t) = \alpha_i \cdot p_i(t)\big\{L_i(t)N[d_1] - RN[d_2]\big\}, \quad i = 1,\ldots,N, \tag{50}$$

where

$$d_1 = \frac{1}{\Sigma_i(t,T_{i-1})}\left[\ln\left(\frac{L_i(t)}{R}\right) + \frac{1}{2}\Sigma_i^2(t,T_{i-1})\right], \tag{51}$$

$$d_2 = d_1 - \Sigma_i(t,T_{i-1}), \tag{52}$$

with Σ_i defined by (49).

We thus see that each caplet price is given by a Black type formula.

Remark 5.1. Sometimes it is more convenient of working with a LIBOR model of the form

$$\mathrm{d}L_i(t) = L_i(t)\sigma_i(t)\,\mathrm{d}W^i(t), \quad i = 1,\ldots,N, \tag{53}$$

where $\sigma_i(t)$ is a *scalar* deterministic function, W^i is a *scalar* Q^i-Wiener process. Then the formulas above still hold if we replace $\|\sigma_i\|^2$ by σ_i^2. We can also allow for correlation between the various Wiener processes, but this will not affect the pricing of caps and floors. Such a correlation *will* however affect the pricing of more complicated products.

5.4 Terminal measure dynamics and existence

We now turn to the question whether there always exists a LIBOR market model for any given specification of the deterministic volatilities $\sigma_1,\ldots,\sigma_N$. In order to even get started we first have to specify all LIBOR rates $L_1,\ldots,L_N$ under **one** common measure, and the canonical choice is the **terminal measure** Q^N.

Our problem is then basically that of carrying out a two stage program.

- Specify all LIBOR rates under Q^N with dynamics of the form

$$\mathrm{d}L_i(t) = L_i(t)\mu_i\big(t,L(t)\big)\,\mathrm{d}t + L_i(t)\sigma_i(t)\,\mathrm{d}W^N(t), \quad i = 1,\ldots,N. \tag{54}$$

where $L(t) = [L_1(t), \ldots, L_N(t)]^\star$, and μ_i is some deterministic function.
- Show that, for some suitable choice of $\mu_1, \ldots, \mu_N$, the Q^N dynamics in (54) will imply Q^i dynamics of the form (47).

In order to carry out this program we need to see how W^N is transformed into W^i as we change measure from Q^N to Q^i. We do this inductively by studying the effect of the Girsanov transformation from Q^i to Q^{i-1}.

Remark 5.2. We have a small but irritating notational problem. LIBOR rates are typically denoted by the letter "L", but this is also a standard notation for a likelihood process. In order to avoid confusion we therefore introduce the notational convention that, *in this chapter only*, likelihood processes will be denoted by the letter η. In particular we introduce the notation

$$\eta_i^j(t) = \frac{\mathrm{d}Q^j}{\mathrm{d}Q^i}, \quad \text{on } \mathcal{F}_t \text{ for } i, j = 1, \ldots, N. \tag{55}$$

In order to get some idea of how we should choose the Q^N drifts of the LIBOR rates in (54) we will now perform some informal calculations. We thus (informally) assume that the LIBOR dynamics are of the form (54) under Q^N and that they are also of the form (47) under their own martingale measure. It is easily seen that the Radon–Nikodym derivative η_i^j is given by

$$\eta_i^j(t) = \frac{p_i(0)}{p_j(0)} \cdot \frac{p_j(t)}{p_i(t)}, \tag{56}$$

and in particular

$$\eta_i^{i-1}(t) = a_i \cdot \frac{p_{i-1}(t)}{p_i(t)} = a_i\big(1 + \alpha_i L_i(t)\big), \tag{57}$$

where $a_i = p_i(0)/p_{i-1}(0)$. From this formula we can now easily compute the η_i^{i-1} dynamics under Q^i as

$$\mathrm{d}\eta_i^{i-1}(t) = a_i\alpha_i\,\mathrm{d}L_i(t). \tag{58}$$

Assuming (still informally) that the L_i-dynamics are as in (47), and using (37) we then obtain

$$\mathrm{d}\eta_i^{i-1}(t) = a_i\alpha_i L_i(t)\sigma_i(t)\,\mathrm{d}W^i(t) \tag{59}$$

$$= a_i\alpha_i\frac{1}{\alpha_i}\left(\frac{p_{i-1}(t)}{p_i(t)} - 1\right)\sigma_i(t)\,\mathrm{d}W^i(t) \tag{60}$$

$$= \eta_i^{i-1}(t)a_i\alpha_i\frac{1}{\eta_i^{i-1}(t)}\left(\frac{p_{i-1}(t)}{p_i(t)} - 1\right)\sigma_i(t)\,\mathrm{d}W^i(t). \tag{61}$$

Using (57) we finally obtain

$$\mathrm{d}\eta_i^{i-1}(t) = \eta_i^{i-1}(t)\frac{\alpha_i L_i(t)}{1+\alpha_i L_i(t)}\sigma_i(t)\,\mathrm{d}W^i(t). \tag{62}$$

Thus the Girsanov kernel for η_i^{i-1} is given by

$$\frac{\alpha_i L_i(t)}{1+\alpha_i L_i(t)}\sigma_i^\star(t), \tag{63}$$

so from the Girsanov Theorem we have the relation

$$\mathrm{d}W^i(t) = \frac{\alpha_i L_i(t)}{1+\alpha_i L_i(t)}\sigma_i^\star(t)\,\mathrm{d}t + \mathrm{d}W^{i-1}(t). \tag{64}$$

Applying this inductively we obtain

$$\mathrm{d}W^i(t) = -\sum_{k=i+1}^{N}\frac{\alpha_k L_k(t)}{1+\alpha_k L_k(t)}\sigma_k^\star(t)\,\mathrm{d}t + \mathrm{d}W^N(t), \tag{65}$$

and plugging this into (47) we can finally obtain the Q^N dynamics of L_i (see (66) below).

All this was done under the informal assumption that there actually existed a LIBOR model satisfying both (47) and (54). We can however easily turn the argument around and we have the following existence result.

Proposition 5.2. *Consider a given volatility structure σ_1, σ_N, where each σ_i is assumed to be bounded, a probability measure Q^N and a standard Q^N-Wiener process W^N. Define the processes $L_1, \ldots, L_N$ by*

$$\begin{aligned}\mathrm{d}L_i(t) = &-L_i(t)\left(\sum_{k=i+1}^{N}\frac{\alpha_k L_k(t)}{1+\alpha_k L_k(t)}\sigma_k(t)\sigma_i^\star(t)\right)\mathrm{d}t\\ &+ L_i(t)\sigma_i(t)\,\mathrm{d}W^N(t),\end{aligned} \tag{66}$$

for $i = 1, \ldots, N$ where we use the convention $\sum_N^N(\ldots) = 0$. Then the Q^i-dynamics of L_i are given by (47). Thus there exists a LIBOR model with the given volatility structure.

Proof. Given that (66) has a solution for $i = 1, \ldots, N$, and that the Girsanov kernel in (63) satisfies the Novikov condition, the proof consists of exactly the calculations above. As for the existence of a solution of (66), this is trivial for $i = N$ since then the equation reads

$$\mathrm{d}L_N(t) = L_i(t)\sigma_N(t)\,\mathrm{d}W^N(t),$$

which is just GBM and since σ_N is bounded a solution does exist. Assume now that (66) admits a solution for $k = i+1, \ldots, N$. We can then write the ith

component of (66) as

$$dL_i(t) = L_i(t)\mu_i\big[t, L_{i+1}(t), \ldots, L_N(t)\big]\,dt + L_i(T)\sigma_i(t)\,dW^N(t),$$

where the point is that μ_i does only depend on L_k for $k = i+1, \ldots, N$ and not on L_i. Denoting the vector $(L_{i+1}, \ldots, L_N)$ by L^N_{i+1} we thus have the explicit solution

$$L_i(t) = L_i(0)\exp\left\{\int_0^t \left(\mu_i\big[s, L^N_{i+1}(s)\big] - \frac{1}{2}\|\sigma_i\|^2(s)\right)ds\right\}$$

$$\times \exp\left\{\int_0^t \mu_i\big[s, L^N_{i+1}(s)\big]\,dW^N(s)\right\},$$

thus proving existence by induction. It also follows by induction that, given an initial positive LIBOR term structure, all LIBOR rate processes will be positive. From this we see that the Girsanov kernel in (63) is also bounded and thus it satisfies the Novikov condition. □

5.5 *Calibration and simulation*

Suppose that we want to price some exotic (i.e. not a cap or a floor) interest rate derivative, like a Bermudan swaption, performing this with a LIBOR model means that we typically carry out the following two steps.

- Use implied Black volatilities in order to calibrate the model parameters to market data.
- Use Monte Carlo (or some other numerical method) to price the exotic instrument.

In this section we mainly discuss the calibration part, and only comment briefly on the numerical aspects. For numerics and simulation see the Notes.

Let us thus assume that, for the resettlement dates $T_0, \ldots, T_N$, we are given an empirical term structure of implied forward volatilities, $\bar{\sigma}_1, \ldots, \bar{\sigma}_N$, i.e. the implied Black volatilities for all caplets. For simplicity we assume that we are standing at time $t = 0$. Comparing the Black formula (40) with (50) we see that in order to calibrate the model we have to choose the deterministic LIBOR volatilities $\sigma_1(\cdot), \ldots, \sigma_N(\cdot)$, such that

$$\bar{\sigma}_i = \frac{1}{T_i}\int_0^{T_{i-1}} \big\|\sigma_i(s)\big\|^2\,ds, \quad i = 1, \ldots, N. \tag{67}$$

Alternatively, if we use a scalar Wiener process for each LIBOR rate we must choose the scalar function $\sigma_i(\cdot)$ such that

$$\bar{\sigma}_i = \frac{1}{T_i} \int_0^{T_{i-1}} \sigma_i^2(s)\, \mathrm{d}s, \quad i = 1, \ldots, N. \tag{68}$$

This is obviously a highly under determined system, so in applications it is common to make some structural assumption about the shape of the volatility functions. Below is a short and incomplete list of popular specifications. We use the formalism with a scalar Wiener process for each forward rate, and we recall that L_i lives on the time interval $0 \leqslant t \leqslant T_{i-1}$. We also introduce the temporary convention that $T_{-1} = 0$.

1. For each $i = 1, \ldots, N$, assume that the corresponding volatility is constant in time, i.e. that

   $$\sigma_i(t) = \sigma_i$$

   for $0 \leqslant t \leqslant T_{i-1}$.
2. For each $i = 1, \ldots, N$, assume that σ_i is piecewise constant, i.e. that

   $$\sigma_i(t) = \sigma_{ij}, \quad \text{for } T_{j-1} < t \leqslant T_j,\ j = 0, \ldots, i-1.$$

3. As in item 2, but with the requirement that the volatility only depends on the number of resettlement dates left to maturity, i.e. that

   $$\sigma_{ij} = \beta_{i-j}, \quad \text{for } T_{j-1} < t \leqslant T_j,\ j = 0, \ldots, i-1$$

   where $\beta_1, \ldots, \beta_N$ are fixed numbers.
4. As in item 2, with the further specification that

   $$\sigma_{ij} = \beta_i \gamma_j, \quad \text{for } T_{j-1} < t \leqslant T_j, j = 0, \ldots, i-1$$

   where β_i and γ_j are fixed numbers.
5. Assume some simple functional parameterized form of the volatilities such as for example

   $$\sigma_i(t) = q_i(T_{i-1} - t)e^{\beta_i(T_{i-1}-t)}$$

   where $q_i(\cdot)$ is some polynomial and β_i is a real number.

Assuming that the model has been calibrated to market data, Monte Carlo simulation is the standard tool for computing prices of exotics. Since the SDE (66) is to complicate to allow an analytical solution, we have to resort to simulation of discretized versions of the equations.

6 Notes

The basic papers on the LIBOR and swap market models are Miltersen et al. (1997), Brace et al. (1997), and Jamshidian (1997). Since these basic papers were published there has appeared a huge literature on the subject. Very readable accounts can be found in Hunt and Kennedy (2000), Protter (2000) and the almost encyclopedic Brigo and Mercurio (2001).

7 Geometric interest rate theory

The object of this section is to give an introduction to some recent works on the geometric aspects of interest rate theory.

7.1 Setup

We consider a given forward rate model under a risk neutral martingale measure Q. We will adopt the Musiela parameterization and use the notation

$$r(t, x) = f(t, t + x).$$

We recall the following result which is the HJM drift condition in the Musiela parameterization.

Proposition 7.1 *(The Forward Rate Equation). Under the martingale measure Q the r-dynamics are given by*

$$\mathrm{d}r(t, x) = \left\{ \frac{\partial}{\partial x} r(t, x) + \sigma(t, x) \int_0^x \sigma(t, u)^\star \, \mathrm{d}u \right\} \mathrm{d}t + \sigma(t, x) \, \mathrm{d}W(t), \tag{69}$$

$$r(0, x) = r^o(0, x). \tag{70}$$

where $\star$ denotes transpose.

7.2 Main problems

Suppose now that we are give 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

$$dZ_t = a(Z_t)\,dt + b(Z_t)\,dW_t, \tag{71}$$

$$r(t, x) = G(Z_t, x) \tag{72}$$

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 (69).

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.

8 Consistency and invariant manifolds

In this section we study when a given submanifold of forward rate curves is consistent (in the sense described above) with 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.

We 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 $\mathcal{M}$?

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 8.1 *(Invariant Manifold)*. Take as given the forward rate process dynamics (69). 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.

8.1 The formalized problem

8.1.1 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 8.2. 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

$$r : R_+ \to R$$

satisfying the norm condition $\|r\|_\gamma < \infty$. Here the norm is defined as

$$\|r\|_\gamma{}^2 = \int_0^\infty r^2(x)e^{-\gamma x}\,\mathrm{d}x + \int_0^\infty \left(\frac{\mathrm{d}r}{\mathrm{d}x}(x)\right)^2 e^{-\gamma x}\,\mathrm{d}x.$$

Remark 8.1. The variable x is as before interpreted as time to maturity. With the inner product

$$(r, q) = \int_0^\infty r(x)q(x)e^{-ax}\mathrm{d}x + \int_0^\infty \left(\frac{\mathrm{d}r}{\mathrm{d}x}(x)\right)\left(\frac{\mathrm{d}q}{\mathrm{d}x}(x)\right)e^{-\gamma x}\,\mathrm{d}x,$$

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 $\mathcal{H}_\gamma$.

8.1.2 The forward curve manifold

We consider as given a mapping

$$G : \mathcal{Z} \to \mathcal{H}, \tag{73}$$

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

$$G : \mathcal{Z} \times R_+ \to R. \tag{74}$$

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 8.3. The **forward curve manifold** $\mathcal{G} \subseteq \mathcal{H}$ is defined as

$$\mathcal{G} = Im(G).$$

8.1.3 The interest rate model

We take as given a volatility function σ of the form

$$\sigma : \mathcal{H} \times R_+ \to R^m$$

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.

$$\mathrm{d}r_t(x) = \left\{ \frac{\partial}{\partial x} r_t(x) + \sigma(r_t, x) \int_0^x \sigma(r_t, u)^\star \, \mathrm{d}u \right\} \mathrm{d}t + \sigma(r_t, x) \, \mathrm{d}W_t. \quad (75)$$

Remark 8.2. 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 Cristensen, 1999.

We need some regularity assumptions, and the main ones are as follows. See Björk and Cristensen (1999) for technical details.

Assumption 8.1. We assume the following.

- The volatility mapping $r \longmapsto \sigma(r)$ is smooth.
- The mapping $z \longmapsto G(z)$ is a smooth embedding, so in particular the Frechet 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 Eq. (75).

8.1.4 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 (75).
  - A mapping G, specifying a forward curve manifold $\mathcal{G}$.
- Is $\mathcal{G}$ then invariant under the action of r?

8.2 The invariance conditions

In order to study the invariance problem we need to introduce some compact notation.

Definition 8.4. We define $\mathbf{H}\sigma$ by

$$\mathbf{H}\sigma(r, x) = \int_0^x \sigma(r, s) \, \mathrm{d}s.$$

Suppressing the x-variable, the Itô dynamics for the forward rates are thus given by

$$dr_t = \left\{ \frac{\partial}{\partial x} r_t + \sigma(r_t) \mathbf{H} \sigma(r_t)^\star \right\} dt + \sigma(r_t)\, dW_t \tag{76}$$

and we write this more compactly as

$$dr_t = \mu_0(r_t)\, dt + \sigma(r_t)\, dW_t, \tag{77}$$

where the drift μ_0 is given by the bracket term in (76). To get some intuition we now formally "divide by dt" and obtain

$$\frac{dr}{dt} = \mu_0(r_t) + \sigma(r_t)\dot{W}_t, \tag{78}$$

where the formal time derivative $\dot{W}_t$ is interpreted as an "input signal" chosen by chance. We are thus led to study the associated deterministic control system

$$\frac{dr}{dt} = \mu_0(r_t) + \sigma(r_t)u_t. \tag{79}$$

The intuitive idea is now that $\mathcal{G}$ is invariant under (77) if and only if $\mathcal{G}$ is invariant under (79) 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 $Im[G'_z(z)]$, where G'_z denotes the Frechet derivative (Jacobian), we are led to conjecture that $\mathcal{G}$ is invariant if and only if the condition

$$\mu_0(r) + \sigma(r)u \in Im\big[G'_z(z)\big]$$

is satisfied for all $u \in R^m$. This can also be written

$$\mu_0(r) \in Im\big[G'_z(z)\big],$$
$$\sigma(r) \in Im\big[G'_z(z)\big],$$

where the last inclusion is interpreted component wise 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 (79), and Itô calculus, which governs (77). In order to bridge this gap we have to rewrite the analysis in terms of Stratonovich integrals instead of Itô integrals.

Definition 8.5. 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

$$\int_0^t X_s \circ dY_s = \int_0^t X_s\, dY_s + \frac{1}{2}\langle X, Y\rangle_t. \tag{80}$$

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 (80) by

$$\mathrm{d}\langle X, Y\rangle_t = \mathrm{d}X_t\,\mathrm{d}Y_t, \tag{81}$$

with the usual 'multiplication rules' $\mathrm{d}W \cdot \mathrm{d}t = \mathrm{d}t \cdot \mathrm{d}t = 0$, $\mathrm{d}W \cdot \mathrm{d}W = \mathrm{d}t$. We now recall the main result and *raison d'être* for the Stratonovich integral.

Proposition 8.1 *(Chain Rule). Assume that the function $F(t, y)$ is smooth. Then we have*

$$\mathrm{d}F(t, Y_t) = \frac{\partial F}{\partial t}(t, Y_t)\,\mathrm{d}t + \frac{\partial F}{\partial y} \circ \mathrm{d}Y_t. \tag{82}$$

Thus, in the Stratonovich calculus, the Itô formula takes the form of the standard chain rule of ordinary calculus.

Returning to (77), the Stratonovich dynamics are given by

$$\begin{aligned}\mathrm{d}r_t = &\left\{\frac{\partial}{\partial x}r_t + \sigma(r_t)\mathbf{H}\sigma(r_t)^\star\right\}\mathrm{d}t - \frac{1}{2}\mathrm{d}\big\langle\sigma(r_t), W_t\big\rangle \\ &+ \sigma(r_t) \circ \mathrm{d}W_t.\end{aligned} \tag{83}$$

In order to compute the Stratonovich correction term above we use the infinite dimensional Itô formula (see Da Prato and Zabczyk, 1992) to obtain

$$\mathrm{d}\sigma(r_t) = \{\ldots\}\,\mathrm{d}t + \sigma_r'(r_t)\sigma(r_t)\,\mathrm{d}W_t, \tag{84}$$

where σ_r' denotes the Frechet derivative of σ w.r.t. the infinite dimensional r-variable. From this we immediately obtain

$$\mathrm{d}\big\langle\sigma(r_t), W_t\big\rangle = \sigma_r'(r_t)\sigma(r_t)\,\mathrm{d}t. \tag{85}$$

Remark 8.3. If the Wiener process W is multidimensional, then σ is a vector $\sigma = [\sigma_1, \ldots, \sigma_m]$, and the RHS of (85) 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 (83) becomes

$$\begin{aligned}\mathrm{d}r_t = &\left\{\frac{\partial}{\partial x}r_t + \sigma(r_t)\mathbf{H}\sigma(r_t)^\star - \frac{1}{2}\sigma_r'(r_t)\sigma(r_t)\right\}\mathrm{d}t \\ &+ \sigma(r_t) \circ \mathrm{d}W_t.\end{aligned} \tag{86}$$

We now write (86) as

$$\mathrm{d}r_t = \mu(r_t)\,\mathrm{d}t + \sigma(r_t) \circ \mathrm{d}W_t, \tag{87}$$

where

$$\mu(r,x) = \frac{\partial}{\partial x} r(x) + \sigma(r_t, x) \int_0^x \sigma(r_t, u)^\star \, \mathrm{d}u - \frac{1}{2}\big[\sigma_r'(r_t)\sigma(r_t)\big](x). \quad (88)$$

Given the heuristics above, our main result is not surprising. The formal proof, which is somewhat technical, is left out. See Björk and Cristensen (1999).

Theorem 8.1 *(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*

$$G_x'(z) + \sigma(r)\mathbf{H}\sigma(r)^\star - \frac{1}{2}\sigma_r'(r)\sigma(r) \in Im\big[G_z'(z)\big], \quad (89)$$

$$\sigma(r) \in Im\big[G_z'(z)\big] \quad (90)$$

hold for all $z \in \mathcal{Z}$ with $r = G(z)$.

Here, G_z' and G_x' denote the Frechet derivative of G with respect to z and x, respectively. The condition (90) is interpreted component wise for σ. Condition (89) is called *the consistent drift condition*, and (90) is called *the consistent volatility condition*.

Remark 8.4. It is easily seen that if the family G is invariant under shifts in the x-variable, then we will automatically have the relation

$$G_x'(z) \in Im\big[G_z'(z)\big],$$

so in this case the relation (89) can be replaced by

$$\sigma(r)\mathbf{H}\sigma(r)^\star - \frac{1}{2}\sigma_r'(r)\sigma(r) \in Im\big[G_z'(z)\big],$$

with $r = G(z)$ as usual.

8.3 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.

8.3.1 The Nelson–Siegel family

The Nelson–Siegel (henceforth NS) forward curve manifold $\mathcal{G}$ is parameterized by $z \in R^4$, the curve $x \longmapsto G(z,x)$ as

$$G(z,x) = z_1 + z_2 e^{-z_4 x} + z_3 x e^{-z_4 x}. \quad (91)$$

For $z_4 \neq 0$, the Frechet derivatives are easily obtained as

$$G'_z(z,x) = \left[1,\ e^{-z_4 x},\ xe^{-z_4 x},\ -(z_2 + z_3 x)xe^{-z_4 x}\right], \tag{92}$$

$$G'_x(z,x) = (z_3 - z_2 z_4 - z_3 z_4 x)e^{-z_4 x}. \tag{93}$$

In the degenerate case $z_4 = 0$, we have

$$G(z,x) = z_1 + z_2 + z_3 x, \tag{94}$$

We return to this case below.

8.3.2 *The Hull–White and Ho–Lee models*

As our test case, we analyze the Hull and White (henceforth HW) extension of the Vasiček model. On short rate form the model is given by

$$\mathrm{d}R(t) = \left\{\Phi(t) - aR(t)\right\}\mathrm{d}t + \sigma\,\mathrm{d}W(t), \tag{95}$$

where $a, \sigma > 0$. As is well known, the corresponding forward rate formulation is

$$\mathrm{d}r(t,x) = \beta(t,x)\,\mathrm{d}t + \sigma e^{-ax}\,\mathrm{d}W_t. \tag{96}$$

Thus, the volatility function is given by $\sigma(x) = \sigma e^{-ax}$, and the conditions of Theorem 8.1 become

$$G'_x(z,x) + \frac{\sigma^2}{a}\left[e^{-ax} - e^{-2ax}\right] \in Im\left[G'_z(z,x)\right], \tag{97}$$

$$\sigma e^{-ax} \in Im\left[G'_z(z,x)\right]. \tag{98}$$

To investigate whether the NS manifold is invariant under HW dynamics, we start with (98) and fix a z-vector. We then look for constants (possibly depending on z) A, B, C, and D, such that for all $x \geqslant 0$ we have

$$\sigma e^{-ax} = A + Be^{-z_4 x} + Cxe^{-z_4 x} - D(z_2 + z_3 x)xe^{-z_4 x}. \tag{99}$$

This is possible if and only if $z_4 = a$, and since (98) 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 8.2 *(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 Cristensen, 1999.

Remark 8.5. It is an easy exercise to see that the minimal manifold which is consistent with HW is given by

$$G(z,x) = z_1 e^{-ax} + z_2 e^{-2ax}.$$

8.4 The Filipović state space approach to consistency

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. In a very general setting, inverse consistency problems like this has been studied in great detail by Filipović (Filipović, 1999, 2000a, 2000b). In this section we will give an introduction to the Filipović state space approach to the (inverse) consistency problem, and we will also study a small laboratory example.

The study will be done within the framework of a factor model.

Definition 8.6. A **factor model** for the forward rate process r consists of the following objects and relations.

- A d-dimensional **factor** or **state** process Z with Q-dynamics of the form
$$\mathrm{d}Z_t = a(Z_t)\,\mathrm{d}t + b(Z_t)\,\mathrm{d}W_t, \tag{100}$$
where W is an m-dimensional Wiener process. We denote by a_i the ith component of the column vector a, and by b_i the ith row of the matrix b.
- A smooth **output mapping**
$$G : R^d \to \mathcal{H}.$$
For each $z \in R^d$, $G(z)$ is thus a real valued C^∞ function and it's value at the point $x \in R$ is denoted by $G(z, x)$.
- The forward rate process is then defined by
$$r_t = G(Z_t), \tag{101}$$
or on component form
$$r_t(x) = G(Z_t, x). \tag{102}$$

Since we have given the Z dynamics under the *martingale measure* Q, it is obvious that there has to be some consistency requirements on the relations between a, b and G in order for r in (101) to be a specification of the forward rate process under Q. The obvious way of deriving the consistency requirements is to compute the r dynamics from (100)–(101) and then to compare the result with the general form of the forward rate equation in (69). For ease of notation we will use the shorthand notation

$$G_x = \frac{\partial G}{\partial x}, \qquad G_i = \frac{\partial G}{\partial z_i}, \qquad G_i = \frac{\partial^2 G}{\partial z_i \partial z_j}. \tag{103}$$

From the Itô formula, (100), and (101) we obtain

$$\mathrm{d}r_t = \left\{ \sum_{i=1}^{d} G_i(Z_t) a_i(Z_t)\,\mathrm{d}t + \frac{1}{2} \sum_{i,j=1}^{d} G_{ij}(Z_t) b_i(Z_t) b_j^\star(Z_t) \right\} \mathrm{d}t \tag{104}$$

$$+\sum_{i=1}^{d} G_i(Z_t)b_i(Z_t)\,\mathrm{d}W_t \tag{105}$$

where * denotes transpose. Going back to the forward rate equation (69), we can identify the volatility process as

$$\sigma_t = \sum_{i=1}^{d} G_i(Z_t)b_i(Z_t).$$

We now insert this into the drift part of (69). We then use (101) to deduce that $\mathbf{F}r_t = G_x(Z_t)$ and also insert this expression into the drift part of (69). Comparing the resulting equation with (104) gives us the required consistency conditions.

Proposition 8.3 *(Filipović). Under a martingale measure Q, the following relation must hold identically in* (z, x).

$$G_x(z,x) + \sum_{i,j=1}^{d} b_i(z)b_j^\star(z)G_i(z,x)\int_0^x G_j(z,s)\,\mathrm{d}s$$
$$= \sum_{i=1}^{d} G_i(z,x)a_i(z) + \frac{1}{2}\sum_{i,j=1}^{d} G_{ij}(z,x)b_i(z)b_j^\star(z). \tag{106}$$

We can view the consistency equation (106) in three different ways.

- We can check consistency for a given specification of G, a, b.
- We can specify a and b. Then (106) is a PDE for the determination of a consistent output function G.
- We can specify G, i.e. we can specify a finite dimensional manifold of forward rate curves, and then use (106) to investigate whether there exist an underlying consistent state vector process Z, and if so, to find a and b.

We will focus on the last inverse problem above, and to see how the consistency equation can be used, we now go on to study to two simple laboratory examples.

Example 8.1. In this example we consider the 2-dimensional manifold of **linear** forward rate curves, i.e. the output function G defined by

$$G(z,x) = z_1 + z_2 x. \tag{107}$$

This is not a very natural example from a finance point of view, but it is a good illustration of technique. The question we ask is whether there exist some forward rate model consistent with the class of linear forward rate curves and

if so what the factor dynamics look like. For simplicity we restrict ourselves to the case of a scalar driving Wiener process, but the reader is invited to analyze the (perhaps more natural) case with a two-dimensional W.

We thus model the factor dynamics as

$$\mathrm{d}Z_{1,t} = a_1(Z_t)\,\mathrm{d}t + b_1(Z_t)\,\mathrm{d}W_t, \tag{108}$$

$$\mathrm{d}Z_{2,t} = a_1(Z_t)\,\mathrm{d}t + b_2(Z_t)\,\mathrm{d}W_t. \tag{109}$$

In this case we have

$$G_x(z,x) = z_2, \qquad G_1(z,x) = 1, \qquad G_2(z,x) = x,$$
$$G_{11}(z,x) = 0, \qquad G_{12}(z,x) = 0, \qquad G_{22}(z,x) = 0,$$

and

$$\int_0^x G_1(z,s)\,\mathrm{d}s = x, \qquad \int_0^x G_2(z,s)\,\mathrm{d}s = \frac{1}{2}x^2,$$

so the consistency equation (106) becomes

$$z_2 + b_1^2(z)x + b_1(z)b_2(z)\frac{1}{2}x^2 + b_2(z)b_1(z)x^2 + b_2^2(z)\frac{1}{2}x^3 = a_1(z) + a_2(z)x. \tag{110}$$

Identifying coefficients we see directly that $b_2 = 0$ so the equation reduces to

$$z_2 + b_1^2(z)x = a_1(z) + a_2(z)x \tag{111}$$

which gives us the relations $a_1 = z_2$ and $a_2 = b_1^2$. Thus we see that for this choice of G there does indeed exist a class of consistent factor models, with factor dynamics given by

$$\mathrm{d}Z_{1,t} = Z_{2,t}\,\mathrm{d}t + b_1(Z_t)\,\mathrm{d}W_t, \tag{112}$$

$$\mathrm{d}Z_{2,t} = b_1^2(Z_t)\,\mathrm{d}t. \tag{113}$$

Here b_1 can be chosen completely freely (subject only to regularity conditions). Choosing $b_1(z) = 1$, we see that the factor Z_2 is essentially running time, and the model is then in fact a special case of the Ho–Lee model.

8.5 Notes

The section is largely based on Björk and Cristensen (1999) and Filipović (1999). 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 been studied by Filipović in great depth (Filipović, 2001, 2000b). 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 (2001).

9 Existence of nonlinear realizations

We now turn to Problem 2 in Section 7.2, i.e. the problem 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. 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 (2001).

We now take as given a volatility $\sigma : \mathcal{H} \to \mathcal{H}$ and consider the induced forward rate model (on Stratonovich form)

$$\mathrm{d}r_t = \mu(r_t)\,\mathrm{d}t + \sigma(r_t) \circ \mathrm{d}W_t \tag{114}$$

where as before (see Section 8.2),

$$\mu(r) = \frac{\partial}{\partial x} r + \sigma(r)\mathbf{H}\sigma(r)^{\star} - \frac{1}{2}\sigma_r'(r)\sigma(r). \tag{115}$$

Remark 9.1. The reason for our choice of $\mathcal{H}$ as the underlying space, is that the linear operator $\mathbf{F} = \partial/\partial x$ is bounded in this space. Together with the assumptions above, this implies that both μ and σ are smooth vector fields on $\mathcal{H}$, thus ensuring the existence of a strong local solution to the forward rate equation for every initial point $r^o \in \mathcal{H}$.

9.1 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 $G : R^d \to \mathcal{H}$ such that r, locally in time, has the representation

$$\mathrm{d}Z_t = a(Z_t)\,\mathrm{d}t + b(Z_t)\,\mathrm{d}W_t, \quad Z_0 = z_0, \tag{116}$$

$$r(t, x) = G(Z_t, x). \tag{117}$$

Remark 9.2. 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 9.1. *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 Cristensen (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 (116)–(117), then every forward rate curve that will be produced by the model is of the form $x \longmapsto G(z, x)$ for some choice of z. Thus there exists a finite dimensional invariant submanifold $\mathcal{G}$ containing the initial forward rate curve r^o, namely $\mathcal{G} = Im\, G$. □

Using Theorem 8.1 we immediately obtain the following geometric characterization of the existence of a finite realization.

Corollary 9.1. *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.

9.2 The main result

Given the volatility vector field σ, and hence also the field μ, we now are faced with the problem of determining if 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 $\mathcal{H}$ (or any other Hilbert space), and a smooth vector field f on the space. For each fixed point $r^o \in \mathcal{H}$ we now ask if 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

$$\frac{dr_t}{dt} = f(r_t), \tag{118}$$

$$r_0 = r^o. \tag{119}$$

If r_t is the solution, at time t, of this ODE, we use the notation

$$r_t = e^{ft} r^o.$$

We have thus defined a group of operators $\{e^{ft}\colon\ t \in R\}$, and we note that the set $\{e^{ft} r^o\colon\ t \in R\} \subseteq \mathcal{H}$ 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 $r^o \in \mathcal{H}$ 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 an **tangential manifold** for the vector fields. At a first glance it would seem that there always exists an 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 $\{e^{f_1 s} r^o\colon\ s \geqslant 0\}$. 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 $\mathcal{H}$. This is our obvious candidate for an 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 9.1. Given smooth vector fields f and g on $\mathcal{H}$, the Lie bracket $[f, g]$ is a new vector field defined by

$$[f, g](r) = f'(r)g(r) - g'(r)f(r). \tag{120}$$

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 9.2. Let $f_1, \ldots, f_n$ 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

$$[f_i, f_j](x) \in \operatorname{span}\{f_1(x), \ldots, f_n(x)\}, \quad \forall i, j,$$

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 9.1 *(Frobenius). Let $f_1, \ldots, f_k$ and be independent smooth vector fields in $\mathcal{H}$ and consider a fixed point $r^o \in \mathcal{H}$. 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 $f_1, \ldots, f_k$ of vector fields is (locally) involutive.*

Proof. See Björk and Svensson (2001), 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 9.3. Take the vector fields $f_1, \ldots, f_k$ as given. The **Lie algebra** generated by $f_1, \ldots, f_k$ is the smallest linear space (over R) of vector fields which contains $f_1, \ldots, f_k$ and is closed under the Lie bracket. This Lie algebra is denoted by

$$\mathcal{L} = \{f_1, \ldots, f_k\}_{LA}.$$

The **dimension** of $\mathcal{L}$ is defined, for each point $r \in \mathcal{H}$ as

$$\dim\big[\mathcal{L}(r)\big] = \dim \operatorname{span}\big\{f_1(r), \ldots, f_k(r)\big\}.$$

Putting all these results together, we have the following main result on finite dimensional realizations.

Theorem 9.2 *(Main Result). Take the volatility mapping $\sigma = (\sigma_1, \ldots, \sigma_m)$ as given. Then the forward rate model generated by σ generically admits a finite dimensional realization if and only if*

$$\dim\{\mu, \sigma_1, \ldots, \sigma_m\}_{LA} < \infty$$

in a neighborhood of r^o.

When computing the Lie algebra generated by μ and σ, the following observations are often useful.

Lemma 9.1. *Take the vector fields $f_1, \dots, f_k$ as given. The Lie algebra $\mathcal{L} = \{f_1, \dots, f_k\}_{LA}$ 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 $[f, f] = 0$. □

9.3 Applications

In this section we give some simple applications of the theory developed above, but first we need to recall some facts about quasi-exponential functions.

Definition 9.4. A **quasi-exponential** (or QE) function is by definition any function of the form

$$f(x) = \sum_i e^{\lambda_i x} + \sum_j e^{\alpha_i x}\big[p_j(x)\cos(\omega_j x) + q_j(x)\sin(\omega_j x)\big], \quad (121)$$

where λ_i, α_1, ω_j are real numbers, whereas p_j and q_j are real polynomials.

QE functions will turn up over and over again, so we list some simple well known properties.

Lemma 9.2. *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 $f(x) = ce^{Ax}b$, where c is a row vector, A is a square matrix and b is a column vector.*
- *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.*

9.3.1 Constant volatility

We start with the simplest case, which is when the volatility $\sigma(r, x)$ is a constant vector in $\mathcal{H}$, 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

$$\mu(r, x) = \mathbf{F}r(x) + \sigma(x) \int_0^x \sigma(s)\, \mathrm{d}s,$$

$$\sigma(r, x) = \sigma(x).$$

where as before $\mathbf{F} = \frac{\partial}{\partial x}$.

The Frechet 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

$$\mu_r' = \mathbf{F},$$
$$\sigma_r' = 0.$$

Thus the Lie bracket $[\mu, \sigma]$ is given by

$$[\mu, \sigma] = \mathbf{F}\sigma,$$

and in the same way we have

$$\big[\mu, [\mu, \sigma]\big] = \mathbf{F}^2\sigma.$$

Continuing in the same manner it is easily seen that the relevant Lie algebra $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L} = \{\mu, \sigma\}_{LA} &= \operatorname{span}\big\{\mu, \sigma, \mathbf{F}\sigma, \mathbf{F}^2\sigma, \ldots\big\} \\ &= \operatorname{span}\big\{\mu, \mathbf{F}^n\sigma\,;\, n = 0, 1, 2, \ldots\big\}.\end{aligned}$$

It is thus clear that $\mathcal{L}$ is finite dimensional (at each point r) if and only if the function space

$$\operatorname{span}\big\{\mathbf{F}^n\sigma\,;\, n = 0, 1, 2, \ldots\big\}$$

is finite dimensional. We have thus obtained tho following result.

Proposition 9.2. *Under the above assumptions, there exists a finite dimensional realization if and only if σ is a quasi-exponential function.*

9.3.2 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

$$\sigma(r, x) = \varphi(r)\lambda(x). \tag{122}$$

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 9.1. 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

$$\mu(r) = \mathbf{F}r + \varphi^2(r)D - \frac{1}{2}\varphi'(r)[\lambda]\varphi(r)\lambda, \tag{123}$$

where $\varphi'(r)[\lambda]$ denotes the Frechet derivative $\varphi'(r)$ acting on the vector λ, and where the constant vector $D \in \mathcal{H}$ is given by

$$D(x) = \lambda(x) \int_0^x \lambda(s)\,\mathrm{d}s.$$

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

$$\mu(r) = \mathbf{F}r + \varphi^2(r)D - \frac{1}{2}\varphi'(r)[\lambda]\varphi(r)\lambda,$$
$$\sigma(r) = \varphi(r)\lambda,$$

is finite dimensional. Under Assumption 9.1 we can use Lemma 9.1, to see that the Lie algebra is in fact generated by the simpler system of vector fields

$$f_0(r) = \mathbf{F}r + \Phi(r)D,$$
$$f_1(r) = \lambda,$$

where we have used the notation

$$\Phi(r) = \varphi^2(r).$$

Since the field f_1 is constant, it has zero Frechet 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

$$\big[[f_0, f_1], f_1\big] = \Phi''(r)[\lambda; \lambda]D.$$

Note that $\Phi''(r)[\lambda; \lambda]$ is the second order Frechet derivative of Φ operating on the vector pair $[\lambda; \lambda]$. This pair is to be distinguished from (notice the semicolon) the Lie bracket $[\lambda, \lambda]$ (with a comma), which if course would be equal to zero. We now make a further assumption.

Assumption 9.2. We assume that $\Phi''(r)[\lambda; \lambda] \neq 0$ for all $r \in \mathcal{H}$.

Given this assumption we may again use Lemma 9.1 to see that the Lie algebra is generated by the following vector fields

$$f_0(r) = \mathbf{F}r,$$
$$f_1(r) = \lambda,$$
$$f_3(r) = \mathbf{F}\lambda,$$
$$f_4(r) = D.$$

Of these vector fields, all but f_0 are constant, so all brackets are easy. After elementary calculations we see that in fact

$$\{\mu, \sigma\}_{LA} = \text{span}\{\mathbf{F}r, \mathbf{F}^n\lambda,\ \mathbf{F}^n D;\ n = 0, 1, \ldots\}.$$

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 $\{\mathbf{F}^n\lambda;\ n \geqslant 0\}$ is finite dimensional. This occurs if and only if λ is quasi-exponential. If, on the other hand, λ is quasi-exponential, then we know from Lemma 9.2, that also D is quasi-exponential, since it is the integral of the QE function λ multiplied by the QE function λ. Thus the space $\{\mathbf{F}^n D;\ n = 0, 1, \ldots\}$ is also finite dimensional, and we have proved the following result.

Proposition 9.3. *Under Assumptions 9.1 and 9.2, 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.*

9.4 Notes

The section is largely based on Björk and Svensson (2001) 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 fall 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 (2001), occurs when σ is a finite point evaluation, i.e. when $\sigma(t, r) = h(t, r(x_1), \ldots, r(x_k))$ for fixed benchmark maturities $x_1, \ldots, x_k$. In Chiarella and Kwon (2001) it is studied when the corresponding finite set of benchmark forward rates is Markovian.

The Lie theory can also be used to determine when a HJM model leads to a Markovian short rate. A classic paper on Markovian short rates is Carverhill (1994), where a deterministic volatility of the form $\sigma(t, x)$ is considered. The first to state and prove a general result was Jeffrey (1995). See Eberlein and Raible (1999) for an example with a driving Levy process.

The problem of how to construct a concrete realization, given knowledge of the structure of the (finite dimensional) Lie algebra has been studied in Björk and Landén (2002). Stochastic volatility models are treated in Björk et al. (2002).

The functional analytical framework above has been extended considerably by Filipovć and Teichmann (Filipović and Teichmann, 2002, 2003, 2004). In particular, Filipović and Teichmann prove the remarkable result that any forward rate model admitting an FDR must necessarily have an affine term structure.

The geometric ideas presented above and in Björk and Svensson (2001) 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 Brocket (1981) for an overview of these areas.

10 Potentials and positive interest

In the previous sections, all modeling has been done under an equivalent martingale measure Q. It is of course also possible to model the relevant stochastic processes under the objective probability measure P, provided one can link it to the (not necessarily unique) martingale measure Q. This way of modeling is in fact what is done routinely in theoretical and empirical asset pricing theory, where one uses a *stochastic discount factor* (or SDF for short) instead of a martingale measure, and the purpose of this section is to present two approaches to interest rate theory based on stochastic discount factor, and relating bond pricing to stochastic potential theory.

Another appealing aspects of the approaches described below is that they both generate **positive term structures**, i.e. a system of bond prices for which all induced forward rates are positive. It is easily seen that positivity is equivalent to a positive short rate, and also equivalent to the conditions $0 \leqslant p(t,T) \leqslant 1$ and $p_T(t,T) < 0$ where p_T denotes the partial derivative of the bond price w.r.t. maturity.

At the end of the section we will also present a new approach to positive interest rate modeling, based on a study of the "term structure density." Although this approach is not based directly on potential we include it in the present section because of its connection to positive interest rates.

10.1 Generalities

As a general setup we consider a standard filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$ where P is the objective measure. We now need an assumption about how he market prices various assets.

Assumption 10.1. We assume that the market prices all assets, underlying and derivative, using a fixed martingale measure Q (with the money account as the numeraire).

We now recall that for a T-claim Y the arbitrage free price process is given by

$$\Pi(t;X) = E^Q\Big[e^{-\int_t^T r_s\,\mathrm{d}s}\cdot Y \mid \mathcal{F}_t\Big],$$

and, in particular, we have the price at $t=0$ as

$$\Pi(0;X) = E^Q\Big[e^{-\int_0^T r_s\,\mathrm{d}s}\cdot Y\Big]. \tag{124}$$

We denote the likelihood process for the transition from the objective measure P to the martingale measure Q by L, i.e.

$$L_t = \frac{\mathrm{d}Q_t}{\mathrm{d}P_t}, \tag{125}$$

where subindex t denotes the restriction of P and Q to $\mathcal{F}_t$. We may of course also write the price in (124) as an expected value under P:

$$E^P\Big[e^{-\int_0^T r_s\,\mathrm{d}s}\cdot L_T\cdot Y\Big] = E^P[Z_T\cdot Y]. \tag{126}$$

This leads us to the following definition.

Definition 10.1. The **state price density process**, or **stochastic discount factor** Z is defined by

$$Z(t) = e^{-\int_0^t r_s\,\mathrm{d}s}\cdot L_t. \tag{127}$$

We now have the following basic pricing result.

Proposition 10.1. *For any T-claim Y, the arbitrage free price process is given by*

$$\Pi(t;X) = \frac{E^P[Z_T Y \mid \mathcal{F}_t]}{Z_t}. \tag{128}$$

In particular, bond prices are given by

$$p(t,T) = \frac{E^P[Z_T \mid \mathcal{F}_t]}{Z_t}. \tag{129}$$

Proof. From the Bayes formula we obtain

$$\Pi(t;X) = E^Q\Big[e^{-\int_t^T r_s\,\mathrm{d}s}Y \mid \mathcal{F}_t\Big] = \frac{E^P[e^{-\int_t^T r_s\,\mathrm{d}s}L_T Y \mid \mathcal{F}_t]}{E^Q[L_T \mid \mathcal{F}_t]}$$

$$= \frac{E^P[e^{-\int_t^T r_s\,\mathrm{d}s} L_T Y \mid \mathcal{F}_t]}{L_t} = \frac{E^P[Z_T Y \mid \mathcal{F}_t]}{Z_t}. \qquad \square$$

We now have the following fact which we will use extensively.

Proposition 10.2. *Assume that the short rate is strictly positive and that the economically natural condition* $p(0,T) \to 0$ *as* $T \to \infty$ *is satisfied. Then the stochastic discount factor* Z *is a probabilistic* ***potential****, i.e.*

- Z *is a supermartingale.*
- $E[Z_t] \to 0$ *as* $t \to \infty$.

Conversely one can show that any potential will serve as a stochastic discount factor. The moral is thus that modeling bond prices in a market with positive interest rates is equivalent to modeling a potential, and in the next sections we will describe two ways of doing this.

We end by noticing that we can easily recover the short rate from the dynamics of Z.

Proposition 10.3. *If the dynamics of* Z *are written as*

$$\mathrm{d}Z_t = -h_t\,\mathrm{d}t + \mathrm{d}M_t \tag{130}$$

where h *is nonnegative and* M *is a martingale, then the short rate is given by*

$$r_t = Z_t^{-1} h_t. \tag{131}$$

Proof. Applying the Itô formula to the definition of Z we obtain

$$\mathrm{d}Z_t = -r_t Z_t\,\mathrm{d}t + e^{-\int_0^t r_s\,\mathrm{d}s}\,\mathrm{d}L_t. \tag{132}$$

$\square$

10.2 The Flesaker–Hughston fractional model

Given a stochastic discount factor Z and a positive short rate we may, for each fixed T, define the process $\{X(t,T);\ 0 \leqslant t \leqslant T\}$ by

$$X(t,T) = E^P[Z_T \mid \mathcal{F}_t], \tag{133}$$

and thus, according to (129) write bond prices as

$$p(t,T) = \frac{X(t,T)}{X(t,t)}. \tag{134}$$

We now have the following result.

Proposition 10.4. *For each fixed t, the mapping $T \longmapsto X(t,T)$ is smooth, and in fact*

$$\frac{\partial}{\partial T}X(t,T) = -E^P[r_T Z_T \mid \mathcal{F}_t]. \tag{135}$$

Furthermore, for each fixed T, the process

$$X_T(t,T) = \frac{\partial}{\partial T}X(t,T)$$

is a negative P-martingale satisfying

$$X_T(0,T) = -p_T(0,T), \quad \text{for all } T \geqslant 0. \tag{136}$$

Proof. Using the definition of Z and the Itô formula, we obtain

$$\mathrm{d}Z_s = -r_s Z_s\,\mathrm{d}s + B_s^{-1}\,\mathrm{d}L_s,$$

so

$$Z_T = Z_t - \int_t^T r_s Z_s\,\mathrm{d}s + \int_t^T B_s^{-1}\,\mathrm{d}L_s.$$

Since L is a martingale, this gives us

$$E^P[Z_T \mid \mathcal{F}_t] = -E^P\left[\int_t^T r_s Z_s\,\mathrm{d}s \mid \mathcal{F}_t\right],$$

and (135) follows immediately. The martingale property now follows directly from (135). □

We can now state the basic result from Flesaker–Hughston.

Theorem 10.1. *Assume that the term structure is positive. Then there exists a family of positive martingales $M(t,T)$ indexed by T and a positive deterministic function Φ such that*

$$p(t,T) = \frac{\int_T^\infty \Phi(s)M(t,s)\,\mathrm{d}s}{\int_t^\infty \Phi(s)M(t,s)\,\mathrm{d}s}. \tag{137}$$

The M family can, up to multiplicative scaling by the Φ process, be chosen as

$$M(t,T) = -X_T(t,T) = E^P[r_T Z_T \mid \mathcal{F}_t]. \tag{138}$$

In particular, Φ can be chosen as

$$\Phi(s) = -p_T(0,s), \tag{139}$$

in which case the corresponding M is normalized to $M(0,s) = 1$ for all $s \geqslant 0$.

Proof. A positive short rate implies $X(t,T) \to 0$ as $T \to \infty$, so we have

$$X(t,T) = -\int_T^\infty X_T(t,s)\,\mathrm{d}s,$$

and thus we obtain from (134)

$$p(t,T) = \frac{\int_T^\infty X_T(t,s)\,\mathrm{d}s}{\int_t^\infty X_T(t,s)\,\mathrm{d}s}. \tag{140}$$

If we now define $M(t,T)$ by

$$M(t,T) = -X_T(t,T), \tag{141}$$

then (137) follows from (140) with $\Phi \equiv 1$. The function Φ is only a scale factor which can be chosen arbitrarily, and the choice in (139) is natural in order to normalize the M family. Since X_T is negative, M is positive and we are done. □

There is also a converse of the result above.

Proposition 10.5. *Consider a given family of positive martingales $M(t,T)$ indexed by T and a positive deterministic function Φ. Then the specification*

$$p(t,T) = \frac{\int_T^\infty \Phi(s)M(t,s)\,\mathrm{d}s}{\int_t^\infty \Phi(s)M(t,s)\,\mathrm{d}s} \tag{142}$$

will define an arbitrage free positive system of bond prices. Furthermore, the stochastic discount factor Z generating the bond prices is given by

$$Z_t = \int_t^\infty \Phi(s)M(t,s)\,\mathrm{d}s. \tag{143}$$

Proof. Using the martingale property of the M family, we obtain

$$E^P[Z_T \mid \mathcal{F}_t] = \int_T^\infty E^P\big[\Phi(s)M(T,s) \mid \mathcal{F}_t\big]\,\mathrm{d}s = \int_T^\infty \Phi(s)M(t,s)\,\mathrm{d}s.$$

This implies, by the positivity of M and Φ, that Z is a potential and can thus serve as a stochastic discount factor. The induced bond prices are thus given by

$$p(t,T) = \frac{E^P[Z_T \mid \mathcal{F}_t]}{Z_t},$$

and the calculation above shows that the induced (arbitrage free) bond prices are given by (142). □

We can also easily compute forward rates.

Proposition 10.6. *With bond prices given by (142), forward rates are given by*

$$f(t,T) = \Phi(T)M(t,T), \tag{144}$$

and the short rate has the form

$$r_t = \Phi(t)M(t,t). \tag{145}$$

Proof. Follows directly from (142) and the formula $f(t,T) = \frac{\partial}{\partial T} \ln p(t,T)$. □

The most used instance of a Flesaker–Hughston model is the so-called **rational model**. In such a model we consider a given martingale K and two deterministic positive functions $\alpha(t)$ and $\beta(t)$. We then define the M family by

$$M(t,T) = \alpha(T) + \beta(T)K(t). \tag{146}$$

With this specification of M it is easily seen that bond prices will have the form

$$p(t,T) = \frac{A(T) + B(T)K(t)}{A(t) + B(t)K(t)} \tag{147}$$

where

$$A(t) = \int_t^\infty \Phi(s)\alpha(s)\,\mathrm{d}s, \qquad B(t) = \int_t^\infty \Phi(s)\beta(s)\,\mathrm{d}s.$$

We can specialize this further by assuming K to be of the form

$$K(t) = e^{\int_0^t \gamma(s)\,\mathrm{d}W_s - \frac{1}{2}\int_0^t \gamma^2(s)\,\mathrm{d}s}$$

where γ is deterministic. Then K will be a lognormal martingale, and the entire term structure will be analytically very tractable.

10.3 Changing base measure

The arguments above do not at all depend upon the fact that P was assumed to be the objective probability measure. If instead we work with another base measure $P^0 \sim P$, we will of course have a Flesaker–Hughston representation of bond prices of the form

$$p(t,T) = \frac{\int_T^\infty \Phi(s)M^0(t,s)\,\mathrm{d}s}{\int_t^\infty \Phi(s)M^0(t,s)\,\mathrm{d}s}, \tag{148}$$

where $M^0(t,T)$ is a family of positive P^0 martingales, and the question is how M^0 relates to M.

Proposition 10.7. *With notation as above, we have*

$$M^0(t,T) = \frac{M(t,T)}{L_t^0}, \tag{149}$$

where

$$L_t^0 = \frac{\mathrm{d}P^0}{\mathrm{d}P} \quad \text{on } \mathcal{F}_T. \tag{150}$$

Proof. From (138) we have, modulo scaling, the relation

$$M(t,T) = E^P\big[r_T B_T^{-1} L_T \mid \mathcal{F}_t\big]$$

where $L = \mathrm{d}Q/\mathrm{d}P$, so by symmetry we have

$$M^0(t,T) = E^{P^0}\left[r_T B_T^{-1}\left\{\frac{\mathrm{d}Q}{\mathrm{d}P^0}\right\}_T \,\middle|\, \mathcal{F}_t\right]$$

where

$$\left\{\frac{\mathrm{d}Q}{\mathrm{d}P^0}\right\}_T = \frac{\mathrm{d}Q}{\mathrm{d}P^0} \quad \text{on } \mathcal{F}_t.$$

We now obtain, using the Bayes formula,

$$\begin{aligned} M^0(t,T) &= E^{P^0}\left[r_T B_T^{-1}\left\{\frac{\mathrm{d}Q}{\mathrm{d}P^0}\right\}_T \,\middle|\, \mathcal{F}_t\right] \\ &= E^{P^0}\left[r_T B_T^{-1}\left\{\frac{dQ}{dP}\right\}_T \left\{\frac{dP}{dP^0}\right\}_T \,\middle|\, \mathcal{F}_t\right] \\ &= \left\{\frac{\mathrm{d}P}{\mathrm{d}P^0}\right\}_t \cdot E^P\left[r_T B_T^{-1}\left\{\frac{dQ}{dP}\right\}_T \,\middle|\, \mathcal{F}_t\right] = \frac{M(t,T)}{L_t^0}. \end{aligned}$$

□

10.4 Multi-currency models

The potential setup above can easily be extended to a multi-currency situation. Let us assume that we have N countries, that the economy is arbitrage free and complete, that there is a money account B^i for each country, and that there are no frictions on international trade.

Definition 10.2. The exchange rate process Y_t^{ij} denotes the price, at time t in currency j, of one unit of currency i, and let Z^i denote the SDF under P for country No. i (w.r.t. its own currency).

Now choose an arbitrary but fixed T-claim Ψ expressed in currency i. The arbitrage free price of this claim, at $t = 0$ in currency i will of course be given by the expression

$$\Pi(0;\Psi)^i = E^P\big[Z_T^i \Psi\big], \tag{151}$$

so the arbitrage free price expressed in currency j will be given by

$$\Pi(0;\Psi)^j = Y_0^{ij} E^P\big[Z_T^i \Psi\big]. \tag{152}$$

On the other hand, the claim Ψ is, at time T, worth exactly

$$Y_T^{ij}\Psi$$

units of currency j, so we can also compute the j price at $t=0$ as

$$\Pi(0;\Psi)^j = E^P\big[Z_T^j Y_T^{ij}\Psi\big]. \tag{153}$$

We thus have

$$Y_0^{ij} E^P\big[Z_T^i\Psi\big] = E^P\big[Z_T^j Y_T^{ij}\Psi\big],$$

and if this holds for every T-claim Ψ, we must have

$$Y_0^{ij} Z_T^i = Z_T^j Y_T^{ij}$$

and we have the following result.

Proposition 10.8. *With assumptions as above, exchange rates are related to SDFs according to the formula*

$$Y_t^{ij} = \frac{Z_t^i}{Z_t^j} Y_0^{ij}. \tag{154}$$

The moral of this is that exchange rates are determined completely by a consistent specification of stochastic discount factors for each country. All this can also be expressed in the Flesaker–Hughston terminology. Denoting by $M^i(t,T)$ and Φ^i the Flesaker–Hughston family of martingales and scaling processes in country i, we immediately have, from (143)

$$Y_t^{ij} = Y_0^{ij} \cdot \frac{\int_t^\infty \Phi^i(s) M^i(t,s)\,\mathrm{d}s}{\int_t^\infty \Phi^j(s) M^j(t,s)\,\mathrm{d}s}. \tag{155}$$

10.5 Connections to the Riesz decomposition

In Section 10.1 we saw that any SDF generating a nice bond market is a potential, so from a modeling point of view it is natural to ask how one can construct potentials from scratch.

The main result used is the following.

Proposition 10.9 *(Riesz Decomposition). If Z is a potential, then it admits a representation as*

$$Z_t = -A_t + M_t, \tag{156}$$

where A is an increasing process, and M is a martingale defined by

$$M_t = E^P[A_\infty \mid \mathcal{F}_t]. \tag{157}$$

To construct a potential, let us assume that we define A as

$$A_t = \int_0^t a_s \, ds \tag{158}$$

for some integrable nonnegative process a. Then we easily obtain

$$Z_t = E^P\left[\int_0^\infty a_s \, ds \mid \mathcal{F}_t\right] - \int_0^t a_s \, ds = \int_t^\infty E^P[a_s \mid \mathcal{F}_t] \, ds. \tag{159}$$

We can now connect this to the Flesaker–Hughston framework. The family of processes $X(t, T)$ defined in (133) will, in the present framework, have the form

$$X(t, T) = E^P\left[\int_T^\infty E^P[a_s \mid \mathcal{F}_T] \, ds \;\middle|\; \mathcal{F}_t\right] = \int_T^\infty E^P[a_s \mid \mathcal{F}_t] \, ds, \tag{160}$$

so the basic family of Flesaker–Hughston martingales are given by

$$M(t, T) = -\frac{\partial}{\partial T} X(t, T) = E^P[a_T \mid \mathcal{F}_t]. \tag{161}$$

10.6 Conditional variance potentials

An alternative way of representing potentials which have been studied in depth by Hughston and co-authors is through conditional variances.

Consider a fixed random variable $X_\infty \in L^2(P, \mathcal{F}_\infty)$. We can then define a martingale X by setting

$$X_t = E^P[X_\infty \mid \mathcal{F}_t]. \tag{162}$$

Now let us define the process Z by

$$Z_t = E^P\big[(X_\infty - X_t)^2 \;\big|\; \mathcal{F}_t\big]. \tag{163}$$

An easy calculation shows that

$$Z_t = E^P\big[X_\infty^2 \;\big|\; \mathcal{F}_t\big] - X_t^2. \tag{164}$$

Since the first term is a martingale and the second is a submartingale, the difference is a supermartingale, which by definition is positive and it is in fact a potential.

The point of this is that the potential Z, and thus the complete interest rate model generated by Z, is in fact fully specified by a specification of the single random variable X_∞. A very interesting idea is now to expand X_∞ in Wiener chaos. See the Notes below.

10.7 The Rogers Markov potential approach

As we have seen above, in order to generate an arbitrage free bond market model it is enough to construct a positive supermartingale to act as stochastic discount factor (SDF), and in the previous section we saw how to do this using the Riesz decomposition. In this section we will present a systematic way of constructing potentials along the lines above, in terms of Markov processes and their resolvents. The ideas are due to Rogers (1994), and we largely follow his presentation.

We consider a time homogeneous Markov process X under the objective measure P, with infinitesimal generator $\mathcal{G}$.

For any positive real valued sufficiently integrable function g and any positive number α we can now define the process A in the Riesz decomposition (156) as

$$A_t = \int_0^t e^{-\alpha s} g(X_s)\,\mathrm{d}s, \tag{165}$$

where the exponential is introduced in order to allow at least all bounded functions g. In terms of the representation (158) we thus have

$$a_t = e^{-\alpha t} g(X_t), \tag{166}$$

and a potential Z is, according to (159), obtained by

$$Z_t = \int_t^\infty e^{-\alpha s} E^P\big[g(X_s) \mid \mathcal{F}_t\big]\,\mathrm{d}s. \tag{167}$$

Using the Markov assumption we thus have

$$Z_t = E^P\left[\int_t^\infty e^{-\alpha s} g(X_s)\,\mathrm{d}s \mid X_t\right], \tag{168}$$

and this expression leads to a well known probabilistic object.

Definition 10.3. For any nonnegative α the **resolvent** R_α is an operator, defined for any bounded measurable function g by the expression

$$R_\alpha g(x) = E_x^P\left[\int_0^\infty e^{-\alpha s} g(X_s)\,\mathrm{d}s\right] \tag{169}$$

where subindex x denotes the conditioning $X_0 = x$.

We can now connect resolvents to potentials.

Proposition 10.10. *For any bounded nonnegative g, the process*

$$Z_t = e^{-\alpha t}\frac{R_\alpha g(X_t)}{R_\alpha g(X_0)} \tag{170}$$

is a potential with $Z_0 = 1$.

Proof. The normalizing factor is trivial so we disregard it in the rest of the proof. Using time invariance we have, from (168),

$$Z_t = E^P\left[\int_0^\infty e^{-\alpha(t+s)}g(X_{t+s})\,\mathrm{d}s \mid X_t\right] = e^{-\alpha t}R_\alpha g(X_t). \tag{171}$$

□

Given a SDF of the form above, we can of course compute bond prices, and the short rate can easily be recovered.

Proposition 10.11. *If the stochastic discount factor Z is defined by (170) then bond prices are given by*

$$p(t,T) = e^{-\alpha(T-t)}\frac{E^P[R_\alpha g(X_T) \mid \mathcal{F}_t]}{R_\alpha g(X_t)} \tag{172}$$

and the short rate is given by

$$r_t = \frac{g(X_t)}{R_\alpha g(X_t)}. \tag{173}$$

Proof. The formula (172) follows directly from (170) and the general formula (129). From (167) we easily obtain

$$\mathrm{d}Z_t = -e^{-\alpha t}g(X_t)\,\mathrm{d}t + \mathrm{d}M_t,$$

where M is a martingale defined by

$$M_t = E^P\left[\int_0^\infty e^{-\alpha s}g(X_s) \mid \mathcal{F}_t\right]\mathrm{d}s$$

and (173) now follows from Proposition 10.3. □

One problem with this scheme is that, for a concrete case, it may be very hard to compute the quotient in (173). To overcome this difficulty we recall the following standard result.

Proposition 10.12. *With notation as above we have essentially*

$$R_\alpha = (\alpha - \mathcal{G})^{-1}. \tag{174}$$

The phrase "essentially" above indicates that the result is "morally" correct, but that care has to be taken concerning the domain of the operators. We now provide a lighthearted heuristic argument for the result. The point of the argument below is not to provide a precise proof, but to build an intuition which allows us to guess the result. For a rigorous proof, see any textbook on Markov processes.

Proof sketch. Define a family of operators $\{S_t;\ t \geqslant 0\}$ by

$$S_t g(x) = E_x^P\big[g(X_t)\big]. \tag{175}$$

By time invariance it is now easily seen that we also can write

$$S_t g(x) = E^P\big[g(X_{u+t}) \mid X_u = x\big]. \tag{176}$$

In particular we can use the Markovian assumption to obtain

$$S_{t+s} g(x) = S_t S_s g(x). \tag{177}$$

Thus we have $S_{t+s} = S_t S_s$, so the family S_t forms a semigroup. By definition, the infinitesimal generator $\mathcal{G}$ is given by

$$\mathcal{G} g(x) = \frac{\mathrm{d}}{\mathrm{d}t} S_t g(x), \quad \text{at } t = 0. \tag{178}$$

Using (177) and $S_0 = I$ where I is the identity operator we have

$$\frac{S_{t+h} - S_t}{h} = \frac{S_h - I}{h} S_t, \tag{179}$$

which (at least formally) gives us the Kolmogorov equation

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = \mathcal{G} S_t. \tag{180}$$

Since $\mathcal{G}$ is a linear operator we expect to be able to solve this ODE as

$$S_t = e^{\mathcal{G}t}, \tag{181}$$

and this can indeed be shown to be correct at a certain technical cost. (If we have a finite state space for X, then $\mathcal{G}$ is a square matrix and there are no problems, but in the general case $\mathcal{G}$ is an unbounded operator and care has to be taken.)

With this formalism we should, at least morally, be able to write

$$R_\alpha g(x) = \int_0^\infty e^{-\alpha s} e^{\mathcal{G}s} g(x)\,\mathrm{d}s = \left(\int_0^\infty e^{(\mathcal{G}-\alpha)s}\,\mathrm{d}s\right) g(x). \tag{182}$$

By formal integration (acting as if $\mathcal{G}$ is a real number), we obtain

$$\int_0^\infty e^{(\mathcal{G}-\alpha)s}\,\mathrm{d}s = (\mathcal{G} - \alpha)^{-1}\{0 - I\} = (\alpha - \mathcal{G})^{-1}, \tag{183}$$

which is the result.
The end of proof sketch. □

We now go back to the Rogers scheme and using the identity $R_\alpha = (\alpha - \mathcal{G})^{-1}$ we see that with $f = R_\alpha g$ we have

$$\frac{g(X_t)}{R_\alpha g(X_t)} = \frac{(\alpha - \mathcal{G})f(X_t)}{f(X_t)},$$

where it usually is a trivial task to compute the last quotient.

This led Rogers to use the following scheme.

1. Fix a Markov process X, number α and a nonnegative function f.
2. Define g by
$$g = (\alpha - \mathcal{G})f.$$
3. Choose α (and perhaps the parameters of f) such that g is nonnegative.
4. Now we have $f = R_\alpha g$, and the short rate can be recaptured by
$$r(t) = \frac{(\alpha - \mathcal{G})f(X_t)}{f(X_t)}.$$

In this way Rogers produces a surprising variety of concrete analytically tractable nonnegative interest rate models and, using arguments of the type in Section 10.4 above, exchange rate models are also treated within the same framework.

For illustration we consider the simplest possible example of a potential model, where the underlying Markov process is an n-dimensional Gaussian diffusion of the form

$$\mathrm{d}X_t = -AX_t\,\mathrm{d}t + \mathrm{d}W_t. \tag{184}$$

In this case we have

$$\mathcal{G}f(x) = \frac{1}{2}\Delta f(x) - \nabla f(x)Ax \tag{185}$$

where Δ is the Laplacian and ∇f is the gradient viewed as a row vector.

We now define f by

$$f(x) = e^{cx}$$

for some row vector $c \in R^n$. We immediately obtain

$$g(x) = (\alpha - \mathcal{G})f(x) = f(x)\left(\alpha - \frac{1}{2}\|c\|^2 + cAx\right).$$

The corresponding short rate is given by

$$r_t = \alpha - \frac{1}{2}\|c\|^2 + cAx, \tag{186}$$

so we have a Gaussian multi factor model.

We end this section by connecting the Rogers theory to the Flesaker–Hughston framework, and this is quite straightforward. Comparing (161) to (166) we have

$$M(t,T) = e^{-\alpha T} E^P\big[g(X_T) \mid \mathcal{F}_t\big] \tag{187}$$

10.8 The term structure density approach

One of the major problem of the Heath–Jarrow–Morton framework for forward rates is that there is no easy way to specify forward rate volatilities in such a way that they guarantee positive interest rates. To address this problem (and others), Brody and Hughston have in a series of papers studied interest rate models based on the "term structure density" and we now go on to present some of the main ideas.

Definition 10.4. Let, as usual, $p(t,T)$ denote the price at t of a zero coupon bond maturing at t. Now define $p_t(x)$ and $q_t(x)$ by

$$p_t(x) = p(t, t+x), \tag{188}$$

$$q_t(x) = -\frac{\partial}{\partial x} p_t(x). \tag{189}$$

Thus $p_t(x)$ is the bond price in Musiela parameterization.

We now notice that for a model with positive interest rates, p_t as a function of x has the property that it is decreasing, $p_t(0) = 1$ and $p_t(\infty) = 0$. In other words, p_t has all the properties of a complementary probability distribution. The object q_t is the associated density (always assumed to exist), and the idea of Brody and Hughston is to study the evolution of q_t in a Wiener driven framework.

Working under the objective measure P we can easily derive the p_t dynamics. From general theory we know that the $p(t,T)$ dynamics are of the form

$$\mathrm{d}p(t,T) = p(t,T) r_t \,\mathrm{d}t + \Sigma^0(t,T)\{\mathrm{d}W_t + \lambda_t \,\mathrm{d}t\} \tag{190}$$

where Σ^0 denotes the total volatility and λ is the market price of risk process. Arguing as in the derivation of the Musiela equation for forward rates, we can easily obtain the following dynamics for p_t:

$$\mathrm{d}p_t(x) = \big\{\mathbf{F}p_t(x) + r_t p_t(x)\big\}\,\mathrm{d}t + \Sigma_t(x)\{\mathrm{d}W_t + \lambda_t \mathrm{d}t\}, \tag{191}$$

where

$$\mathbf{F} = \frac{\partial}{\partial x}, \qquad \Sigma_t(x) = \Sigma^0(t, t+x).$$

Noticing that $\mathbf{F}p_t = q_t$, and taking x-derivatives in (190), we obtain the equation

$$\mathrm{d}q_t(x) = \big\{\mathbf{F}q_t(x) + r_t q_t(x)\big\}\,\mathrm{d}t + \omega_t(x)\{\mathrm{d}W_t + \lambda_t \,\mathrm{d}t\}, \tag{192}$$

where

$$\omega_t(x) = \frac{\partial}{\partial x}\Sigma_t(x).$$

Brody and Hughston now observe that since $p_t(\infty) = 0$ in any positive model, we must have $\Sigma_t(\infty) = 0$. We also know from (5) that $\Sigma_t(0) = 0$ so the volatility ω has the property that

$$\int_0^\infty \omega_t(x)\,\mathrm{d}x = 0. \tag{193}$$

Thus we cannot choose ω freely, so in order to isolate the degrees of freedom we instead write ω_t as

$$\omega_t(x) = q_t(x)\big(\nu_t(x) - \bar{\nu}_t\big), \tag{194}$$

where ν_t is chosen freely and

$$\bar{\nu}_t = \int_0^\infty q_t(x)\nu_t(x)\,\mathrm{d}x. \tag{195}$$

We also notice that the bond price volatility $\Sigma_t(x)$ is invariant under all transformations of the form $\nu_t(x) \to \nu_t(x) + \beta_t$ so we can normalize by setting $\bar{\nu}_t = \lambda_t$. We have thus derived the following basic result, which is the starting point of further investigations of Brody and Hughston.

Proposition 10.13. *For a positive interest rate model, the q dynamics are of the form*

$$\mathrm{d}q_t(x) = \big\{\mathbf{F}q_t(x) + r_t q_t(x)\big\}\,\mathrm{d}t + q_t(x)\big(\nu_t(x) - \bar{\nu}_t\big)\{\mathrm{d}W_t + \bar{\nu}_t\,\mathrm{d}t\}. \tag{196}$$

10.9 Notes

For general information on stochastic discount factors and asset pricing, see the textbook Cochrane (2001). The Flesaker–Hughston fractional model was developed in Flesaker and Hughston (1996) and Flesaker and Hughston (1997), and the connection between general potential theory and the Flesaker–Hughston approach is discussed in Jin and Glasserman (2001). In Brody and Hughston (2004) and Hughston and Rafailidis (2005) the conditional variance approach and Wiener chaos expansions are investigated in depth. The Rogers Markovian potential approach was first presented in Rogers (1994). The term structure density model has been studied in Brody and Hughston (2001) and Brody and Hughston (2002).

References

Bhar, R., Chiarella, C. (1997). Transformation of Heath–Jarrow–Morton models to Markovian systems. *The European Journal of Finance* 3, 1–26.

Bingham, N., Kiesel, R. (2004). *Risk Neutral Valuation*, second ed. Springer.

Björk, T. (2003). *Arbitrage Theory in Continuous Time*, second ed. Oxford Univ. Press.

Björk, T., Cristensen, B. (1999). Interest rate dynamics and consistent forward rate curves. *Mathematical Finance* 9 (4), 323–348.

Björk, T., Landén, C. (2002). On the construction of finite dimensional realizations for nonlinear forward rate models. *Finance and Stochastics* 6 (3), 303–331.

Björk, T., Svensson, L. (2001). On the existence of finite dimensional realizations for nonlinear forward rate models. *Mathematical Finance* 11 (2), 205–243.

Björk, T., Landén, C., Svensson, L. (2002). Finite dimensional Markovian realizations for stochastic volatility forward rate models. Working paper in Economics and Finance. Stockholm School of Economics.

Brace, A., Musiela, M. (1994). A multifactor Gauss Markov implementation of Heath, Jarrow, and Morton. *Mathematical Finance* 4, 259–283.

Brace, A., Gatarek, D., Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance* 7, 127–154.

Brigo, D., Mercurio, F. (2001). *Interest Rate Models*. Springer.

Brocket, P. (1981). Nonlinear systems and nonlinear estimation theory. In: Hazewinkel, M., Willems, J. (Eds.), *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*. Reidel.

Brody, D., Hughston, L. (2001). Interest rates and information geometry. *Proceedings of the Royal Society of London, Series A* 457, 1343–1363.

Brody, D., Hughston, L. (2002). Entropy and information in the interest rate term structure. *Quantitative Finance* 2, 70–80.

Brody, D., Hughston, L. (2004). Chaos and coherence: A new framework for interest rate modelling. *Proceedings of the Royal Society of London, Series A* 460, 85–110.

Cairns, A.D. (2004). *Interest Rate Models*. Princeton Univ. Press.

Carverhill, A. (1994). When is the spot rate Markovian? *Mathematical Finance* 4, 305–312.

Chiarella, C., Kwon, O.K. (2001). Forward rate dependent Markovian transformations of the Heath–Jarrow–Morton term structure model. *Finance and Stochastics* 5, 237–257.

Cochrane, J. (2001). *Asset Pricing*. Princeton Univ. Press.

Da Prato, G., Zabzcyk, J. (1992). *Stochastic Equations in Infinite Dimensions*. Cambridge Univ. Press.

Duffie, D. (2001). *Dynamic Asset Pricing Theory*, third ed. Princeton Univ. Press.

Eberlein, E., Raible, S. (1999). Term structure models driven by general Levy processes. *Mathematical Finance* 9 (1), 31–53.

Filipović, D. (1999). A note on the Nelson–Siegel family. *Mathematical Finance* 9 (4), 349–359.

Filipović, D. (2000a). Exponential-polynomial families and the term structure of interest rates. *Bernoulli* 6, 1–27.

Filipović, D. (2000b). Invariant manifolds for weak solutions of stochastic equations. *Probability Theory and Related Fields* 118, 323–341.

Filipović, D. (2001). *Consistency Problems for Heath–Jarrow–Morton Interest Rate Models*. *Springer Lecture Notes in Mathematics, vol. 1760*. Springer-Verlag.

Filipović, D., Teichmann, J. (2002). On finite dimensional term structure models. Working paper.

Filipović, D., Teichmann, J. (2003). Existence of invariant manifolds for stochastic equations in infinite dimension. *Journal of Functional Analysis* 197, 398–432.

Filipović, D., Teichmann, J. (2004). On the geometry of the term structure of interest rates. *Proceedings of the Royal Society* 460, 129–167.

Flesaker, B., Hughston, L. (1996). Positive interest. *Risk* 9, 46–49.

Flesaker, B., Hughston, L. (1997). International models for interest rates and foreign exchange. *Net Exposure* 3, 55–79.

Heath, D., Jarrow, R., Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica* 60, 77–105.

Hughston, L., Rafailidis, A. (2005). A chaotic approach to interest rate modelling. *Finance and Stochastics* 9, 43–65.

Hunt, P., Kennedy, J. (2000). *Financial Derivatives in Theory and Practice*. Wiley.

Inui, K., Kijima, M. (1998). A Markovian framework in multi-factor Heath–Jarrow–Morton models. *Journal of Financial and Quantitative Analysis* 33, 423–440.

Isidori, A. (1989). *Nonlinear Control Systems*. Springer-Verlag.

Jamshidian, F. (1997). Libor and swap market models and measures. *Finance and Stochastics* 1, 293–330.

Jeffrey, A. (1995). Single factor Heath–Jarrow–Morton term structure models based on Markov spot interest rate dynamics. *Journal of Financial and Quantitative Analysis* 30, 619–642.

Jin, Y., Glasserman, P. (2001). Equilibrium positive interest rates: A unified view. *Review of Financial Studies* 14, 187–214.

Miltersen, K., Sandmann, K., Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance* 52, 409–430.

Musiela, M. (1993). Stochastic PDEs and term structure models. Preprint.

Nelson, C., Siegel, A. (1987). Parsimonious modelling of yield curves. *Journal of Business* 60, 473–489.

Protter, P. (2000). *Efficient Methods of Valuing Interest Rate Derivatives*. Springer-Verlag.

Ritchken, P., Sankarasubramanian, L. (1995). Volatility structures of forward rates and the dynamics of the term structure. *Mathematical Finance* 5 (1), 55–72.

Rogers, L. (1994). The potential approach to the term structure of interest rates and foreign exchange rates. *Mathematical Finance* 7, 157–176.

Warner, F. (1979). *Foundations of Differentiable Manifolds and Lie Groups*. Scott, Foresman, Hill.

Zabczyk, J. (2001). Stochastic invariance and consistency of financial models. Preprint. Scuola Normale Superiore, Pisa.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15010-6

Chapter 10
Calculating Portfolio Credit Risk*

Paul Glasserman
403 Uris Hall, Columbia Business School, New York, NY 10027, USA
E-mail: pg20@columbia.edu

Abstract

This chapter provides an overview of modeling and computational issues associated with portfolio credit risk. We consider the problem of calculating the loss distribution in a portfolio of assets exposed to credit risk, such as corporate bonds or bank loans. We also discuss the pricing of portfolio credit derivatives, such as basket default swaps and collateralized debt obligations. A portfolio view of credit risk requires capturing dependence between the assets in the portfolio; we discuss models of dependence and associated computational techniques. A standard modeling framework takes the assets to be conditionally independent given a set of underlying factors, and this is convenient for computational purposes. We discuss a recursive convolution technique, transform inversion, saddlepoint approximation, and importance sampling for Monte Carlo simulation.

1 Introduction

This chapter provides an overview of modeling and computational issues associated with portfolio credit risk. A simple example of the type of setting we consider is a portfolio of corporate bonds or bank loans. The promised cashflows of these underlying assets are known, but there is a risk that an issuer or borrower will default and fail to make the promised payments. This type of risk should be contrasted with market risk – the fluctuations in the market levels of interest rates, exchange rates, stock prices and other asset prices. Of course, a corporate bond is also subject to interest rate risk and other types of risk, but we focus on the credit risk resulting from the possibility of default.

Credit risk models are generally somewhat simpler than the models used for the dynamics of asset prices, in large part because of the limited data available

* This work is supported by NSF grants DMI0300044 and DMS0410234.

to support complex models. For equities and currencies, continuous trading both in quantity and time is a reasonable approximation of reality, and new price information arrives virtually continuously. This leads to diffusion models of prices and models for derivative pricing based on dynamic replication or hedging. In contrast, defaults are comparatively rare events and the mechanics of default depend on the inner workings of a firm, which are at best only partly observable. The most commonly used models and those we consider here are fairly simple, often just modeling the occurrence of default through a binary random variable or the time to default through a nonnegative random variable.

Taking a portfolio view of credit risk entails capturing dependence between the creditworthiness of different borrowers or *obligors*, and that is the focus of this chapter. Defaults exhibit dependence because firms operate in a common economic environment (defaults are more common during recessions) and may share exposure to risks associated with a particular geographic region or industry. These types of factors underlie portfolio models. A model of the creditworthiness of a single firm often incorporates balance sheet information and describes the mechanisms leading to the firm's default. This type of analysis is used to determine credit ratings and to determine the fundamental value of a corporate bond or a convertible security. In contrast, a portfolio model often takes information about the credit quality of individual firms (through ratings or bond spreads, for example) as inputs and combines this with a model of dependence to arrive at the overall credit risk in a portfolio.

Portfolio credit risk models are used both for risk management and for the pricing of derivative securities that are exposed to multiple sources of credit risk. In risk management applications, the primary objective is measuring the distribution of losses in a portfolio over a fixed horizon. This loss distribution is typically summarized through a single risk measure, such as value-at-risk or expected shortfall. A further risk management objective is decomposing the overall risk in a portfolio into risk contributions associated with individual counterparties or transactions. Examples of portfolio credit derivatives include basket credit default swaps and collateralized debt obligations (described later in this chapter); the cashflows of these securities depend on the timing of defaults of a set of underlying bonds or firms.

In both types of applications, dependence between defaults complicates the computation of the distribution of losses. In the models we consider, defaults are conditionally independent, given a set of common factors. This structure is important to the computational procedures we discuss. In its simplest form, the computational problem we consider reduces to finding the distribution of a sum of conditionally independent random variables.

The rest of this chapter is organized as follows. Section 2 gives a more detailed description of the risk management and pricing applications we consider. Section 3 describes models of dependence between defaults. Section 4 discusses the calculation of conditional loss distributions – i.e., the loss distribution conditional on the factors that make defaults independent. Section 5 extends the methods of Section 4 to unconditional loss distributions. Section 6

discusses importance sampling for portfolio credit risk and Section 7 concludes the chapter.

2 Problem setting

Throughout, we consider a portfolio or individual security that is exposed to the risk of default by any of m different *obligors*, also referred to as *counterparties, credits*, or simply *names*. We let

$$\tau_k = \text{the time of default of obligor } k, \quad k = 1, \ldots, m,$$

with $\tau_k = \infty$ if the kth obligor never defaults. For a fixed horizon T, we set

$$Y_k = \mathbf{1}\{\tau_k \leqslant T\}, \quad k = 1, \ldots, m,$$

the indicator that the kth obligor defaults in $[0, T]$.

A default triggers a loss. In the simplest case, the default of an obligor results in the loss of a fixed amount known in advance – the full amount of a loan or the value of a bond, for example. In practice, several considerations may make the loss upon default depend on the time of default and other factors. If the obligor is making regular payments on a loan or coupon payments on a bond, then the size of the loss depends on the time of default. Also, creditors will often recover some of the value of loans, bonds and other credit instruments after default through bankruptcy proceedings or by selling defaulted assets; the loss is then reduced by the amount recovered. We will not consider the problem of modeling losses upon default but instead set

$$V_k = \text{loss upon default of obligor } k, \quad k = 1, \ldots, m.$$

In fact, we will often take these to be constants, in which case we denote them by v_k. We take the V_k to be nonnegative, though it is sometimes possible for the default of a counterparty in a swap, for example, to result in a gain rather than a loss.

Two of the central problems in credit risk are (i) modeling the marginal distribution of the time to default for a single obligor and (ii) modeling the dependence between the default times of multiple obligors. We will detail models used for the second issue in Section 3. The first problem is often addressed by specifying a hazard rate or, more generally, a default intensity. For example, the distribution of the time to default is often represented as

$$P(\tau_k \leqslant t) = 1 - \exp\left(-\int_0^t h_k(s)\,\mathrm{d}s\right), \tag{1}$$

with h_k a (deterministic) hazard rate with the interpretation that the probability of default in $(t, t + \delta)$ given that default has not occurred by time t is $h_k(t)\delta + o(\delta)$. In a more general stochastic intensity model of default times,

the hazard rate is replaced by a stochastic process; see, e.g., Duffie and Singleton (1999), Duffie and Garleanu (2001), Giesecke and Tomecek (2005), and Jarrow et al. (1997).

Default hazard rates are typically inferred from yield spreads on corporate bonds and spreads in credit default swaps, which are, in effect, default insurance contracts. Information about the time to default can also be gleaned from economic variables, an obligor's credit rating and its financial statements, when available. To be precise, we must distinguish between the distribution of time to default under a risk-adjusted probability measure and the distribution observed empirically. The first is potentially observable in market prices and credit spreads and is relevant to pricing credit risky instruments. The second is relevant to risk management applications – it is the actual distribution of time to default that matters for measuring risk – but it cannot be inferred from market prices without correcting for the market risk premium for bearing default risk. Das et al. (2005) estimate models of actual and risk-adjusted default probabilities and find that the relation between the two varies widely over time, with the latter often much larger than the former. Some models, such as Jarrow et al. (1997), adopt simple assumptions on the nature of the risk-adjustment for tractability; and in the absence of sufficient data, the distinction between the two types of probabilities is sometimes ignored.

We focus here on credit losses resulting from default, but at least two other aspects of credit risk are important in practice. Fluctuations in spreads in, e.g., corporate bonds, are a source of risk derived at least in part from changes in creditworthiness of the issuer; changes in spreads affect prices continually, not just upon default. (See, e.g., the models in Schonbucher, 2003.) Changes in an issuer's credit rating also affect the market value of a credit-sensitive instrument. A natural extension of a model based on default times allows for multiple states associated with ratings and transitions between them; see, in particular, Jarrow et al. (1997). From this perspective, a default-time model is the special case of a two-state specification. The credit risk associated with spread fluctuations and ratings transitions results from changes in market prices, rather than the failure of an obligor to make a contractual payment.

We turn now to a description of some of the main problems that arise in credit risk management and the pricing of credit derivatives.

2.1 Measuring portfolio credit risk

Over a fixed horizon T, the losses from default suffered by a fixed portfolio can be written as

$$L = Y_1V_1 + Y_2V_2 + \cdots + Y_mV_m, \tag{2}$$

where, as before, the Y_k are default indicators and V_k is the loss triggered by the default of obligor k. The problem of measuring portfolio credit risk is the problem of calculating the distribution of L. This distribution is often summarized through a single measure of risk. One widely used measure is

value-at-risk (VaR_α) at confidence level $1-\alpha$, often with $\alpha = 1\%$ or $\alpha = 0.01\%$. This is simply the $1-\alpha$ quantile of the distribution – the infimum over all x for which

$$P(L > x) \leqslant \alpha.$$

Thus, in the case of a continuous loss distribution, we have

$$P(L > VaR_\alpha) = \alpha.$$

A closely related risk measure is the expected shortfall ES_α, defined by

$$ES_\alpha = \mathsf{E}[L \mid L \geqslant VaR_\alpha].$$

Key to calculating either of these risk measures is accurate measurement of the tail of the loss random variable L. The difficulty of the problem depends primarily on the dependence assumed among the Y_k and V_k.

2.2 Measuring marginal risk contributions

The measurement of portfolio credit risk is often followed by a process of decomposing the total risk into a sum of *risk contributions* associated with individual obligors, subportfolios, or transactions. This decomposition is used to allocate capital and measure profitability. A standard decomposition of VaR $(= VaR_\alpha)$ sets

$$\begin{aligned} VaR &= \mathsf{E}[L \mid L = VaR] \\ &= \mathsf{E}[Y_1V_1 + Y_2V_2 + \cdots + Y_mV_m \mid L = VaR] \\ &= \mathsf{E}[Y_1V_1 \mid L = VaR] + \mathsf{E}[Y_2V_2 \mid L = VaR] + \cdots \\ &\quad + \mathsf{E}[Y_mV_m \mid L = VaR]. \end{aligned} \tag{3}$$

The risk contribution for obligor k is then $\mathsf{E}[Y_kV_k \mid L = VaR]$. This of course assumes that the event $\{L = VaR\}$ has positive probability; in practice, it may be necessary to condition on $|L-VaR| < \epsilon$, for some $\epsilon > 0$. By a similar argument, a portfolio's expected shortfall can be decomposed into risk contributions of the form $\mathsf{E}[Y_kV_k \mid L \geqslant VaR]$, $k = 1, \ldots, m$.

Both VaR and expected shortfall belong to the class of *positively homogeneous* risk measures (as in Artzner et al., 1999), and such risk measures admit convenient decompositions as a consequence of Euler's theorem for positively homogeneous functions. Suppose, for example, that the losses V_k are constants v_k and consider some measure of risk ρ viewed as a function of $v_1, \ldots, v_m$ with the joint distribution of $Y_1, \ldots, Y_m$ held fixed. Positive homogeneity means that

$$\rho(\gamma v_1, \gamma v_2, \ldots, \gamma v_m) = \gamma\rho(v_1, v_2, \ldots, v_m)$$

for all $\gamma \geqslant 0$. In addition to VaR and ES, the standard deviation of the loss has this property. Assuming differentiability of the risk measure and differentiating

both sides with respect to γ at $\gamma = 1$, we get

$$\sum_{k=1}^{m} v_k \frac{\partial \rho}{\partial v_k} = \rho(v_1, v_2, \ldots, v_m).$$

Thus, as observed in Garman (1999) and Litterman (1999), the weighted sum of sensitivities on the left gives a decomposition over obligors of the total risk on the right. In this decomposition, $v_k \partial\rho/\partial v_k$ is the kth obligor's contribution to the total risk in the portfolio.

Subject only to modest regularity conditions, the risk contributions for *VaR* and *ES* defined through conditional expectations (as in (3)) coincide with those obtained as weighted sensitivities of each risk measure; see Gourieroux et al. (2000), Kurth and Tasche (2003) and Tasche (1999). These decompositions based on weighted sensitivities have also been shown to satisfy sets of axioms for sensible risk allocation; see Denault (2001) and Kalkbrener (2005). For computational purposes, the representation as conditional expectations is the most transparent, but calculating a large number of conditional expectations, all conditioned on a rare event, remains a computationally demanding problem; see Glasserman (2005), Kalkbrener et al. (2004) and Martin et al. (2001b).

2.3 Pricing nth-to-default swaps

Credit default swaps are among the most actively traded credit derivatives. A credit default swap functions like an insurance contract on a bond. The protection buyer in the swap makes regular premium payments to the protection seller until either the expiration of the swap or the default of the obligor referenced in the swap. If such a default occurs during the life of the swap, the protection seller makes a payment to the buyer; this payment would often equal the face value of a bond referenced in the swap. Thus, the swap insures the buyer against a loss from default.

The owner of a portfolio of bonds might consider buying such protection for every bond in the portfolio, but the payments required on the various swaps would make this expensive. An alternative to insuring each bond separately would be to enter into a basket default swap that references a basket of bonds. The simplest such structure is a first-to-default swap, which provides protection against the first default in a basket. The swap terminates upon the occurrence of the first default (or the expiration of the swap), and provides no protection against subsequent defaults. This is far less expensive than buying separate protection for each bond and provides adequate coverage if the chance of multiple defaults is small.

More generally, an nth-to-default swap provides protection against only the nth default in a basket. Let $\tau_1, \ldots, \tau_m$ be the default times of the names in the basket and let

$$\tau_{(1)} \leqslant \tau_{(2)} \leqslant \cdots \leqslant \tau_{(m)}$$

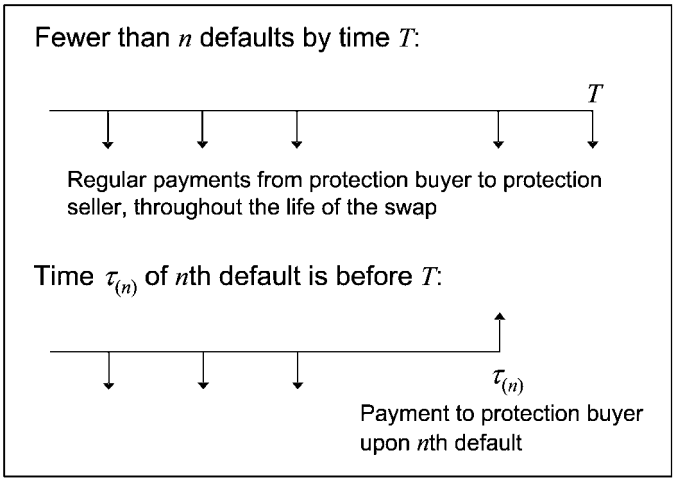


Fig. 1. Cashflows of an nth-to-default swap.

be their order statistics, so that $\tau_{(n)}$ is the time of the nth default. The protection buyer in an nth-to-default swap makes regular payments to the protection seller until $\min\{\tau_{(n)}, T\}$ – that is, until the nth default or the maturity T, whichever comes first. If $\tau_{(n)} \leqslant T$, then upon the nth default the protection seller makes a payment to the protection buyer to compensate for the loss resulting from the nth default; if $\tau_{(n)} > T$, the protection seller pays nothing (see Fig. 1). Thus, if $n > 1$, the protection buyer assumes the risk of the first $n - 1$ defaults.

The price of an nth-to-default swap depends primarily on the probability that the nth default occurs before time T. (The relevant default probabilities for pricing are the risk-adjusted or risk-neutral probabilities, not the empirical probabilities.) We say "primarily" because the timing of defaults also matters. The probability that $\tau_{(n)}$ is less than T is strongly influenced by the dependence between the individual default times $\tau_1, \ldots, \tau_m$.

The marginal distribution of each τ_i is, of course, also important. The (risk-adjusted) distribution of τ_i is at least partly determined by the market prices of credit default swaps on the ith obligor: a credit default swap with maturity T referencing obligor i contains information about the probability that $\tau_i < T$. (The prices of corporate bonds contain such information as well; see Duffie and Singleton (1999) and Jarrow and Turnbull (1995).) Thus, modeling for basket default swaps primarily deals with combining information about marginal default probabilities observed in market prices with a mechanism for capturing dependence between defaults.

2.4 *Pricing collateralized debt obligations*

A collateralized debt obligation or CDO is a collection of notes backed by a portfolio of bonds or other credit instruments. Each note carries with

it promised payments that are financed by coupon payments received from the underlying bond portfolio. Defaults of the underlying bonds reduce the payments received by the CDO and – through a set of allocation rules – the payments made to the holders of the notes issued by the CDO.

The notes are grouped into tranches of different levels of seniority, with senior tranches protected from defaults of the underlying bond by junior tranches. A CDO can create highly rated senior notes from a portfolio of low quality credits through the rules used to allocate cashflows from the underlying assets to the tranches. (In a synthetic CDO, the underlying assets are replaced by credit default swaps.)

Consider, for simplicity, the case of a CDO with a single payment date. Each tranche of the CDO is defined by upper and lower loss limits u and l, with $u = \infty$ for the most senior tranche and $l = 0$ for the most junior (equity) tranche. Let L denote the loss suffered by the portfolio underlying the CDO up to the payment date. Then the loss experienced by a tranche with limits u and l is

$$\min\{u, \max\{L - l, 0\}\}. \tag{4}$$

Thus, the tranche suffers no losses unless the portfolio loss L exceeds the lower limit l, and the tranche is wiped out if the portfolio's loss exceeds u. The equity tranche has lower limit $l = 0$ and absorbs the initial losses. Losses exceeding the upper limit of the equity tranche are absorbed by the next tranche and so on. In this way, more senior tranches are protected from losses by more junior tranches and can therefore achieve higher credit ratings than the underlying debt instruments.

In the case of a single payment date, the value of a tranche is the difference between the present value of the promised payment and the expected discounted value of the tranche loss (4). Thus, the key to valuing the tranche is determining the (risk-adjusted) distribution of the portfolio loss L.

In practice, the notes backed by a CDO make regular coupon payments. However, a note with multiple payments may be viewed as a portfolio of single-payment notes and then priced as a linear combination of the prices of these single-payment notes. Thus, valuing a tranche with regular payments reduces to finding the loss L at each coupon date and then finding the expected discounted value of (4) for each coupon date. For more on CDOs and other credit derivatives, see Duffie and Singleton (2003) and Schonbucher (2003).

3 Models of dependence

In this section, we describe mechanisms used to model dependence among the default times $\tau_1, \ldots, \tau_m$ and among the default indicators $Y_1, \ldots, Y_m$. We discuss structural models, stochastic intensity models, copula models (particularly the widely used Gaussian copula) and a mixed Poisson model.

3.1 Structural models

A structural model of default is one that describes how the inner workings of a firm's finances lead to default. Starting with Merton (1974), the general outline of these models goes as follows. Through a basic accounting identity, the value of a firm's equity and the value of its debt sum to equal the value of the assets owned by the firm. Equity carries limited liability and therefore cannot have negative value. Should the value of the firm's assets fall below the value of the debt, the equityholders may simply surrender the firm to the bondholders. Thus, the equityholders have a put option on the firm's assets, and default occurs when they exercise this put.

In Merton's (1974) model, firm value $A(t)$ is described by a geometric Brownian motion

$$A(t) = A(0)\exp\bigl(\mu t + \sigma W(t)\bigr),$$

for some parameters $\mu, \sigma > 0$ and a standard Brownian motion W. The firm's debt matures at T with a face value of D, and no payments are due prior to T. The firm defaults at T if $A(T) < D$. This occurs with probability

$$\begin{aligned} P\bigl(A(T) < D\bigr) &= P\bigl(\mu T + \sigma W(T) < \log\bigl(D/A(0)\bigr)\bigr) \\ &= \Phi\left(\frac{\log(D/A(0)) - \mu T}{\sigma\sqrt{T}}\right), \end{aligned} \tag{5}$$

with Φ the cumulative normal distribution.

This model has been generalized in several ways. One line of work has tried to capture more features of a firm's finances; another has sought to add realism to the dynamics of A. In the model of Black and Cox (1976), default occurs the first time firm value drops belows a boundary (rather than at a fixed maturity date T), so the time to default is a first-passage time. In Leland (1994) and Leland and Toft (1996), this boundary is chosen by the equityholders to maximize the value of equity. Models with dynamics for firm value that go beyond geometric Brownian motion include Chen and Kou (2005), Hilberink and Rogers (2002), Kijima and Suzuki (2001) and Linetsky (2006). Duffie and Lando (2001) and Giesecke (2004) link structural models with intensity-based models by limiting the information about the firm available to the market.

A small step (at least in principle) takes us from a structural model of default of a single firm to a model of dependence of default times across multiple firms. For m firms, we may specify a multivariate process $(A_1(t), \ldots, A_m(t))$ of firm values and then posit that the default of firm i is triggered by a decline in A_i in accordance with any of the existing single-firm models. This approach transfers the problem of specifying dependence between default times to the ostensibly simpler problem of specifying dependence between firm values: firm value is more directly related to economic, geographic and industry factors. Though conceptually appealing, this approach is not directly applied in practice because firm values are not observable and because no fully satisfactory

structural model of a single firm's default is yet available. Nevertheless, the structural perspective provides an important conceptual framework on which to build simpler models.

3.2 Copula models

A copula function is simply a way of describing the dependence structure in a multivariate distribution. If random variables $X_1, \ldots, X_m$ have joint distribution function F, meaning that

$$P(X_1 \leqslant x_1, \ldots, X_m \leqslant x_m) = F(x_1, \ldots, x_m),$$

for any $x_1, \ldots, x_m$, then there exists a function $C : [0, 1]^m \to [0, 1]$ for which

$$F(x_1, \ldots, x_m) = C\big(F_1(x_1), \ldots, F_m(x_m)\big), \tag{6}$$

for all $x_1, \ldots, x_m$, where F_i is the marginal distribution of X_i, $i = 1, \ldots, m$. The function C (which is essentially unique) is the copula function associated with the multivariate distribution F. The representation (6) distinguishes purely marginal features of F (the F_i) from the dependence structure, which is entirely determined by C. See Embrechts et al. (2000), Li (2000), and Nelsen (1999) for background on copulas.

The marginal distributions of default times are often taken to be roughly known through the hazard rate representation in (1) and the link between hazard rates and credit spreads. In building a model of the *joint* distribution of default times, it is therefore natural to approach the problem through the copula representation (6).

The Gaussian copula is particularly convenient because it is completely summarized by a correlation matrix Σ. Write Φ_Σ for the m-dimensional normal distribution with correlation matrix Σ in which all marginals have zero mean and unit variance. Write Φ for the one-dimensional standard normal distribution. Then the normal copula function C_Σ associated with correlation matrix Σ satisfies

$$\Phi_\Sigma(x_1, \ldots, x_m) = C_\Sigma\big(\Phi(x_1), \ldots, \Phi(x_m)\big).$$

Thus,

$$C_\Sigma(u_1, \ldots, u_m) = \Phi_\Sigma\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m)\big), \tag{7}$$

for any $u_i \in (0, 1)$, $i = 1, \ldots, m$. The function C_Σ extracts the dependence structure from the multivariate normal distribution for use with other marginal distributions. The Gaussian copula model for default times specifies

$$P(\tau_1 \leqslant t_1, \ldots, \tau_m \leqslant t_m) = C_\Sigma\big(F_1(t_1), \ldots, F_m(t_m)\big), \tag{8}$$

where F_i is the marginal distribution of the time to default for obligor i, $i = 1, \ldots, m$.

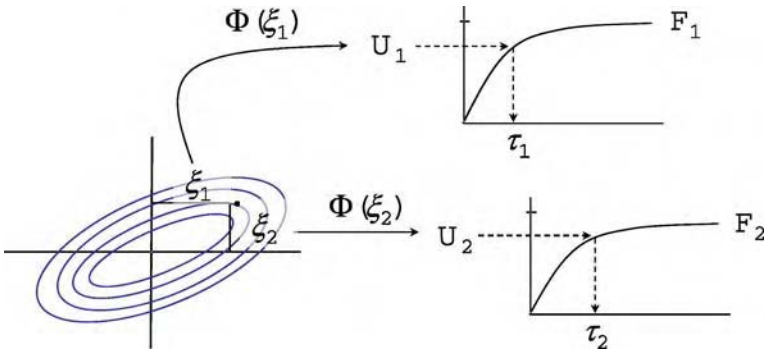


Fig. 2. Bivariate illustration of Gaussian copula construction of default times.

The interpretation of this specification is clarified by the mechanism used to simulate default times under this model. Figure 2 gives a bivariate illustration. One would start by finding an $m \times m$ matrix A for which $AA^\top = \Sigma$, for example through Cholesky factorization. Next, one would generate independent standard normal random variables $Z_1, \ldots, Z_m$ and set $\xi = AZ$, with $Z = (Z_1, \ldots, Z_m)^\top$. The vector ξ now has distribution $N(0, \Sigma)$. In Figure 2, the ellipses illustrate the contour lines of the bivariate normal distribution of (ξ_1, ξ_2). Next, set $U_i = \Phi(\xi_i)$, $i = 1, \ldots, m$; each U_i is uniformly distributed between 0 and 1, but the U_i are clearly not independent – they retain the dependence of the Gaussian copula. The final step is to set $\tau_i = F_i^{-1}(U_i)$, $i = 1, \ldots, m$, where F_i is the marginal distribution of τ_i. If (1) holds with $h_i > 0$, this means solving for τ_i in the equation

$$1 - \exp\left(-\int_0^{\tau_i} h_i(s)\,\mathrm{d}s\right) = U_i.$$

This algorithm ensures that each τ_i has the correct marginal distribution and imposes a joint distribution on $\tau_1, \ldots, \tau_m$ that is completely specified by Σ.

A similar construction defines the Gaussian copula for the default indicators $Y_1, \ldots, Y_m$ as well (as in Gupton et al., 1997). Indeed, if we simply set $Y_i = \mathbf{1}\{\tau_i \leqslant T\}$, $i = 1, \ldots, m$, then the default indicators inherit the Gaussian copula from the default times. But we can also construct the default indicators directly (without generating the default times) by setting

$$Y_i = \mathbf{1}\{\xi_i \leqslant x_i\}, \quad x_i = \Phi^{-1}(p_i), \quad i = 1, \ldots, m, \tag{9}$$

with $p_i = P(\tau_i \leqslant T)$, the marginal default probability for the ith obligor. This is illustrated in Fig. 3. Observe that this construction is consistent with the Merton (1974) default mechanism (5) if we set

$$\xi_i = \frac{\log(A_i(T)/A_i(0)) - \mu_i T}{\sigma_i \sqrt{T}}, \qquad x_i = \frac{\log(D_i/A_i(0)) - \mu_i T}{\sigma_i \sqrt{T}},$$

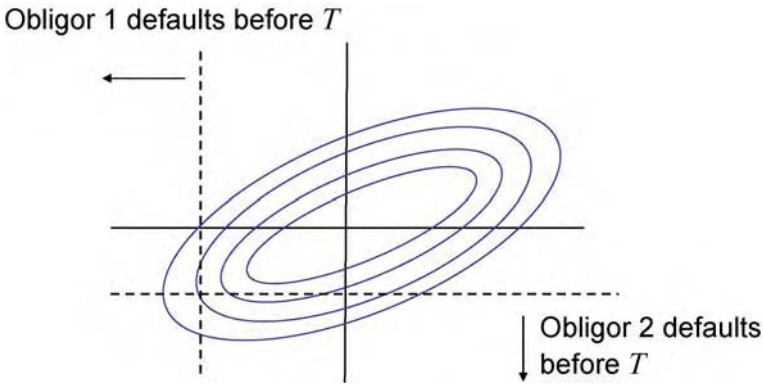


Fig. 3. Bivariate illustration of Gaussian copula construction of default indicators.

the subscript i indicating parameters of (5) associated with the ith firm.

In this setting, we assume that the losses $V_1, \ldots, V_m$ are independent of each other and of everything else. But we could bring these into the Gaussian copula as well by replacing $Y_k V_k$ with a more general loss random variable X_k having a known marginal distribution and then linking $X_1, \ldots, X_m$ through the Gaussian copula.

The popularity of the Gaussian copula model is due at least as much to its simplicity as to its empirical validity. In particular, the fact that its dependence structure is summarized through a correlation matrix allows a parsimonious representation and also provides a vehicle for translating correlations in equity returns (for which a great deal of data is available) to dependence between default times (for which much less data is available).

The underlying correlation matrix is often assumed (or estimated) to have a factor structure, meaning that we can write

$$\Sigma = AA^\top + B^2,$$

with B an $m \times m$ diagonal matrix and A an $m \times d$ matrix and $d \ll m$. In this case, we can write

$$\xi_k = a_{k1} Z_1 + \cdots + a_{kd} Z_d + b_k \epsilon_k, \quad k = 1, \ldots, m, \tag{10}$$

where

- $Z_1, \ldots, Z_d$ are independent standard normal random variables (the factors);
- the a_{kj} (the factor loadings) are the elements of the matrix A (the loading matrix);
- the ϵ_k (representing idiosyncratic risks) are standard normal random variables, independent of each other and of the factors;
- and $b_k = \sqrt{1 - a_{k1}^2 - \cdots - a_{kd}^2}$, so that ξ_k has unit variance.

The resulting $\xi_1, \ldots, \xi_m$ have the multivariate normal distribution $N(0, \Sigma)$. A feature of the factor representation is that the ξ_k (and thus the τ_k and the Y_k) are conditionally independent given $Z_1, \ldots, Z_d$. The factor structure is sometimes derived from economic factors (associated with industry sectors or geographic regions, for example) affecting default probabilities.

In pricing applications, a single-factor ($d = 1$) structure is sometimes assumed and, indeed, the coefficients a_{k1} all taken to be identical. In this case, all the off-diagonal elements of Σ are equal to a^2, where a is the common value of all the coefficients a_{k1}. This gives rise to the notion of an *implied* correlation for a credit derivative such as an nth-to-default swap: the implied correlation is the value of a^2 (if one exists) that equates the prices in the model to the price in the market. The market prices of CDOs typically produce different implied correlations for different tranches, indicating that the Gaussian copula (or at least the single-factor Gaussian copula with constant loading) cannot fully describe the joint distribution of default times. Nevertheless, implied correlations and the Gaussian copula remain a standard (and convenient) way to compare, price and hedge credit derivatives.

The specification in (7)–(8) can be extended to other copulas as well. For example, a t-copula is given by

$$C_{\Sigma,\nu}(u_1, \ldots, u_m) = F_{\Sigma,\nu}\big(F_\nu^{-1}(u_1), \ldots, F_\nu^{-1}(u_m)\big), \tag{11}$$

where $F_{\Sigma,\nu}$ is the multivariate t distribution with ν degrees of freedom and "correlation" matrix Σ, and F_ν is the univariate t distribution with ν degrees of freedom. The multivariate t density is given by

$$f_{\Sigma,\nu}(x) \propto \big(1 + x^\top \Sigma^{-1} x\big)^{-(m+\nu)}, \quad x \in \mathbb{R}^m;$$

Σ is the correlation matrix if $\nu > 2$ but otherwise the second-order moments of the distribution are infinite.

The Gaussian copula is a limiting case of the t copula with infinite degrees of freedom. But the cases $\nu < \infty$ and $\nu = \infty$ exhibit a qualitative difference in the dependence they introduce between extreme events. Except in the case of perfect correlation, the Gaussian copula exhibits zero *extreme tail dependence*, whereas t copulas have positive extreme tail dependence. Roughly speaking, this means that a pair of random variables linked through a Gaussian copula become independent when we condition on an extreme value of one of the two. In contrast, they remain dependent when linked through a t copula, the strength of the extreme tail dependence increasing as ν decreases; cf. Embrechts et al. (2000). Mashal and Zeevi (2003) find that this makes the t copula a better description of market returns. Hull and White (2004) suggest a "double t copula" for CDO pricing. Kalemanova et al. (2007) find that the multivariate normal inverse Gaussian distribution has features similar to the multivariate t while offering greater tractability. A different copula is constructed from the normal inverse Gaussian distribution in Guegan and Houdain (2005).

3.3 Mixed Poisson model

An alternative way of introducing dependence uses a mixed Poisson model, as in the CreditRisk$^+$ framework developed by Credit Suisse First Boston (Wilde, 1997). Although this model could also be cast in the copula framework, it is sufficiently distinctive to merit separate consideration.

The mixed Poisson model achieves greater tractability than, e.g., the Gaussian copula model but requires an important approximation from the outset: each default time τ_k is replaced with a (conditionally) Poisson process of times

$$\tau_k^1 = \tau_k < \tau_k^2 < \tau_k^3 < \cdots. \tag{12}$$

Only the first of these is meaningful, but if the intensity for the Poisson process is small, the probability of observing more than one arrival within the interval $[0, T]$ of interest may be small enough to be neglected. Similarly, in this setting the default indicators get replaced with (conditionally) Poisson random variables counting the number of arrivals in $[0, T]$. A Poisson random variable with a very small mean is nearly an indicator variable because the probability that the Poisson variable takes a value other than 0 or 1 is small.

We describe the time points (12) as conditionally Poisson because they become the epochs of a Poisson process only when we condition on a set of underlying random variables $\Gamma_1, \ldots, \Gamma_d$. We interpret these as underlying risk factors (much as the Z_i in (10)), but we require that they be mutually independent and positive; in fact, we will later specialize to the case in which each Γ_i has a gamma distribution.

Conditional on these random variables, each Y_k has a Poisson distribution with mean R_k,

$$R_k = a_{k0} + a_{k1}\Gamma_1 + \cdots + a_{kd}\Gamma_d, \tag{13}$$

for some positive coefficients $a_{k0}, \ldots, a_{kd}$. Thus, each Y_k may be viewed as a Poisson random variable with a random mean — a mixed Poisson random variable. We normalize $\Gamma_1, \ldots, \Gamma_d$ to have mean 1 and variances $\sigma_1^2, \ldots, \sigma_d^2$.

The times in (12) are similarly the points of a mixed Poisson process — a Poisson process with randomized intensity. Because Y_k counts the number of arrivals in $[0, T]$, the arrival rate for (12) consistent with (13) is R_k/T.

Applications of mixed Poisson models have long history; see, e.g., the discussion in Johnson et al. (1993, Section 8.3.2). Using gamma random variables for the mixing variables leads to some tractability and allows calculation of the distribution of L through numerical inversion of its probability generating function, as we will see in Section 5.3. To help fix ideas, we briefly describe simulation of the model.

In each replication, we first generate the common risk factors $\Gamma_j \geqslant 0$, $j = 1, \ldots, d$, independently. For example, we will choose Γ_j to have the gamma distribution with shape parameter α_j and scale parameter β_j, $j = 1, \ldots, d$,

with

$$\alpha_j = \frac{1}{\sigma_j^2}, \qquad \beta_j = \sigma_j^2, \quad j = 1, \ldots, d.$$

This gives Γ_j mean 1 and variance σ_j^2. Then we evaluate R_k as in (13) and draw Y_k from the Poisson distribution with mean R_k.

In the original CreditRisk$^+$ model, the losses upon default are fixed amounts v_k and must be integer valued. Each loss amount v_k is then interpreted as a multiple of a base amount and all losses must be rounded to such a multiple. This can be extended to allow random (integer-valued) losses V_k independent of each other and of the Y_k. The total portfolio loss is given by (2).

4 Conditional loss distributions

A general approach to calculating the loss distribution in a credit portfolio (and also to pricing credit derivatives tied to a portfolio) in the Gaussian copula and similar models proceeds by calculating the loss distribution conditional on a set of underlying factors that make defaults independent and then integrating over the distribution of the underlying factors to get the unconditional loss distribution. In this section, we discuss methods for carrying out the inner part of this calculation – finding the conditional loss distribution when conditioning makes defaults independent.

Consider, for example, the Gaussian copula with the factor structure in (10). Defaults become independent once we condition on the vector $Z = (Z_1, \ldots, Z_d)^\top$ of underlying factors. Thus, in calculating the conditional loss distribution

$$P(L \leqslant x \mid Z) = P(Y_1 V_1 + \cdots + Y_m V_m \mid Z)$$

we are calculating the distribution of the sum of independent random variables. The conditional default probability for obligor k is given by

$$p_k(Z) = P(Y_k = 1 \mid Z) = P(\xi_k \leqslant x_k \mid Z) = \Phi\left(\frac{\Phi^{-1}(p_k) - a_k Z}{b_k}\right), \tag{14}$$

with $a_k = (a_{k1}, \ldots, a_{kd})$ the row vector of coefficients in (10). We have thus far assumed that the losses upon default, V_k, are independent of each other and of the factors; so long as the losses $Y_1 V_1, \ldots, Y_m V_m$ are conditionally independent given Z, the conditional distribution of L is given by the distribution of the sum of independent random variables.

In order to keep our discussion generic and to lighten notation, for the rest of this section we will simply take $Y_1, V_1, \ldots, Y_m, V_m$ to be independent and omit explicit reference to the conditioning that makes them independent. We write p_k (rather than, e.g., $p_k(Z)$) for the kth obligor's default probability.

The remainder of this section thus deals with methods for calculating the distribution of sums of independent random variables.

4.1 Recursive convolution

Andersen et al. (2003) develop a recursive method to compute the loss distribution for independent (or conditionally independent) defaults; a similar idea is used in Hull and White (2004). In its simplest form, this method is applicable when the losses upon default are constants $v_1, \ldots, v_m$ and in fact integers. Losses are thus measured as multiples of a basic unit. The loss distribution admits a probability mass function π on the integers from 0 to $\ell_{\max} = v_1 + \cdots + v_m$.

The recursion produces a sequence of probability mass functions $\pi_1, \ldots, \pi_m$ in which π_k records the distribution of losses from just the first k obligors and $\pi_m = \pi$. The first of these distributions is simply

$$\pi_1(x) = \begin{cases} p_1, & x = v_1; \\ 0, & \text{otherwise.} \end{cases}$$

When we add the kth obligor, the portfolio loss increases by v_k with probability p_k remains unchanged with probability $1 - p_k$. Thus, for $k = 2, \ldots, m$, and all integers x from 0 to $v_1 + \cdots + v_k$,

$$\pi_k(x) = p_k \pi_{k-1}(x - v_k) + (1 - p_k)\pi_{k-1}(x). \tag{15}$$

The method generalizes to a model in which the default of the kth obligor results in possible losses $v_{k1}, \ldots, v_{kn_k}$ with probabilities $q_{k1}, \ldots, q_{kn_k}$. We now initialize by setting

$$\pi_1(v_{1j}) = p_1 q_{1j}, \quad j = 1, \ldots, n_1,$$

and $\pi_1(x) = 0$ for all other x. At the kth step, the recursion gives

$$\pi_k(x) = p_k \sum_{j=1}^{n_k} q_{kj} \pi_{k-1}(x - v_{kj}) + (1 - p_k)\pi_{k-1}(x). \tag{16}$$

As observed in Andersen et al. (2003), this method lends itself to the calculation of sensitivities. Consider, for example, the effect of small changes in the default probabilities (which, in turn, may result from small changes in credit spreads through (1)). Suppose the probabilities p_k depend smoothly on a parameter θ and use a dot to indicate derivatives with respect to this parameter. Then

$$\dot{\pi}_1(v_{1j}, \theta) = \dot{p}_1(\theta) q_{1j}, \quad j = 1, \ldots, n_1,$$

and $\dot{\pi}_1(x, \theta) = 0$ for all other x. Differentiating both sides of (16) yields

$$\dot{\pi}_k(x, \theta) = \dot{p}_k(\theta) \left(\sum_{j=1}^{n_k} q_{kj} \pi_{k-1}(x - v_{kj}, \theta) - \pi_{k-1}(x, \theta) \right) \tag{17}$$

$$+ p_k(\theta) \sum_{j=1}^{n_k} q_{kj} \dot{\pi}_{k-1}(x - v_{kj}, \theta) + \big(1 - p_k(\theta)\big) \dot{\pi}_{k-1}(x, \theta). \tag{18}$$

4.2 Transform inversion

The recursive algorithm in Section 4.2 computes a sequence of explicit convolutions, at the kth step convolving the loss distribution for a subportfolio of obligors $1, \ldots, k$ with the loss distribution of the $(k+1)$st obligor. An alternative to computing convolutions is to compute either a Fourier or Laplace transform of the distribution of a sum of random variables and then numerically invert the transform. This approach applies whether the individual obligor losses are fixed or stochastic.

The distribution of the loss L depends on the distribution of the individual losses $V_1, \ldots, V_m$. The distribution of each V_i may be characterized through its cumulant generating function

$$\Lambda_i(\theta) = \log \mathsf{E}\big[\exp(\theta V_i)\big], \quad \theta \in \mathbb{R}, \tag{19}$$

the logarithm of the moment generating function of V_i. Each V_i is a random fraction of a largest possible loss upon the default of obligor i, so it is reasonable to take the V_i to be bounded random variables. This is more than sufficient to ensure that $\Lambda_i(\theta)$ is finite for all real θ.

Recall that in this section we take the obligors to be independent of each other. (More commonly, they are only conditionally independent given a set of underlying factors, in which case the loss distributions we compute should be interpreted as conditional loss distributions.) As a consequence of independence we find that the moment generating function of the loss L is

$$\begin{aligned} \phi_L(\theta) &= \mathsf{E}\left[\exp\left(\theta \sum_{k=1}^{m} Y_k V_k\right)\right] \\ &= \prod_{k=1}^{m} \mathsf{E}\big[\exp(\theta Y_k V_k)\big] \\ &= \prod_{k=1}^{m} \big(1 + p_k\big[\exp\big(\Lambda_i(\theta)\big) - 1\big]\big). \end{aligned}$$

The Laplace transform of the distribution of L is the function $s \mapsto \phi_L(-s)$ and the characteristic function of L is the function $\omega \mapsto \phi_L(i\omega)$, with $i = \sqrt{-1}$.

We apply Laplace transform inversion to the calculate the tail of the loss distribution,

$$1 - F_L(x) = P(L > x).$$

But first we review Laplace transform inversion more generically. Thus, let f be a function on $[0, \infty)$ with Laplace transform

$$\hat{f}(s) = \int_0^\infty e^{-st} f(t)\, \mathrm{d}t.$$

The original function f can be recovered from the transform $\hat{f}$ through the Bromwich inversion integral

$$f(t) = \frac{1}{2\pi i} \int_{b-i\infty}^{b+i\infty} e^{st} \hat{f}(s)\, \mathrm{d}s, \tag{20}$$

where b is any real number such that all singularities of $\hat{f}$ (viewed as a function of the complex plane) lie to the left of the line from $b - i\infty$ to $b + i\infty$. Noting that the Laplace transform of $1 - F_L$ is given by

$$\frac{1 - \phi_L(-s)}{s},$$

with ϕ_L the moment generating function of L, as above, the Bromwich inversion integral gives

$$P(L > t) = \frac{1}{2\pi i} \int_{b-i\infty}^{b+i\infty} e^{st} \left(\frac{1 - \phi_L(-s)}{s} \right) \mathrm{d}s.$$

Moreover, this is valid for all values of b, because $[1-\phi_L(-s)]/s$ is well-defined and finite for all values of s; the limit as $s \to 0$ is finite because ϕ_L is differentiable at the origin.

Abate and Whitt (1995) develop numerical procedures for Laplace transform inversion that allow the user to control roundoff and approximation error. Their method rewrites (20) as

$$f(t) = \frac{2e^{bt}}{\pi} \int_0^\infty \mathrm{Re}\big(\hat{f}(b + iu)\big) \cos(ut)\, \mathrm{d}u,$$

where Re gives the real part of a complex number. They then use a trapezoidal rule with step size h to approximate the integral as

$$f(t) \approx f_h(t) = \frac{he^{bt}}{\pi} \mathrm{Re}\big(\hat{f}(b)\big) + \frac{2he^{bt}}{\pi} \sum_{k=1}^{\infty} \mathrm{Re}\big(\hat{f}(b + ikh)\big) \cos(kht).$$

In practice, the infinite sum must be truncated and Abate and Whitt (Abate and Whitt, 1995) bound the truncation error. They also apply Euler summation to accelerate convergence. To use the Abate–Whitt method to calculate the

loss distribution, set

$$\hat{f}(s) = \frac{1 - \phi_L(-s)}{s}.$$

Glasserman and Ruiz-Mata (2006) report experiments with this approach in the Gaussian copula model. Gregory and Laurent (2003) suggest the use of Fourier inversion but do not specify a particular inversion technique.

4.3 Saddlepoint approximation

Saddlepoint approximation uses transforms to approximate the distribution of a random variable without numerical integration. This approach is most conveniently formulated in terms of the cumulant generating function of the random variable, which is the logarithm of its moment generating function. The cumulant generating function of L is thus

$$\psi_L(\theta) = \log \phi_L(\theta) = \sum_{k=1}^{m} \log\bigl(1 + p_k\bigl[\exp\bigl(\Lambda_k(\theta)\bigr) - 1\bigr]\bigr). \tag{21}$$

This function (like the cumulant generating function of any nonnegative random variable) is increasing, convex, infinitely differentiable and takes the value zero at $\theta = 0$. The saddlepoint associated with a loss level $x > 0$ is the root θ_x of the equation

$$\psi'_L(\theta_x) = x. \tag{22}$$

The derivative of ψ_L is an increasing function with $\psi'_L(\theta) \to 0$ as $\theta \to -\infty$ and $\psi'_L(\theta) \to \infty$ as $\theta \to \infty$, so this equation has exactly one root for each $x > 0$.

The significance of the saddlepoint θ_x is best explained through a more general examination of the cumulant generating function ψ_L. The derivatives of ψ_L at the origin give the cumulants of L, and the first cumulant is just the mean; indeed,

$$\psi'_L(0) = \sum_{k=1}^{m} p_k \Lambda'_k(0) = \sum_{k=1}^{m} p_k \mathsf{E}[V_k] = \mathsf{E}[L].$$

Consider, now, the case of fixed losses $V_k \equiv v_k$, $k = 1, \ldots, m$. In this case,

$$\psi'_L(\theta) = \sum_{k=1}^{m} \frac{p_k e^{\theta v_k}}{1 + p_k(e^{\theta v_k} - 1)} v_k; \tag{23}$$

this may be interpreted as the expected loss when the original default probabilities p_k are replaced with

$$p_{k,\theta} = \frac{p_k e^{\theta v_k}}{1 + p_k(e^{\theta v_k} - 1)}. \tag{24}$$

In the case of stochastic V_k, we have

$$\psi'_L(\theta) = \sum_{k=1}^{m} \frac{p_k e^{\Lambda_k(\theta)}}{1 + p_k(e^{\Lambda_k(\theta)} - 1)} \Lambda'_k(\theta).$$

This corresponds to the expected loss when the default probabilities p_k are replaced with

$$p_{k,\theta} = \frac{p_k e^{\Lambda_k(\theta)}}{1 + p_k(e^{\Lambda_k(\theta)} - 1)}$$

and the expected loss $\mathsf{E}[V_k]$ is replaced with $\Lambda'_k(\theta)$. The original expected loss $\mathsf{E}[V_k]$ coincides with $\Lambda'_k(0)$, because Λ_k is the cumulant generating function of V_k.

Thus, each value of θ determines a modified set of default probabilities and a modified loss given default for each obligor. For each θ, $\psi'_L(\theta)$ is the expected portfolio loss under the portfolio parameters determined by θ. The saddlepoint θ_x in (22) is the value of θ that shifts the expected loss to x.

Saddlepoint approximations can be derived as approximations to contour integrals that arise in inverting transforms to obtain probability distributions. They can also often be interpreted as the result of shifting (or exponentially twisting or tilting) probabilities by θ_x and then applying a normal approximation. A standard (cf. Jensen (1995)) saddlepoint approximation, for $x > \mathsf{E}[L]$, gives

$$P(L > x) \approx \exp\bigl(-\theta_x x + \psi_L(\theta_x) + (1/2)\psi''_L(\theta_x)\bigr)\Phi\bigl(-\theta_x\sqrt{\psi''_L(\theta_x)}\,\bigr),$$

where Φ is the cumulative normal distribution. This is used in Martin et al. (2001a, 2001b). The closely related Lugannani–Rice approximation is

$$P(L > x) \approx 1 - \Phi\bigl(r(x)\bigr) + \phi\bigl(r(x)\bigr)\left(\frac{1}{\lambda(x)} - \frac{1}{r(x)}\right),$$

with

$$r(x) = \sqrt{2\bigl(\theta_x x - \psi_L(\theta_x)\bigr)} \quad \text{and} \quad \lambda(x) = \theta_x\sqrt{\psi''_L(\theta_x)}.$$

Gordy (2002) applies this to the CreditRisk$^+$ model. A modification of the Lugannani–Rice formula is

$$P(L > x) \approx 1 - \Phi\left(r(x) + \frac{1}{r(x)}\log\frac{\lambda}{r(x)}\right). \tag{25}$$

This modification has the advantage that it always produces a value between 0 and 1. All of these approximations give quite accurate results, particularly when the loss threshold x is large, when the number of obligors m is large, and the obligors losses $Y_k V_k$, $k = 1, \ldots, m$, are similarly distributed.

5 Unconditional loss distributions

5.1 Factor models

In Section 4, we discussed methods for computing loss distributions when obligors are independent. The motivation for considering this lies in its application to factor models in which obligors are *conditionally* independent, given a set of underlying factors. The methods of Section 4 then apply to the calculation of conditional loss distributions, and finding the unconditional loss distribution is a matter of integrating out the factors.

A simple model might specify two sets of parameters – corresponding to a high-default regime and a low-default regime, with independent obligors in each regime – and then model portfolio losses using a mixture of the two sets of parameters. In this case, the underlying "factor" is the regime and the unconditional loss distribution may be computed as a mixture of the two conditional loss distributions. This case is straightforward, as is any finite mixture of independent-obligor models.

In the Gaussian copula model, using (10), obligors become independent conditional on the normally distributed factors $Z_1, \ldots, Z_d$. Thus, finding the loss distribution in this model entails integrating out the effect of the factors. If the number of factors d is large, this almost invariably requires Monte Carlo simulation.

A straightforward simulation algorithm repeats the following steps for each replication:

- generate factors $Z = (Z_1, \ldots, Z_d)$;
- generate latent variables $\xi_1, \ldots, \xi_m$, using (10);
- evaluate default indicators $Y_1, \ldots, Y_m$ as in (9) and, for each k for which $Y_k = 1$, generate the loss given default V_k;
- calculate portfolio loss $L = Y_1V_1 + Y_2V_2 + \cdots + Y_mV_m$.

From multiple independent replications of the portfolio loss, one can estimate the loss distribution and any quantities (such as value-at-risk) derived from the distribution.

An alternative approach uses simulation only for the factors and then applies a deterministic numerical method (as in Section 4) to compute or approximate the conditional loss distribution. This yields the following steps for each replication, for a given set of loss thresholds $x_1, \ldots, x_n$:

- generate factors $Z = (Z_1, \ldots, Z_d)$;
- calculate or approximate $P(L > x_i|Z)$, $i = 1, \ldots, n$.

Averaging over independent draws of the factors yields estimates of the unconditional loss probabilities $P(L > x_i)$, $i = 1, \ldots, n$.

The second step in this algorithm may be carried out through convolution, transform inversion, or saddlepoint approximation, as described in Section 4.

A rough but fast alternative is to apply a normal approximation to the conditional distribution of L given Z, using the first two moments of the conditional loss, an idea explored in Shelton (2004) and Zheng (2004).

Glasserman and Ruiz-Mata (2006) compare the computational efficiency of ordinary Monte Carlo simulation with methods that combine simulation for the factors with the techniques of Section 4. Their comparison invites the following remarks:

(i) Numerical transform inversion and saddlepoint approximation involve some error in the calculation of conditional loss probabilities, so the estimates they produce when combined with Monte Carlo simulation of the factors are biased. The investigation in (Glasserman and Ruiz-Mata, 2006) therefore compares the mean square errors obtained in a fixed amount of computing time as a figure of merit.
(ii) Because each replication using convolution, transform inversion or saddlepoint approximation takes longer than each replication using ordinary simulation, these methods complete fewer (indeed, far fewer) replications in a fixed amount of computing time.
(iii) The recursive convolution method computes the full conditional loss distribution on each replication, whereas transform inversion and saddlepoint approximation must in practice be limited to a smaller number of loss thresholds $x_1, \ldots, x_n$.
(iv) Using ordinary Monte Carlo, the same replications can be used to estimate loss probabilities $P(L > x)$ at a large number of loss thresholds with little additional effort.
(v) Using the saddlepoint approximation requires solving for multiple saddlepoint parameters θ_{x_i} on each replication.
(vi) The computing time required using recursive convolution grows quickly with the number of obligors m.

As a consequence of these properties, Glasserman and Ruiz-Mata (2006) find that, with the total computing time held fixed, ordinary Monte Carlo often produces a smaller mean square error than methods that combine simulation with the techniques of Section 4, except at large loss levels.

When the number of factors is small – up to three or four, say – one can replace Monte Carlo sampling of the factors with a numerical integration procedure, such as Gaussian quadrature. For a moderate number of dimensions – five to twenty, say – quasi-Monte Carlo sampling of the factors may be attractive. As an alternative to integrating out the factors, Glasserman (2004) proposes approximations that use a single "most important" value of the factors, as in the classical Laplace approximation for integrals. Shelton (2004) and Zheng (2004) approximate the conditional loss distribution using a normal distribution by matching two moments and then compute the unconditional distribution through numerical integration, assuming a small number of factors.

5.2 Homogeneous approximation

Full computation of the unconditional loss distribution in the Gaussian copula model can be somewhat time consuming; an alternative is to approximate the unconditional distribution using a simpler family of distributions.

One approach is to look at the limiting distribution for a large portfolio of homogeneous obligors. Suppose, therefore, that the default indicators Y_k, $k = 1, 2, \ldots$, are conditionally i.i.d. given a single factor Z and that the losses V_k, $k = 1, 2, \ldots$ are i.i.d. and independent of Z. The latent variables have the form

$$\xi_k = \rho Z + \sqrt{1-\rho^2}\epsilon_k, \quad k = 1, 2, \ldots, \tag{26}$$

with $0 < \rho < 1$. Thus, all obligors are identical and conditionally independent. The portfolio loss $L \equiv L_m = Y_1V_1 + Y_2V_2 + \cdots + Y_mV_m$ satisfies

$$L_m/m \to \mathsf{E}[Y_1V_1 \mid Z] = P(Y_1 = 1 \mid Z)v \equiv p(Z)v, \quad v = \mathsf{E}[V_1],$$

with probability 1. Also, if the unconditional default probability is $P(Y = 1) = p$, then

$$p(Z) = P\big(\xi_k < \Phi^{-1}(p) \mid Z\big) = \Phi\left(\frac{\Phi^{-1}(p) - \rho Z}{\sqrt{1-\rho^2}}\right). \tag{27}$$

Thus, for $0 < q < 1$,

$$\begin{aligned} &P(L_m/m \leqslant vq) \to P\big(p(Z) \leqslant q\big) \\ &\quad = G_{p,\rho}(q) = \Phi\left(\frac{\sqrt{1-\rho^2}\Phi^{-1}(q) - \Phi^{-1}(p)}{\rho}\right). \end{aligned} \tag{28}$$

This limiting loss distribution was identified by Vasicek (1991) and is applied in regulatory standards (Basel Committee on Bank Supervision, 2003), among many other places. With the expected loss given default (per obligor) fixed at v, this defines a two-parameter family of distributions that can be used to approximate loss distributions in more general portfolios by setting

$$P(L_m \leqslant x) \approx G_{p,\rho}\left(x \Big/ \sum_{k=1}^{m} v_k\right), \tag{29}$$

for some p and ρ.

The parameter p is the mean of the distributions $G_{p,\rho}$,

$$\int_0^\infty u \, \mathrm{d}G_{p,\rho}(u) = p.$$

In approximating the loss distribution of an arbitrary portfolio with default probabilities p_k and conditional expected losses v_k, $k = 1, \ldots, m$, it is therefore natural to use p to match the mean; this is accomplished by setting

$$p = \sum_{k=1}^{m} p_k v_k \Big/ \sum_{k=1}^{m} v_k.$$

The parameter ρ can be chosen to match the variance of the distribution. Glasserman (2004) shows that the variance of $G_{p,\rho}$ is

$$\sigma_{p,\rho}^2 = 2\Phi_2\big(0, \Phi^{-1}(p); -\sqrt{1-\rho^2}/\sqrt{2}\big) - p^2, \tag{30}$$

where $\Phi_2(\cdot, \cdot; r)$ is the standard bivariate normal distribution with correlation r. The variance of the actual loss distribution is

$$\sigma_L^2 = \mathsf{Var}[L] = \sum_{k=1}^{m} \mathsf{Var}[Y_k V_k] + 2\sum_{k=1}^{m-1} \sum_{j=k+1}^{m} v_k v_j \mathsf{Cov}[Y_k, Y_j],$$

with

$$\mathsf{Var}[Y_k V_k] = p_k \mathsf{E}\big[V_k^2\big] - p_k^2 v_k^2$$

and, for $j \neq k$,

$$\begin{aligned}\mathsf{Cov}[Y_k, Y_j] &= P\big(\xi_k \leqslant \Phi^{-1}(p_k), \xi_j \leqslant \Phi^{-1}(p_j)\big) - p_k p_j \\ &= \Phi_2\big(\Phi^{-1}(p_k), \Phi^{-1}(p_j); a_k a_j^\top\big) - p_k p_j,\end{aligned}$$

in light of (10). To match the variance of the actual and approximating distributions in (29), we therefore want ρ to satisfy $(\sum_{k=1}^m v_k)^2 \sigma_{p,\rho}^2 = \sigma_L^2$.

The effectiveness of this two-moment approximation is discussed in Glasserman (2004) along with other methods for choosing ρ in using $G_{p,\rho}$ as an approximating distribution. Gordy (2004) develops "granularity adjustments" for risk measures computed from the limiting distribution $G_{p,\rho}$ as corrections to (29) for finite m. Kalemanova et al. (2007) derive limiting homogeneous approximations under the double t copula of Hull and White (2004) and under the normal inverse Gaussian copula. A different approach to approximating the unconditional loss distribution, using a correlation expansion, is developed in Glasserman and Suchintabandid (2007).

5.3 Mixed Poisson model

Next, we consider the unconditional loss distribution in the mixed Poisson model discussed in Section 3.3. Although the distribution itself is not available in closed form, its Laplace transform (and cumulant generating function) can be made explicit. This contrasts with the Gaussian copula model for which only the conditional loss distribution has an explicit Laplace transform.

If N is a Poisson random variable with mean λ, then its moment generating function is

$$\mathsf{E}\big[\exp(\theta N)\big] = \exp\big(\lambda\big(e^{\theta} - 1\big)\big),$$

for all θ. In the mixed Poisson model of Section 3.3, each Y_k is conditionally Poisson, given the factor random variables $\Gamma_1, \ldots, \Gamma_d$, with conditional mean R_k in (13). Thus,

$$\begin{aligned}\mathsf{E}\big[\exp(\theta Y_k) \mid \Gamma_1, \ldots, \Gamma_d\big] &= \exp\big(R_k\big(e^{\theta} - 1\big)\big) \\ &= \exp\left(\sum_{j=0}^{d} a_{kj}\Gamma_j\big(e^{\theta} - 1\big)\right),\end{aligned}$$

with $\Gamma_0 \equiv 1$. Moreover, $Y_1, \ldots, Y_m$ are conditionally independent given $\Gamma_1, \ldots, \Gamma_d$, so, for the portfolio loss $L = Y_1 v_1 + \cdots + Y_m c_m$, we have

$$\begin{aligned}\mathsf{E}\big[\exp(\theta L) \mid \Gamma_1, \ldots, \Gamma_d\big] &= \prod_{k=1}^{m} \mathsf{E}\big[\exp(\theta v_k Y_k) \mid \Gamma_1, \ldots, \Gamma_d\big] \\ &= \exp\left(\sum_{k=1}^{m}\sum_{j=0}^{d} a_{kj}\Gamma_j\big(e^{v_k\theta} - 1\big)\right).\end{aligned}$$

To complete the calculation of the moment generating function (or the Laplace transform) of the L we suppose that Γ_j has cumulant generating function ψ_j, so that

$$\log \mathsf{E}\big[\exp(\alpha\Gamma_j)\big] = \psi_j(\alpha), \quad j = 1, \ldots, d,$$

and $\psi_0(\alpha) \equiv \alpha$. We suppose that $\psi_j(\alpha)$ is finite for some $\alpha > 0$. Then

$$\phi_L(\theta) \equiv \mathsf{E}\big[\exp(\theta L)\big] = \mathsf{E}\left[\exp\left(\sum_{j=0}^{d} \alpha_j\Gamma_j\right)\right] = \exp\left(\sum_{j=0}^{d} \psi_j(\alpha_j)\right), \quad (31)$$

with

$$\alpha_j = \sum_{k=1}^{m} a_{kj}\big(e^{v_k\theta} - 1\big), \quad j = 0, 1, \ldots, d.$$

Because $\psi_j(\alpha_j)$ is finite for all sufficiently small $\alpha_j > 0$, the moment generating function of L in (31) is finite for all sufficiently small $\theta > 0$.

In the particular case that Γ_j has a gamma distribution with mean 1 and variance σ_j^2, $j = 1, \ldots, d$, we have

$$\psi_j(\alpha) = -\frac{1}{\sigma_j^2}\log\big(1 - \sigma_j^2\alpha\big), \quad \alpha < 1/\sigma_j^2.$$

This case is used in the CreditRisk$^+$ framework (Wilde, 1997). It generalizes a classical model (see, e.g., Greenwood and Yule, 1920, Section IV) used in actuarial science with $d = 1$ (a single underlying factor). The tractability of this case rests, in part, on the fact that a gamma mixture of Poisson random variables has a negative binomial distribution, as explained in Johnson et al. (1993, p. 328).

From the moment generating function ϕ_L in (31), one can evaluate the Laplace transform $s \mapsto \phi_L(-s)$ and the characteristic function $\omega \mapsto \phi_L(\sqrt{-1}\omega)$; either of these can then be inverted numerically to find the distribution of L. One may alternatively apply a saddlepoint approximation to the cumulant generating function $\log \phi_L$, as in Gordy (2002). If all $v_1, \ldots, v_m$ are integers (interpreted as multiples of a basic loss amount), then L is integer-valued with probability generating function $z \mapsto \phi_L(\log z)$. The Panjer (1981) recursion is used in the actuarial literature for this case and applied in Wilde (1997). Haaf et al. (2005) develop a numerically stable alternative procedure for transform inversion. Haaf and Tasche (2002) discuss the calculation of marginal risk contributions.

6 Importance sampling

In Section 4, we discussed methods for computing the loss distribution in a portfolio of independent obligors. These methods also allow calculation of *conditional* loss distributions when obligors become independent conditional on a set of underlying factors. As noted in Section 5.1, ordinary Monte Carlo simulation is often the most effective way to calculate the unconditional loss distribution in a model with multiple underlying factors. But ordinary Monte Carlo simulation is generally inefficient for the estimation of small probabilities associated with large losses, and these are often of primary importance for risk management.

In this section, we discuss the application of importance sampling for rare-event simulation in credit risk. In importance sampling, we change the distribution from which we generate outcomes in order to generate more observations of a rare event; we then weight each scenario to correct for the change in distribution. The appropriate weight to "unbias" the simulation is the likelihood ratio relating the original and new probabilities. The application of importance sampling to credit risk has been suggested in Avranitis and Gregory (2001), Joshi and Kainth (2004), Kalkbrener et al. (2004), Merino and Nyfeler (2004), Morokoff (2004), Glasserman and Li (2005), and Glasserman et al. (2005); we base our discussion on the last two of these references which include theoretical support for the methods they develop.

6.1 Independent obligors

We begin by considering the case of independent obligors. This is the simplest case to introduce and will be a building block in the case of dependent

obligors that become conditionally independent given a set of underlying factors.

Consider, therefore, the case of independent default indicators $Y_1, \ldots, Y_m$, with $P(Y_k = 1) = p_k$. To further simplify the setting, let losses given default be constants $v_1, \ldots, v_m$, so that the portfolio loss is $L = Y_1 v_1 + \cdots + Y_m v_m$. Our objective is to estimate $P(L > x)$ for large x; from precise estimates of the tail of L, we can then estimate VaR and other risk measures. See Glynn (1996) for an analysis of importance sampling for quantile estimation.

In order to generate more scenarios with large losses, we might increase each default probability p_k to some new value q_k and then simulate using the higher default probabilities. The resulting likelihood ratio associated with any outcome $Y_1, \ldots, Y_m$ is the ratio of the probabilities of these outcomes under the original and new default probabilities, which is given by

$$\ell = \prod_{k=1}^{m} \left(\frac{p_k}{q_k}\right)^{Y_k} \left(\frac{1-p_k}{1-q_k}\right)^{1-Y_k}. \tag{32}$$

We obtain an importance sampling estimate of $P(L > x)$ by generating the default indicators Y_k from the new probabilities, evaluating the portfolio loss $L = Y_1 v_1 + \cdots + Y_m v_m$ and returning the estimator $\mathbf{1}\{L > x\}\ell$, with $\mathbf{1}\{\cdot\}$ the indicator of the event in braces. Multiplying by the likelihood ratio ℓ makes this an unbiased estimator for all choices of q_k, $0 < q_k < 1$, $k = 1, \ldots, m$.

Although increasing the default probabilities produces more replications with large losses, it does not guarantee a reduction in variance; indeed, a poor choice for the q_k can easily produce an increase in variance. Some guidance in the choice of probabilities is available from the simulation literature.

The problem of rare-event simulation for sums of independent random variables has been studied extensively, as have generalizations of this problem. It follows, for example, from Sadowsky and Bucklew (1990) that a particularly effective approach to this problem is to apply an *exponential twist* to the distributions of the increments. Applying an exponential twist means multiplying the value of a probability density or mass function at every point x by a factor $\exp(\theta x)$, for some parameter θ, and then normalizing so that the scaled density or mass function has total probability 1.

In our setting, the increments are the random variables $Y_k v_k$. Each takes two values, v_k and 0, with probabilities p_k and $1 - p_k$. Applying an exponential twist with parameter θ thus means multiplying the first probability by $\exp(\theta v_k)$, multiplying the second probability by $\exp(\theta 0) = 1$, and then normalizing so that the two values sum to 1. This produces exactly the values $p_{k,\theta}$ in (24). In other words, the exponentially twisted probabilities are the probabilities used in the saddlepoint approximation.

With $q_k = p_{k,\theta}$, the likelihood ratio in (32) reduces, through algebraic simplification, to

$$\ell = \exp\bigl(-\theta L + \psi_L(\theta)\bigr),$$

where ψ_L the cumulant generating function in (21). (With $V_k \equiv v_k$, we have $\Lambda_k(\theta) = v_k\theta$ in (21).) From this expression, we see that the likelihood ratio is decreasing in L for all $\theta > 0$. This is attractive because the key to reducing variance is making the likelihood ratio small on the event $\{L > x\}$. More explicitly, if we write E_θ to denote expectation using the default probabilities $p_{k,\theta}$, then the second moment of the importance sampling estimator is

$$\mathsf{E}_\theta\big[\mathbf{1}\{L > x\}\ell^2\big] = \mathsf{E}\big[\mathbf{1}\{L > x\}\ell\big];$$

thus, we reduce variance by making ℓ small when $L > x$.

We would like to choose the parameter θ to minimize variance. Doing so exactly is generally infeasible, but we can approximate the second moment of the estimator by the upper bound

$$\begin{aligned}\mathsf{E}\big[\mathbf{1}\{L > x\}\ell\big] &= \mathsf{E}\big[\mathbf{1}\{L > x\}\exp\big(-\theta L + \psi_L(\theta)\big)\big] \\ &\leqslant \exp\big(-\theta x + \psi_L(\theta)\big)\end{aligned}$$

and then minimize this upper bound. Because ψ_L is convex, the upper bound is minimized at the value $\theta = \theta_x$ solving $\psi_L'(\theta_x) = x$. This is precisely the saddlepoint in (22). Recall from (22) and (23) that the probabilities p_{k,θ_x} defined by the saddlepoint have the property that they set the expected loss equal to x. Glasserman and Li (2005) show that using this change of probability produces variance reduction that is asymptotically optimal as the portfolio size m and the threshold x grow in fixed proportion.

The foregoing discussion extends, with minor modification, to the case of random losses V_k. In addition to twisting the default probabilities, we now twist the loss amounts. If, for example, V_k has density f_k, then applying an exponential twist means replacing f_k with

$$f_{k,\theta}(v) = e^{\theta v - \Lambda_k(\theta)} f_k(v),$$

where, as in Section 4.3, Λ_k is the cumulant generating function of V_k. The resulting likelihood ratio is

$$\begin{aligned}\ell &= \prod_{k=1}^m \left(\frac{p_k}{p_{k,\theta}}\right)^{Y_k} \left(\frac{1-p_k}{1-p_{k,\theta}}\right)^{1-Y_k} \prod_{k=1}^m \frac{f_k(V_k)}{f_{k,\theta}(V_k)} \\ &= \exp\big(-\theta L + \psi_L(\theta)\big).\end{aligned}$$

Thus, the likelihood ratio has the same form as before, but the fact that the V_k are random is reflected in the functions Λ_k appearing in ψ_L, as in (21). The argument used in the case of fixed losses again leads us to take $\theta = \theta_x$, the saddlepoint associated with loss level x.

6.2 Importance sampling in the Gaussian copula

A model with independent obligors is a useful place to begin a discussion of importance sampling, but from a practical perspective it is necessary to consider a model with dependent defaults. We consider the case of the Gaussian

copula model of dependence, in which defaults become independent conditional on a set of underlying factors $Z = (Z_1, \ldots, Z_d)^\top$.

If the kth obligor has unconditional default probability p_k, then its conditional default probability, given Z, is $p_k(Z)$, as defined in (14). A straightforward extension of the importance sampling technique described above for the independent case would apply the same ideas conditional on Z. This means replacing the $p_k(Z)$ with

$$p_{k,\theta}(Z) = \frac{p_k(Z)e^{\Lambda_k(\theta)}}{1 + p_k(Z)(e^{\Lambda_k(\theta)} - 1)},$$

and replacing the loss densities f_k with $f_{k,\theta}$. The conditional likelihood ratio for this conditional change of distribution is

$$\ell(Z) = \exp\bigl(-\theta L + \psi_L(\theta, Z)\bigr).$$

Observe that ψ_L now depends on Z because the conditional default probabilities $p_k(Z)$ replace the original default probabilities p_k in (21).

Furthermore, the saddlepoint θ_x itself depends on Z because it solves the equation

$$\frac{\partial}{\partial\theta}\psi_L(\theta_x, Z) = x,$$

which depends on Z. For some values of Z, it is possible to get $\theta_x(Z) < 0$; this happens when $\mathsf{E}[L|Z] > x$. Rather than twist with a negative parameter (which would decrease the conditional default probabilities), it is preferable to replace $\theta_x(Z)$ with 0. The two cases can be combined by twisting by $\theta_x^+(Z) = \max\{\theta_x(Z), 0\}$. Using this parameter, the conditional likelihood ratio becomes

$$\ell(Z) = \exp\bigl(-\theta_x^+(Z)L + \psi_L\bigl(\theta_x^+(Z), Z\bigr)\bigr).$$

This approach parallels (conditional on Z) the technique described above for the case of independent obligors. But whereas this approach produces sizeable variance reduction in the independent case, Glasserman and Li (2005) prove that, by itself, it cannot be very effective once defaults are only conditionally independent. This phenomenon may be explained by noting that large losses occur in two ways in the Gaussian copula – either because Z falls near the origin and many obligors default by chance, despite having small default probabilities; or because Z falls in a region far from the origin in which many of the conditional default probabilities $p_k(Z)$ are large. The conditional exponential twist addresses only the first of these two vehicles: it increases the conditional default probabilities to generate more defaults conditional on Z. But the second vehicle is the more important one: large losses are generally more likely to occur because of exceptional outcomes of a relatively small number of factors, rather than because of exceptional outcomes of a large number of default indicators.

This second mechanism can be incorporated into the importance sampling procedure by changing the distribution of the factors in order to generate more

outcomes in which the conditional default probabilities $p_k(Z)$ are large. Recall that Z has a standard multivariate normal distribution. The simplest change of distribution to consider is to move the mean of Z from the origin to some other point μ. The ratio of the $N(0, I)$ and $N(\mu, I)$ densities at Z is $\exp(-\mu^\top Z + \mu^\top \mu/2)$; if we first change the mean of Z and then twist the conditional default probabilities given Z, the overall likelihood ratio is

$$\exp\bigl(-\theta_x^+(Z)L + \psi_L\bigl(\theta_x^+(Z), Z\bigr)\bigr)\exp\bigl(-\mu^\top Z + \tfrac{1}{2}\mu^\top \mu\bigr).$$

It remains to choose the new mean μ. Glasserman and Li (2005) suggest that μ should be chosen as the value of z maximizing

$$P(L > x \mid Z = z)\exp\bigl(-z^\top z/2\bigr). \tag{33}$$

The second factor is proportional to the standard normal density, so a maximizing z may be interpreted as the "most likely" factor outcome leading to a large loss. In the same spirit, one may seek to choose z to

$$\text{minimize } z^\top z \quad \text{subject to} \quad \mathsf{E}[L \mid Z = z] \geqslant x. \tag{34}$$

Solving (33) is generally impractical, so Glasserman and Li (2005) replace the conditional loss probability with the upper bound

$$P(L > x \mid Z = z) \leqslant \exp\bigl(-\theta_x^+(z)x + \psi_L\bigl(\theta_x^+(z), z\bigr)\bigr) \equiv \exp\bigl(F_x(z)\bigr).$$

By using this upper bound as an approximation in (33), they arrive at the optimization problem

$$\max_z\bigl\{F_x(z) - \tfrac{1}{2}z^\top z\bigr\}$$

and they choose the new mean $\mu = \mu_*$ to solve this problem. Problem (34) may be recast as

$$\text{minimize } z^\top z \quad \text{subject to} \quad F_x(z) \geqslant 0,$$

because of the equivalences

$$\begin{aligned} F_x(z) \geqslant 0 \quad &\Leftrightarrow \quad F_x(z) = 0 \quad \Leftrightarrow \quad \theta_x^+(z) = 0 \quad \Leftrightarrow \quad \theta_x(z) \leqslant 0 \\ &\Leftrightarrow \quad \mathsf{E}[L|Z = z] \geqslant x. \end{aligned}$$

Figure 4 illustrates the solutions μ_* and z_* to these optimization problems for a single-factor homogeneous portfolio with $p_k \equiv 0.02$, $v_k \equiv 1$, $m = 100$ and $x = 10$. Also, we take the factor loading $\rho = -0.3$ (in the notation of (26)), so that the default probabilities (and $F_x(z)$) are increasing functions of z. The function F_x^o in the figure is what we get if we replace θ_x^+ with θ_x in the definition of F_x.

Glasserman and Li (2005) establish asymptotic optimality results using μ_* in single-factor homogeneous models. For multifactor models, Glasserman et al. (2005) use a mixture of mean shifts, using a procedure that generalizes

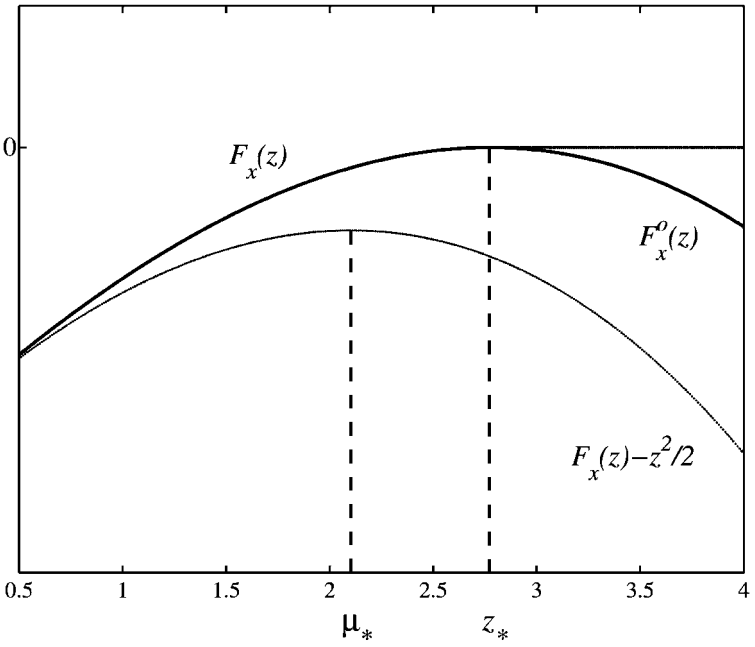


Fig. 4. Graphs of the functions $F_x(z)$, $F_x^o(z)$, and $F_x(z) - z^2/2$ for a single-factor, homogeneous portfolio with $p_k \equiv 0.02$, $v_k \equiv 1$, $\rho \equiv -0.3$, $m = 100$, and $x = 10$. The point μ_* maximizes $F_x(z) - z^2/2$, and z_* is the smallest point for which $F_x(z) = 0$.

(34) for selecting the new means, and establish asymptotic optimality. Numerical results in these papers indicate that the variance reduction achieved is around a factor of 50 for values of x with $P(L > x) \approx 1\%$, and that the variance reduction achieved generally increases with the rarity of the event. Thus, these techniques specifically address the region in which precise estimates are difficult to obtain using ordinary Monte Carlo simulation. Related importance sampling techniques are developed for the t copula and other models in Bassamboo et al. (2006), Kang and Shahabuddin (2005) and Kostadinov (2005).

7 Summary

This chapter provides an overview of some of the primary models and computational methods used in measuring portfolio credit risk and pricing credit derivatives. The models used in these applications combine marginal information – the default probability or the distribution of the time to default for a single obligor – with a mechanism that captures dependence between obligors. We have given particular attention to the Gaussian copula model of dependence and a mixed Poisson model.

A common feature of the Gaussian copula and mixed Poisson models is that defaults become independent conditional on a set of underlying factors. This has important implications for the computational procedures used with these models. In the Gaussian copula model, the losses from default are given by a sum of independent random variables, conditional on the underlying factors. Thus, any method for computing or approximating the distribution of a sum of independent random variables can be applied to find the conditional loss distribution. These methods include recursive convolution, transform inversion, and saddlepoint approximation. Finding the unconditional loss distribution then requires integrating over the distribution of the factors. The greater tractability of the mixed Poisson model allows direct calculation of the unconditional distribution, regardless of the number of underlying factors.

Monte Carlo simulation can be combined with deterministic numerical methods. In the Gaussian copula model, deterministic methods can be used to compute the conditional loss distribution given a set of underlying factors, with simulation then used to integrate over the distribution of the factors. But simulation can also be used to calculate the unconditional loss distribution directly. Ordinary Monte Carlo can be quite effective in estimating the unconditional loss distribution, except at rarely observed large loss levels. Importance sampling techniques designed for rare-event simulation can be very effective in improving Monte Carlo estimates of the upper tail of the loss distribution.

References

Abate, J., Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* 7, 36–48.

Andersen, L., Basu, S., Sidenius, J. (2003). All your hedges in one basket. *Risk* 16 (November), 67–72.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9, 203–228.

Avranitis, A., Gregory, J. (2001). *Credit: The Complete Guide to Pricing, Hedging and Risk Management*. Risk Books, London.

Basel Committee on Bank Supervision (2003). *The New Basel Capital Accord*. Third consultative document, http://www.bis.org.

Bassamboo, A., Juneja, S., Zeevi, A. (2006). Portfolio credit risk with extremal dependence. *Operations Research*, in press.

Black, F., Cox, J. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* 31, 351–367.

Chen, N., Kou, S.G. (2005). Credit spreads, optimal capital structure and implied volatility with endogenous default and jump risk. *Mathematical Finance*, in press.

Das, S., Duffie, D., Kapadia, N., Saita, L. (2005). Common failings: how corporate defaults are correlated. *Working paper*.

Duffie, D., Garleanu, N. (2001). Risk and valuation of collateralized debt obligations. *Financial Analysts Journal* 57, 41–59.

Duffie, D., Lando, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica* 69, 633–664.

Duffie, D., Singleton, K. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies* 12, 687–720.

Duffie, D., Singleton, K. (2003). *Credit Risk: Pricing, Measurement, and Management*. Princeton Univ. Press, Princeton, NJ.

Denault, M. (2001). Coherent allocation of risk capital. *Journal of Risk* 4, 1–34.

Embrechts, P., McNeil, A., Straumann, D. (2000). Correlation and dependence properties in risk management: Properties and pitfalls. In: Embrechts, P. (Ed.), *Extremes and Integrated Risk Management*. Risk Books, London, pp. 71–76.

Garman, M. (1999). Taking VAR to pieces. In: *Hedging with Trees*. Risk Publications, London.

Giesecke, K. (2004). Correlated default with incomplete information. *Journal of Banking and Finance* 28, 1521–1545.

Giesecke, K., Tomecek, P. (2005). Dependent events and changes of time. *Working paper*. School of ORIE, Cornell University..

Glasserman, P. (2004). Tail approximations for portfolio credit risk. *Journal of Derivatives* 12 (Winter), 24–42.

Glasserman, P. (2005). Measuring marginal risk contributions in credit portfolios. *Journal of Computational Finance* 9, 1–41.

Glasserman, P., Li, J. (2005). Importance sampling for portfolio credit risk. *Management Science* 51, 1643–1656.

Glasserman, P., Ruiz-Mata, J. (2006). A comparison of approximation techniques for portfolio credit risk. *Journal of Credit Risk* 2, 33–66.

Glasserman, P., Kang, W., Shahabuddin, P. (2005). Fast simulation of multifactor portfolio credit risk. *Operations Research*, in press.

Glasserman, P., Suchintabandid, S. (2007). Correlation expansions for CDO prising. *Journal of Banking and Finance* 31, 1375–1398.

Glynn, P.W. (1996). Importance sampling for Monte Carlo estimation of quantiles. In: *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the Second St. Petersburg Workshop on Simulation*. St. Petersburg Univ. Press, St. Petersburg, Russia, pp. 180–185.

Gordy, M.B. (2002). Saddlepoint approximation of CreditRisk$^+$. *Journal of Banking and Finance* 26, 1335–1353.

Gordy, M.B. (2004). Granularity adjustment in portfolio credit risk measurement. In: Szegö, G. (Ed.), *Risk Measures for the 21st Century*. Wiley, New Yok.

Gourieroux, C., Laurent, J.P., Scaillet, O. (2000). Sensitivity analysis of values at risk. *Journal of Empirical Finance* 7, 225–245.

Greenwood, M., Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* 83, 255–279.

Gregory, J., Laurent, J.-P. (2003). I will survive. *Risk* 16 (June), 103–107.

Guegan, D., Houdain, J. (2005). Collateralized debt obligations pricing and factor models: A new methodology using normal inverse Gaussian distributions. *Working paper*. Ecole Normale Supérieure, Cachan, France.

Gupton, G., Finger, C., Bhatia, M. (1997). *CreditMetrics Technical Document*. J.P. Morgan & Co., New York.

Haaf, H., Tasche, D. (2002). Calculating value-at-risk contributions in CreditRisk$^+$. *GARP Risk Review* 7, 43–47.

Haaf, H., Reiss, O., Schoenmakers, J. (2005). Numerically stable computation of CreditRisk$^+$. In: Gundlach, M., Lehrbass, F. (Eds.), *CreditRisk$^+$ in the Banking Industry*. Springer-Verlag, Berlin, pp. 67–76.

Hilberink, B., Rogers, L.C.G. (2002). Optimal capital structure and endogenous default. *Finance and Stochastics* 6, 237–263.

Hull, J., White, A. (2004). Valuation of a CDO and nth to default CDS without Monte Carlo. *Journal of Derivatives* 12 (Winter), 8–23.

Jarrow, R.A., Turnbull, S.M. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50, 53–85.

Jarrow, R.A., Lando, D., Turnbull, S.M. (1997). A Markov model for the term structure of credit risk spreads. *Review of Financial Studies* 10, 481–523.

Jensen, J.L. (1995). *Saddlepoint Approximations*. Oxford Univ. Press, Oxford, UK.
Johnson, N.L., Kotz, S., Kemp, A.W. (1993). *Univariate Discrete Distributions*, second ed. Wiley, New York.
Joshi, M., Kainth, D. (2004). Rapid computation of prices and deltas of nth to default swaps in the Li model. *Quantitative Finance* 4, 266–275.
Kalkbrener, M. (2005). An axiomatic approach to capital allocation. *Mathematical Finance* 15, 425–437.
Kalkbrener, M., Lotter, H., Overbeck, L. (2004). Sensible and efficient capital allocation for credit portfolios. *Risk* 17, S19–S24.
Kalemanova, A., Schmid, B., Werner, R. (2007). The normal inverse Gaussian distribution for synthetic CDO pricing. *Journal od Derivatives*, in press.
Kang, W., Shahabuddin, P. (2005). Simulation of multifactor portfolio credit risk in the *t*-copula model. In: *Proceedings of the Winter Simulation Conference*, pp. 1859–1868.
Kijima, M., Suzuki, T. (2001). A jump-diffusion model for pricing corporate debt securities in a complex capital structure. *Quantitative Finance* 1, 611–620.
Kostadinov, K. (2005). Tail approximations for portfolio credit risk with heavy-tailed risk factors. *Journal of Risk* 8, 81–107.
Kurth, A., Tasche, D. (2003). Contributions to credit risk. *Risk* March, 84–88.
Leland, H.E. (1994). Corporate debt value, bond covenants and optimal capital structure. *Journal of Finance* 49, 1213–1252.
Leland, H.E., Toft, K.B. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *Journal of Finance* 51, 987–1019.
Li, D. (2000). On default correlation: A copula function approach. *Journal of Fixed Income* 9, 43–54.
Linetsky, V. (2006). Pricing equity derivatives subject to bankruptcy. *Mathematical Finance* 16, 255–282.
Litterman, R. (1999). Hot spots and hedges. In: *Hedging with Trees*. Risk Publications, London.
Martin, R., Thompson, K., Browne, C. (2001a). Taking to the saddle. *Risk* 14 (June), 91–94.
Martin, R., Thompson, K., Browne, C. (2001b). Who contributes and how much. *Risk* 14 (August), 99–103.
Mashal, R., Zeevi, A. (2003). Beyond correlation: Extreme co-movements between financial assets. *Working paper*. Columbia Business School.
Merino, S., Nyfeler, M.A. (2004). Applying importance sampling for estimating coherent credit risk contributions. *Quantitative Finance* 4, 199–207.
Merton, R.C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449–470.
Morokoff, W.J. (2004). An importance sampling method for portfolios of credit risky assets. In: *Proceedings of the Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1668–1676.
Nelsen, R.B. (1999). *An Introduction to Copulas*. Springer-Verlag, New York.
Panjer, H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* 12, 22–26.
Sadowsky, J.S., Bucklew, J.A. (1990). On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Transactions on Information Theory* 36, 579–588.
Schonbucher, P. (2003). *Credit Derivatives Pricing Models: Models, Pricing and Implementation*. Wiley, Chichester, UK.
Shelton, D. (2004). Back to normal – Proxy integration: A fast accurate method for CDO and CDO-squared pricing. *Citigroup Global Structured Credit Research*, London.
Tasche, D. (1999). Risk contributions and performance measurement. *Working paper*. TU München, Munich, Germany.
Vasicek, O. (1991). Limiting Loan Loss Probability Distribution. KMV Corporation, San Francisco, CA.
Wilde, T. (1997). *CreditRisk$^+$: A Credit Risk Management Framework*. Credit Suisse First Boston, London.
Zheng, H. (2004). Heterogeneous portfolio approximations and applications. Mathematics Department, Imperial College, London.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15011-8

Chapter 11

Valuation of Basket Credit Derivatives in the Credit Migrations Environment*

Tomasz R. Bielecki

Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
E-mail: bielecki@iit.edu

Stéphane Crépey

Département de Mathématiques, Université d'Évry Val d'Essonne, 91025 Évry cedex, France

Monique Jeanblanc

Département de Mathématiques, Université d'Évry Val d'Essonne, 91025 Évry cedex, France

Marek Rutkowski

School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia
and
Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warszawa, Poland

Abstract

The goal of this work is to present a methodology aimed at valuation and hedging of basket credit derivatives, as well as of portfolios of credits/loans, in the context of several possible credit ratings of underlying credit instruments. The methodology is based on a specific Markovian model of a financial market.

* The research of T.R. Bielecki was supported by NSF Grants 0202851 and 0604789, and Moody's Corporation Grant 5-55411.
The research of S. Crepey was supported by Zeliade.
The research of M. Jeanblanc was supported by Zeliade and by Moody's Corporation Grant 5-55411.
The research of M. Rutkowski was supported by the 2005 Faculty Research Grant PS06987.

1 Introduction

The goal of this work is to present some methods and results related to the valuation and hedging of basket credit derivatives, as well as of portfolios of credits/loans, in the context of several possible credit ratings of underlying credit instruments. Thus, we are concerned with modeling dependent credit migrations and, in particular, with modeling dependent defaults. On the mathematical level, we are concerned with modeling dependence between random times and with evaluation of functionals of (dependent) random times; more generally, we are concerned with modeling dependence between random processes and with evaluation of functionals of (dependent) random processes. Modeling of dependent defaults and credit migrations was considered by several authors, who proposed several alternative approaches to this important issue. Since the detailed analysis of these methods is beyond the scope of this text, let us only mention that they can be roughly classified as follows:

- modeling correlated defaults in a static framework using copulae (Hull and White, 2001; Gregory and Laurent, 2004),
- modeling dependent defaults in a "dynamic" framework using copulae (Schönbucher and Schubert, 2001; Laurent and Gregory, 2003; Giesecke, 2004),
- dynamic modeling of credit migrations and dependent defaults via proxies (Douady and Jeanblanc, 2002; Chen and Filipovic (2003, 2005); Albanese et al., 2003; Albanese and Chen, 2004a, 2004b),
- factor approach ((Jarrow and Yu, 2001; Yu, 2007; Frey and Backhaus, 2004); Bielecki and Rutkowski, 2002b, 2003),
- modeling dependent defaults using mixture models (Frey and McNeil, 2003; Schmock and Seiler, 2002),
- modeling of the joint dynamics of credit ratings by a voter process (Giesecke and Weber, 2002),
- modeling dependent defaults by a dynamic approximation (Davis and Esparragoza, 2004).

The classification above is rather arbitrary and by no means exhaustive. In the next section, we shall briefly comment on some of the above-mentioned approaches. In this work, we propose a fairly general Markovian model that, in principle, nests several models previously studied in the literature. In particular, this model covers jump-diffusion dynamics, as well as some classes of Lévy processes. On the other hand, our model allows for incorporating several credit names, and thus it is suitable when dealing with valuation of basket credit products (such as, basket credit default swaps or collateralized debt obligations) in the multiple credit ratings environment. Another practically important feature of the model put forward in this paper is that it refers to market observables only. In contrast to most other papers in this field, we carefully analyze the issue of preservation of the Markovian structure of the model under equivalent changes of probability measures.

1.1 Conditional expectations associated with credit derivatives

We present here a few comments on evaluation of functionals of random times related to financial applications, so to put into perspective the approach that we present in this paper. In order to smoothly present the main ideas we shall keep technical details to a minimum.

Suppose that the underlying probability space is $(\Omega, \mathcal{G}, \mathbb{P})$ endowed with some filtration $\mathbb{G}$ (see Section 2 for details). Let τ_l, $l = 1, 2, \ldots, L$ be a family of finite and strictly positive random times defined on this space. Let also real-valued random variables X and $\widetilde{X}$, as well as real-valued processes A (of finite variation) and Z be given. Next, consider an $\mathbb{R}^k_+$-valued random variable $\zeta := g(\tau_1, \tau_2, \ldots, \tau_L)$ where $g : \mathbb{R}^L_+ \to \mathbb{R}^k_+$ is some measurable function. In the context of valuation of credit derivatives, it is of interest to evaluate conditional expectations of the form

$$\mathbb{E}_{\mathbb{P}^\beta}\Big(\int_{]t,T]} \beta_u^{-1} \, \mathrm{d}D_u \mid \mathcal{G}_t \Big), \tag{1}$$

for some numeraire process β, where the *dividend process* D is given by the following generic formula:

$$D_t = \big(X\alpha_1(\zeta) + \widetilde{X}\alpha_2(\zeta)\big)\mathbb{1}_{\{t \geqslant T\}} + \int_{]0,t]} \alpha_3(u;\zeta) \, \mathrm{d}A_u + \int_{]0,t]} Z_u \, \mathrm{d}\alpha_4(u;\zeta),$$

where the specification of α_is depends on a particular application. The probability measure $\mathbb{P}^\beta$, equivalent to $\mathbb{P}$, is the martingale measure associated with a numeraire β (see Section 4.2 below). We shall now illustrate this general set-up with four examples. In each case, it is easy to identify the processes A, Z as well as the α_is.

Example 1.1 *(Defaultable Bond)*. We set $L = 1$ and $\tau = \tau_1$, and we interpret τ as a time of default of an issuer of a corporate bond (we set here $\zeta = \tau = \tau_1$). The face value of the bond (the promised payment) is a constant amount X that is paid to bondholder at maturity T, provided that there was no default by the time T. In addition, a coupon is paid continuously at the instantaneous rate c_t up to default time or bond's maturity, whichever comes first. In case default occurs by the time T, a recovery payment is paid, either as the lump sum $\widetilde{X}$ at bond's maturity, or as a time-dependent rebate Z_τ at the default time. In the former case, the dividend process of the bond equals

$$D_t = \big(X(1 - H_T) + \widetilde{X}H_T\big)\mathbb{1}_{\{t \geqslant T\}} + \int_{]0,t]} (1 - H_u)c_u \, \mathrm{d}u,$$

where $H_t = \mathbb{1}_{\{\tau \leqslant t\}}$, and in the latter case, we have that

$$D_t = X(1 - H_T)\mathbb{1}_{\{t \geqslant T\}} + \int_{]0,t]} (1 - H_u)c_u \, \mathrm{d}u + \int_{]0,t]} Z_u \, \mathrm{d}H_u.$$

Example 1.2 *(Credit Ratings Triggered Step-Up Corporate Bonds)*. These are corporate coupon bonds for which the coupon payment depends on the issuer's credit quality: the coupon payment increases when the credit quality of the issuer declines. In practice, for such bonds, credit quality is reflected in credit ratings assigned to the issuer by at least one credit ratings agency (such as Moody's-KMV, Fitch Ratings or Standard & Poor's). Let X_t stand for some indicator of credit quality at time t. Assume that t_i, $i = 1, 2, \ldots, n$ are coupon payment dates and let $c_i = c(X_{t_{i-1}})$ be the coupons ($t_0 = 0$). The dividend process associated with the step-up bond equals

$$D_t = X(1 - H_T)\mathbb{1}_{\{t \geqslant T\}} + \int_{]0,t]} (1 - H_u) \, \mathrm{d}A_u + \text{possible recovery payment}$$

where τ, X and H are as in the previous example, and $A_t = \sum_{t_i \leqslant t} c_i$.

Example 1.3 *(Default Payment Leg of a Collateralized Debt Obligation (CDO) Tranche)*. We consider a portfolio of L credit names. For each $l = 1, 2, \ldots, L$, the nominal payment is denoted by N_l, the corresponding default time by τ_l and the associated *loss given default* by $M_l = (1-\delta_l)N_l$, where δ_l is the *recovery rate* for the lth credit name. We set $H_t^l = \mathbb{1}_{\{\tau_l \leqslant t\}}$ for every $l = 1, 2, \ldots, L$, and $\zeta = (\tau_1, \tau_2, \ldots, \tau_L)$. Thus, the *cumulative loss process* equals

$$L_t(\zeta) = \sum_{l=1}^{L} M_l H_t^l.$$

Similarly as in Laurent and Gregory (2003), we consider a cumulative default payments process on the mezzanine tranche of the CDO:

$$M_t(\zeta) = \big(L_t(\zeta) - a\big)\mathbb{1}_{[a,b]}\big(L_t(\zeta)\big) + (b - a)\mathbb{1}_{]b,N]}\big(L_t(\zeta)\big),$$

where a, b are some thresholds such that $0 \leqslant a \leqslant b \leqslant N := \sum_{l=1}^{L} N_l$. If we assume that $M_0 = 0$ then the dividend process corresponding to the default payment leg on the mezzanine tranche of the CDO is $D_t = \int_{]0,t]} \mathrm{d}M_u(\zeta)$.

Example 1.4 *(Default Payment Leg of a kth-to-default CDS)*. Consider a portfolio of L reference defaultable bonds. For each defaultable bond, the notional amount is taken to be deterministic and denoted as N_l; the corresponding recovery rate δ_l is also deterministic. We suppose that the maturities of the bonds are U_l and the maturity of the swap is $T < \min\{U_1, U_2, \ldots, U_L\}$. Here, we set

$\zeta = (\tau_1, \tau_2, \ldots, \tau_L, \tau^{(k)})$, where $\tau^{(k)}$ is the kth order statistics of the collection $\{\tau_1, \tau_2, \ldots, \tau_L\}$.

A special case of the kth-to-default-swap is the case when the protection buyer is protected against only the last default (i.e. the kth default). In this case, the dividend process associated with the default payment leg is

$$D_t = (1 - \delta_{\iota^{(k)}}) N_{\iota^{(k)}} \mathbb{1}_{\{\tau^{(k)} \leqslant T\}} H_t^{(k)},$$

where $H_t^{(k)} = \mathbb{1}_{\{\tau^{(k)} \leqslant t\}}$ and $\iota^{(k)}$ stands for the identity of the kth defaulting credit name. This can be also written as $D_t = \int_{]0,t]} \mathrm{d}N_u(\zeta)$, where

$$N_t(\zeta) = \int_{]0,t]} \sum_{l=1}^{L} (1 - \delta_l) N_l \mathbb{1}_{\tau_l}(u)\, \mathrm{d}H_u^{(k)}.$$

1.2 Existing methods of modeling dependent defaults

It is apparent that in order to evaluate the expectation in (1) one needs to know, among other things, the conditional distribution of ζ given $\mathcal{G}_t$. This, in general, requires the knowledge of conditional dependence structure between random times $\tau_l, \tau_2, \ldots, \tau_L$, so that it is important to be able to appropriately model dependence between these random times. This is not an easy task, in general. Typically, the methodologies proposed in the literature so far handle well the task of evaluating the conditional expectation in (1) for $\zeta = \tau^{(1)} = \min\{\tau_1, \tau_2, \ldots, \tau_L\}$, which, in practical applications, corresponds to *first-to-default* or *first-to-change* type credit derivatives. However, they suffer from more or less serious limitations when it comes to credit derivatives involving subsequent defaults or changes in credit quality, and not just the first default or the first change, unless restrictive assumptions are made, such as conditional independence between the random times in question. In consequence, the existing methodologies would not handle well computation of expectation in (1) with process D as in Examples 1.3 and 1.4, unless restrictive assumptions are made about the model. Likewise, the existing methodologies can't typically handle modeling dependence between credit migrations, so that they can't cope with basket derivatives whose payoffs explicitly depend on changes in credit ratings of the reference names.

Arguably, the best known and the most widespread among practitioners is the copula approach (cf. Li, 2000; Schubert and Schönbucher, 2001; and Laurent and Gregory, 2003, for example). Although there are various versions of this approach, the unifying idea is to use a copula function so to model dependence between some auxiliary random variables, say $\upsilon_1, \upsilon_2, \ldots, \upsilon_L$, which are supposed to be related in some way to $\tau_l, \tau_2, \ldots, \tau_L$, and then to infer the dependence between the latter random variables from the dependence between the former.

It appears that the major deficiency of the copula approach, as it stands now, is its inability to compute certain important conditional distributions. Let us illustrate this point by means of a simple example. Suppose that $L = 2$ and consider the conditional probability $\mathbb{P}(\tau_2 > t + s \mid \mathcal{G}_t)$. Using the copula approach, one can typically compute the above probability (in terms of partial derivatives of the underlying copula function) on the set $\{\tau_1 = t_1\}$ for $t_1 \leqslant t$, but not on the set $\{\tau_1 \leqslant t_1\}$. This means, in particular, that the copula approach is not "Markovian," although this statement is rather vague without further qualifications. In addition, the copula approach, as it stands now, can't be applied to modeling dependence between changes in credit ratings, so that it can't be used in situations involving, for instance, baskets of corporate step-up bonds (cf. Example 1.2). In fact, this approach can't be applied to valuation and hedging of basket derivatives if one wants to account explicitly on credit ratings of the names in the basket. Modeling dependence between changes in credit ratings indeed requires modeling dependence between stochastic processes.

Another methodology that gained some popularity is a methodology of modeling dependence between random times in terms of some proxy processes, typically some Lévy processes (cf. Hull and White, 2001; Albanese et al., 2003; and Chen and Filipović, 2003, 2005, for example). The major problem with these approaches is that the proxy processes are latent processes whose states are unobservable virtual states. In addition, in this approach, when applied to modeling of credit quality, one can't model a succession of credit ratings, e.g., the joint evolution of the current and immediately preceding credit ratings (see Remark 2.1 (ii) below).

2 Notation and preliminary results

The underlying probability space containing all possible events over a finite time horizon is denoted by $(\Omega, \mathcal{G}, \mathbb{P})$, where $\mathbb{P}$ is a generic probability measure. Depending on the context, we shall consider various (mutually equivalent) probability measures on the space $(\Omega, \mathcal{G})$. The probability space $(\Omega, \mathcal{G}, \mathbb{P})$ is endowed with a filtration $\mathbb{G} = \widetilde{\mathbb{H}} \vee \mathbb{F}$, where the filtration $\widetilde{\mathbb{H}}$ carries the information about evolution of credit events, such as changes in credit ratings of respective credit names, and where $\mathbb{F}$ is some *reference filtration*. We shall be more specific about both filtrations later on; at this point, we only postulate that they both satisfy the so-called "usual conditions."

The credit events of fundamental interest to us are changes in credit ratings, in particular – the default event. The evolution of credit ratings can be modeled in terms of an appropriate stochastic process defined on $(\Omega, \mathcal{G}, \mathbb{P})$. Various approaches to the choice of this process have been already proposed in the literature. We shall focus here on the Markov approach, in the sense explained in Section 2.1.1 below.

2.1 Credit migrations

We consider L obligors (or credit names). We assume that current credit rating of the lth reference entity can be classified to one of K_l different rating categories. We let $\mathcal{K}_l = \{1, 2, \ldots, K_l\}$ to denote the set of such categories. However, without a loss of generality, we assume that $\mathcal{K}_l = \mathcal{K} := \{1, 2, \ldots, K\}$ for every $l = 1, 2, \ldots, L$. By convention, the category K corresponds to default.

Let $X^l, l = 1, 2, \ldots, L$ be some processes on $(\Omega, \mathcal{G}, \mathbb{P})$ with values in $\mathcal{K}$. A process X^l represents the evolution of credit ratings of the lth reference entity.

Let us write $X = (X^1, X^2, \ldots, X^L)$. The state space of X is $\mathcal{X} := \mathcal{K}^L$; the elements of $\mathcal{X}$ will be denoted by x. We call the process X the (joint) *migration process*. We assume that $X_0^l \neq K$ for every $l = 1, 2, \ldots, L$, and we define the *default time* τ_l of the lth reference entity by setting

$$\tau_l = \inf\{t > 0\colon\ X_t^l = K\} \tag{2}$$

with the usual convention that $\inf \emptyset = \infty$. We assume that the default state K is absorbing, so that for each name the default event can only occur once. Put another way, for each l the process X^l is stopped at τ_l. Since we are considering a continuous time market then, without loss of practical generality, we assume that simultaneous defaults are not allowed. Specifically, the equality $\mathbb{P}(\tau_{l'} = \tau_l) = 0$ will hold for every $l' \neq l$ in our model.

Remark 2.1. (i) In the special case when $K = 2$, only two categories are distinguished: pre-default ($j = 1$) and default ($j = 2$). We then have $X_t^l = H_t^l + 1$, where $H_t^l = \mathbb{1}_{\{\tau_l \leqslant t\}}$.

(ii) Each credit rating j may include a "history" of transitions. For example, j may be a two-dimensional quantity, say $j = (j', j'')$, where j' represents the current credit rating, whereas j'' represents the immediately preceding credit rating.

2.1.1 Markovian set-up

From now on, we set $\widetilde{\mathbb{H}} = \mathbb{F}^X$, so that the filtration $\widetilde{\mathbb{H}}$ is the natural filtration of the process X. Arguably, the most convenient set-up to work with is the one where the reference filtration $\mathbb{F}$ is the filtration $\mathbb{F}^Y$ generated by relevant (vector) factor process, say Y, and where the process (X, Y) is jointly Markov under $\mathbb{P}$ with respect to its natural filtration $\mathbb{G} = \mathbb{F}^X \vee \mathbb{F}^Y = \widetilde{\mathbb{H}} \vee \mathbb{F}$, so that we have, for every $0 \leqslant t \leqslant s$, $x \in \mathcal{X}$ and any set $\mathcal{Y}$ from the state space of Y,

$$\mathbb{P}(X_s = x, Y_s \in \mathcal{Y} \mid \mathcal{G}_t) = \mathbb{P}(X_s = x, Y_s \in \mathcal{Y} \mid X_t, Y_t). \tag{3}$$

This is the general framework adopted in the present paper. A specific Markov market model will be introduced in Section 3 below.

Of primary importance in this paper will be the kth default time for an arbitrary $k = 1, 2, \ldots, L$. Let $\tau^{(1)} < \tau^{(2)} < \cdots < \tau^{(L)}$ be the ordering (for each ω) of the default times $\tau_1, \tau_2, \ldots, \tau_L$. By definition, the kth default time is $\tau^{(k)}$.

It will be convenient to represent some probabilities associated with the kth default time in terms of the *cumulative default process* H, defined as the increasing process

$$H_t = \sum_{l=1}^{L} H_t^l,$$

where $H_t^l = \mathbb{1}_{\{X_t^l = K\}} = \mathbb{1}_{\{\tau_l \leqslant t\}}$ for every $t \in \mathbb{R}_+$. Evidently $\mathbb{H} \subseteq \widetilde{\mathbb{H}}$, where $\mathbb{H}$ is the filtration generated by the cumulative default process H. It is obvious that the process $S := (H, X, Y)$ has the Markov property under $\mathbb{P}$ with respect to the filtration $\mathbb{G}$. Also, it is useful to observe that we have $\{\tau^{(1)} > t\} = \{H_t = 0\}$, $\{\tau^{(k)} \leqslant t\} = \{H_t \geqslant k\}$ and $\{\tau^{(k)} = \tau_l\} = \{H_{\tau_l} = k\}$ for every $l, k = 1, 2, \ldots, L$.

2.2 Conditional expectations

Although we shall later focus on a Markovian set-up, in the sense of equality (3), we shall first derive some preliminary results in a more general set-up. To this end, it will be convenient to use the notation $\mathcal{F}^{X,t} = \sigma(X_s; s \geqslant t)$ and $\mathcal{F}^{Y,t} = \sigma(Y_s; s \geqslant t)$ for the information generated by the processes X and Y after time t. We postulate that for any random variable $Z \in \mathcal{F}^{X,t} \vee \mathcal{F}_\infty^Y$ and any bounded measurable function g, it holds that

$$\mathbb{E}_{\mathbb{P}}\big(g(Z) \mid \mathcal{G}_t\big) = \mathbb{E}_{\mathbb{P}}\big(g(Z) \mid \sigma(X_t) \vee \mathcal{F}_t^Y\big). \tag{4}$$

This implies, in particular, that the migration process X is *conditionally Markov* with regard to the reference filtration $\mathbb{F}^Y$, that is, for every $0 \leqslant t \leqslant s$ and $x \in \mathcal{X}$,

$$\mathbb{P}(X_s = x \mid \mathcal{G}_t) = \mathbb{P}\big(X_s = x \mid \sigma(X_t) \vee \mathcal{F}_t^Y\big). \tag{5}$$

Note that the Markov condition (3) is stronger than condition (4). We assume from now on that $t \geqslant 0$ and $x \in \mathcal{X}$ are such that $p_x(t) := \mathbb{P}(X_t = x \mid \mathcal{F}_t^Y) > 0$. We begin the analysis of conditional expectations with the following lemma.

Lemma 2.1. *Let $k \in \{1, 2, \ldots, L\}$, $x \in \mathcal{X}$, and let $Z \in \mathcal{F}^{X,t} \vee \mathcal{F}_\infty^Y$ be an integrable random variable. Then we have, for every $0 \leqslant t \leqslant s$,*

$$\mathbb{1}_{\{X_t = x\}} \mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s < k\}} Z | \mathcal{G}_t) = \mathbb{1}_{\{H_t < k, X_t = x\}} \frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s < k, X_t = x\}} Z \mid \mathcal{F}_t^Y)}{p_x(t)}. \tag{6}$$

Consequently,

$$\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s < k\}} Z | \mathcal{G}_t) = \mathbb{1}_{\{H_t < k\}} \sum_{x \in \mathcal{X}} \mathbb{1}_{\{X_t = x\}} \frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s < k, X_t = x\}} Z \mid \mathcal{F}_t^Y)}{p_x(t)}. \tag{7}$$

Proof. Let A_t be an arbitrary event from $\mathcal{G}_t$. We need to check that

$$\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{A_t}\mathbb{1}_{\{X_t=x\}}\mathbb{1}_{\{H_s<k\}}Z) = \mathbb{E}_{\mathbb{P}}\left(\mathbb{1}_{A_t}\mathbb{1}_{\{H_t<k,X_t=x\}}\frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k,X_t=x\}}Z \mid \mathcal{F}_t^Y)}{p_x(t)}\right).$$

Since $\{H_s < k\} \subset \{H_t < k\}$ and the random variable $\tilde{Z} := \mathbb{1}_{\{H_s<k,X_t=x\}}Z$ belongs to $\mathcal{F}^{X,t} \vee \mathcal{F}_\infty^Y$, the left-hand side is equal to

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\big(\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{A_t}\mathbb{1}_{\{H_t<k,X_t=x\}}\mathbb{1}_{\{H_s<k\}}Z \mid \mathcal{G}_t)\big)\\
&\quad= \mathbb{E}_{\mathbb{P}}\big(\mathbb{1}_{A_t}\mathbb{1}_{\{H_t<k,X_t=x\}}\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k,X_t=x\}}Z \mid \mathcal{G}_t)\big)\\
&\quad= \mathbb{E}_{\mathbb{P}}\big(\mathbb{1}_{A_t}\mathbb{1}_{\{H_t<k,X_t=x\}}\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k,X_t=x\}}Z \mid \sigma(X_t) \vee \mathcal{F}_t^Y)\big)\\
&\quad= \mathbb{E}_{\mathbb{P}}\left(\mathbb{1}_{A_t}\mathbb{1}_{\{H_t<k,X_t=x\}}\frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k,X_t=x\}}Z \mid \mathcal{F}_t^Y)}{p_x(t)}\right),
\end{aligned}$$

where the second equality is a consequence of (4), and the last one follows from the equality

$$\mathbb{1}_{\{X_t=x\}}\mathbb{E}_{\mathbb{P}}\big(\widehat{Z} \mid \sigma(X_t) \vee \mathcal{F}_t^Y\big) = \mathbb{1}_{\{X_t=x\}}\frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{X_t=x\}}\widehat{Z} \mid \mathcal{F}_t^Y)}{\mathbb{P}(X_t = x \mid \mathcal{F}_t^Y)},$$

which is valid for any integrable random variable $\widehat{Z}$. Equality (7) is an immediate consequence of (6). □

In the case of a single credit name, that is, in the case of $L = 1$, we have for any $t \geqslant 0$ that $\{H_t < 1\} = \{H_t \neq 1\} = \{X_t \neq K\}$. This leads to the following result.

Corollary 2.1. *Let $L = 1$ and let $Z \in \mathcal{F}^{X,t} \vee \mathcal{F}_\infty^Y$ be an integrable random variable. Then we have, for any $0 \leqslant t \leqslant s$,*

$$\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{X_s\neq K\}}Z \mid \mathcal{G}_t) = \sum_{x=1}^{K-1}\mathbb{1}_{\{X_t=x\}}\frac{\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{X_s\neq K,X_t=x\}}Z \mid \mathcal{F}_t^Y)}{p_x(t)}. \tag{8}$$

For any $0 \leqslant t \leqslant s$, we write

$$\begin{aligned}
q_{k,x;t}(s) &= \mathbb{P}\big(H_s < k, X_t = x \mid \mathcal{F}_t^Y\big) = \mathbb{P}\big(\tau^{(k)} > s, X_t = x \mid \mathcal{F}_t^Y\big),\\
p_{k,x;t}(s) &= \mathbb{P}\big(H_s \geqslant k, X_t = x \mid \mathcal{F}_t^Y\big) = \mathbb{P}\big(\tau^{(k)} \leqslant s, X_t = x \mid \mathcal{F}_t^Y\big),
\end{aligned}$$

so that formally $\mathrm{d}p_{k,x;t}(s) = \mathbb{P}(\tau^{(k)} \in ds, X_t = x|\mathcal{F}_t^Y)$. The following proposition extends Lemma 2.1.

Proposition 2.1. *Let $k \in \{1, 2, \ldots, L\}$ and let Z be an integrable, $\mathbb{F}^Y$-predictable process. Then we have, for every $0 \leqslant t \leqslant s$,*

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{t<\tau^{(k)}\leqslant s\}} Z_{\tau^{(k)}} \mid \mathcal{G}_t)\\ &\quad = \mathbb{1}_{\{H_t<k\}} \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X_t=x\}}}{p_x(t)} \mathbb{E}_{\mathbb{P}}\Big(\int_{]t,s]} Z_u \,\mathrm{d}p_{k,x;t}(u) \mid \mathcal{F}_t^Y \Big). \end{aligned} \tag{9}$$

Proof. Let $t < \alpha < \beta < s$. Let us first establish (9) for a process Z of the form $Z_u = \mathbb{1}_{]\alpha,\beta]}(u) Z_\alpha$ where Z_α is a $\mathcal{F}_\alpha^Y$-measurable, integrable random variable. In this case, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{t<\tau^{(k)}\leqslant s\}} Z_{\tau^{(k)}} \mid \mathcal{G}_t)\\ &\quad = \mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{\alpha<\tau^{(k)}\leqslant \beta\}} Z_\alpha \mid \mathcal{G}_t)\\ &\quad = \mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_\alpha<k\}} Z_\alpha \mid \mathcal{G}_t) - \mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_\beta<k\}} Z_\alpha \mid \mathcal{G}_t)\\ &\quad = \mathbb{1}_{\{H_t<k\}} \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X_t=x\}}}{p_x(t)} \mathbb{E}_{\mathbb{P}}\big(Z_\alpha \big[q_{k,x,t}(\alpha) - q_{k,x,t}(\beta) \big] \mid \mathcal{F}_t^Y \big)\\ &\quad = \mathbb{1}_{\{H_t<k\}} \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X_t=x\}}}{p_x(t)} \mathbb{E}_{\mathbb{P}}\Big(\int_{]t,s]} Z_u \,\mathrm{d}p_{k,x;t}(u) \mid \mathcal{F}_t^Y \Big), \end{aligned}$$

where the third equality follows easily from (7) and the definitions of $q_{k,x;t}(s)$ and $p_{k,x;t}(s)$. The general case follows by standard approximation arguments. □

Corollary 2.2. *Let $L = 1$ and let Z be an integrable, $\mathbb{F}^Y$-predictable stochastic process. Then we have, for every $0 \leqslant t \leqslant s$,*

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{t<\tau\leqslant s\}} Z_\tau \mid \mathcal{G}_t)\\ &\quad = \mathbb{1}_{\{X_t\neq K\}} \sum_{x=1}^{K-1} \frac{\mathbb{1}_{\{X_t=x\}}}{p_x(t)} \mathbb{E}_{\mathbb{P}}\Big(\int_{]t,s]} Z_u \,\mathrm{d}p_{1,k;t}(u) \mid \mathcal{F}_t^Y \Big). \end{aligned} \tag{10}$$

For $K = 2$, Corollaries 2.1 and 2.2 coincide with Lemma 5.1.2(i) and Proposition 5.1.1(i) in Bielecki and Rutkowski (2002a), respectively.

2.2.1 Markovian case

Let us now assume the Markovian set-up of Section 2.1.1. Let Z be a $\mathcal{G}^t = \mathcal{F}^{X,t} \vee \mathcal{F}^{Y,t}$-measurable, integrable random variable. Then formula (7) yields, for every $0 \leqslant t \leqslant s$,

$$\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k\}} Z \mid \mathcal{G}_t) = \mathbb{1}_{\{H_t<k\}} \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X_t=x\}}}{\bar{p}_x(t)} \mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{H_s<k, X_t=x\}} Z \mid Y_t), \tag{11}$$

where $\bar{p}_x(t) = \mathbb{P}(X_t = x \mid Y_t)$, and formula (9) becomes

$$\mathbb{E}_{\mathbb{P}}(\mathbb{1}_{\{t<\tau^{(k)}\leqslant s\}} Z_{\tau^{(k)}} \mid \mathcal{G}_t) = \mathbb{1}_{\{H_t<k\}} \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X_t=x\}}}{\bar{p}_x(t)} \mathbb{E}_{\mathbb{P}}\Big(\int_{]t,s]} Z_u \,\mathrm{d}\bar{p}_{k,x;t}(u) \mid Y_t \Big), \tag{12}$$

where

$$\bar{p}_{k,x;t}(u) = \mathbb{P}(H_u \geqslant k, X_t = x \mid Y_u) = \mathbb{P}(\tau^{(k)} \leqslant u, X_t = x \mid Y_u).$$

3 Markovian market model

We assume that the factor process Y takes values in $\mathbb{R}^n$ so that the state space for the process $M = (X, Y)$ is $\mathcal{X} \times \mathbb{R}^n$. At the intuitive level, we wish to model the process $M = (X, Y)$ as a combination of a Markov chain X modulated by the Lévy-like process Y and a Lévy-like process Y modulated by a Markov chain X. To be more specific, we postulate that the *infinitesimal generator* $\mathbf{A}$ of M is given as

$$\begin{aligned}\mathbf{A}f(x,y) = (1/2)\sum_{i,j=1}^{n} a_{ij}(x,y)\partial_i\partial_j f(x,y) + \sum_{i=1}^{n} b_i(x,y)\partial_i f(x,y) \\ + \gamma(x,y)\int_{\mathbb{R}^n} \big(f\big(x, y+g(x,y,y')\big) - f(x,y)\big)\Pi(x,y;\mathrm{d}y') \\ + \sum_{x'\in\mathcal{X}} \lambda(x,x';y) f(x',y),\end{aligned}$$

where $\lambda(x, x'; y) \geqslant 0$ for every $x = (x^1, x^2, \ldots, x^L) \neq (x'^1, x'^2, \ldots, x'^L) = x'$, and

$$\lambda(x,x;y) = -\sum_{x'\in\mathcal{X}, x'\neq x} \lambda(x,x';y).$$

Here ∂_i denotes the partial derivative with respect to the variable y^i. The existence and uniqueness of a Markov process M with the generator $\mathbf{A}$ will follow (under appropriate technical conditions) from the respective results regarding martingale problems. Specifically, one can refer to Theorems 4.1 and 5.4 in Chapter 4 of Ethier and Kurtz (1986).

We find it convenient to refer to X (Y, respectively) as the *Markov chain component* of M (the *jump-diffusion component* of M, respectively). At any time t, the intensity matrix of the Markov chain component is given as $\Lambda_t =$

$[\lambda(x, x'; Y_t)]_{x,x' \in \mathcal{X}}$. The jump-diffusion component satisfies the SDE:

$$\begin{aligned} \mathrm{d}Y_t &= b(X_t, Y_t)\,\mathrm{d}t + \sigma(X_t, Y_t)\,\mathrm{d}W_t \\ &\quad + \int_{\mathbb{R}^n} g(X_{t-}, Y_{t-}, y')\pi(X_{t-}, Y_{t-}; \mathrm{d}y', \mathrm{d}t), \end{aligned}$$

where, for a fixed $(x, y) \in \mathcal{X} \times \mathbb{R}^n$, $\pi(x, y; \mathrm{d}y', \mathrm{d}t)$ is a Poisson measure with the intensity measure $\gamma(x, y)\Pi(x, y; \mathrm{d}y')\,\mathrm{d}t$, and where $\sigma(x, y)$ satisfies the equality $\sigma(x, y)\sigma(x, y)^{\mathsf{T}} = a(x, y)$.

Remark 3.1. If we take $g(x, y, y') = y'$, and we suppose that the coefficients $\sigma = [\sigma_{ij}]$, $b = [b_i]$, γ, and the measure Π do not depend on x and y then the factor process Y is a Poisson-Lévy process with the characteristic triplet (a, b, ν), where the diffusion matrix is $a(x, y) = \sigma(x, y)\sigma(x, y)^{\mathsf{T}}$, the "drift" vector is $b(x, y)$, and the Lévy measure is $\nu(\mathrm{d}y) = \gamma\Pi(\mathrm{d}y)$. In this case, the migration process X is modulated by the factor process Y, but not vice versa. We shall not study here the "infinite activity" case, that is, the case when the jump measure π is not a Poisson measure, and the related Lévy measure is an infinite measure.

We shall provide with more structure the Markov chain part of the generator **A**. Specifically, we make the following standing assumption.

Assumption (M). The infinitesimal generator of the process $M = (X, Y)$ takes the following form

$$\begin{aligned} \mathbf{A}f(x, y) &= (1/2)\sum_{i,j=1}^{n} a_{ij}(x, y)\partial_i\partial_j f(x, y) + \sum_{i=1}^{n} b_i(x, y)\partial_i f(x, y) \\ &\quad + \gamma(x, y)\int_{\mathbb{R}^n} \big(f\big(x, y + g(x, y, y')\big) - f(x, y)\big)\Pi(x, y; \mathrm{d}y') \\ &\quad + \sum_{l=1}^{L}\sum_{x^{l\prime} \in \mathcal{K}} \lambda^l(x, x'_l; y) f(x'_l, y), \end{aligned} \tag{13}$$

where we write $x'_l = (x^1, x^2, \ldots, x^{l-1}, x'^l, x^{l+1}, \ldots, x^L)$.

Note that x'_l is the vector $x = (x^1, x^2, \ldots, x^L)$ with the lth coordinate x^l replaced by x'^l. In the case of two obligors (i.e., for $L = 2$), the generator becomes

$$\mathbf{A}f(x,y) = (1/2)\sum_{i,j=1}^{n} a_{ij}(x,y)\partial_i\partial_j f(x,y) + \sum_{i=1}^{n} b_i(x,y)\partial_i f(x,y)$$
$$+ \gamma(x,y)\int_{\mathbb{R}^n} \big(f\big(x, y+g(x,y,y')\big) - f(x,y)\big)\Pi(x,y;\mathrm{d}y')$$
$$+ \sum_{x'^1\in\mathcal{K}} \lambda^1(x,x'_1;y)f(x'_1,y) + \sum_{x'^2\in\mathcal{K}} \lambda^2(x,x'_2;y)f(x'_2,y),$$

where $x = (x^1, x^2)$, $x'_1 = (x'^1, x^2)$ and $x'_2 = (x^1, x'^2)$. In this case, coming back to the general form, we have for $x = (x^1, x^2)$ and $x' = (x'^1, x'^2)$

$$\lambda(x,x';y) = \begin{cases} \lambda^1(x,x'_1;y), & \text{if } x^2 = x'^2, \\ \lambda^2(x,x'_2;y), & \text{if } x^1 = x'^1, \\ 0, & \text{otherwise.} \end{cases}$$

Similar expressions can be derived in the case of a general value of L. Note that the model specified by (13) does not allow for simultaneous jumps of the components X^l and $X^{l'}$ for $l \neq l'$. In other words, the ratings of different credit names may not change simultaneously. Nevertheless, this is not a serious lack of generality, as the ratings of both credit names may still change in an arbitrarily small time interval. The advantage is that, for the purpose of simulation of paths of process X, rather than dealing with $\mathcal{X}\times\mathcal{X}$ intensity matrix $[\lambda(x,x';y)]$, we shall deal with L intensity matrices $[\lambda^l(x,x'_l;y)]$, each of dimension $\mathcal{K}\times\mathcal{K}$ (for any fixed y). The structure (13) is assumed in the rest of the paper. Let us stress that within the present set-up the current credit rating of the credit name l directly impacts the intensity of transition of the rating of the credit name l', and vice versa. This property, known as *frailty*, may contribute to default contagion.

Remark 3.2. (i) It is clear that we can incorporate in the model the case when some – possibly all – components of the factor process Y follow Markov chains themselves. This feature is important, as factors such as economic cycles may be modeled as Markov chains. It is known that default rates are strongly related to business cycles.

(ii) Some of the factors $Y^1, Y^2, \ldots, Y^d$ may represent cumulative duration of visits of rating processes X^l in respective rating states. For example, we may set $Y^1_t = \int_0^t \mathbb{1}_{\{X^1_s=1\}}\,\mathrm{d}s$. In this case, we have $b_1(x,y) = \mathbb{1}_{\{x^1=1\}}(x)$, and the corresponding components of coefficients σ and g equal zero.

(iii) In the area of *structural arbitrage*, so called *credit-to-equity* (C2E) models and/or *equity-to-credit* (E2C) models are studied. Our market model nests both types of interactions, that is C2E and E2C. For example, if one of the factors is the price process of the equity issued by a credit name, and if credit migration intensities depend on this factor (implicitly or explicitly) then we have a E2C type interaction. On the other hand, if credit ratings of a given obligor impact

the equity dynamics (of this obligor and/or some other obligors), then we deal with a C2E type interaction.

As already mentioned, $S = (H, X, Y)$ is a Markov process on the state space $\{0, 1, \ldots, L\} \times \mathcal{X} \times \mathbb{R}^d$ with respect to its natural filtration. Given the form of the generator of the process (X, Y), we can easily describe the generator of the process (H, X, Y). It is enough to observe that the transition intensity at time t of the component H from the state H_t to the state $H_t + 1$ is equal to $\sum_{l=1}^{L} \lambda^l(X_t, K; X_t^{(l)}, Y_t)$, provided that $H_t < L$ (otherwise, the transition intensity equals zero), where we write $X_t^{(l)} = (X_t^1, \ldots, X_t^{l-1}, X_t^{l+1}, \ldots, X_t^L)$ and we set $\lambda^l(x^l, x'^l; x^{(l)}, y) = \lambda^l(x, x'_l; y)$.

3.1 Specification of credit ratings transition intensities

One always needs to find a compromise between realistic features of a financial model and its complexity. This issue frequently nests the issues of functional representation of a model, as well as its parameterization. We present here an example of a particular model for credit ratings transition rates, which is rather arbitrary, but is nevertheless relatively simple and should be easy to estimate/calibrate.

Let $\bar{X}_t$ be the average credit rating at time t, so that

$$\bar{X}_t = \frac{1}{L} \sum_{l=1}^{L} X_t^l.$$

Let $\mathcal{L} = \{l_1, l_2, \ldots, l_{\hat{L}}\}$ be a subset of the set of all obligors, where $\hat{L} < L$. We consider $\mathcal{L}$ to be a collection of "major players" whose economic situation, reflected by their credit ratings, effectively impacts all other credit names in the pool. The following exponential-linear "regression" model appears to be a plausible model for the rating transition intensities:

$$\begin{aligned} \ln \lambda^l(x, x'_l; y) = {} & \alpha_{l,0}(x^l, x'^l) + \sum_{j=1}^{n} \alpha_{l,j}(x^l, x'^l) y_j + \beta_{l,0}(x^l, x'^l) h \\ & + \sum_{i=1}^{\bar{L}} \beta_{l,i}(x^l, x'^l) x^i + \tilde{\beta}_l(x^l, x'^l) \bar{x} \\ & + \hat{\beta}_l(x^l, x'^l)(x^l - x'^l), \end{aligned} \tag{14}$$

where h represents a generic value of H_t, so that $h = \sum_{l=1}^{L} \mathbb{1}_{\{K\}}(x^l)$, and $\bar{x}$ represents a generic value of $\bar{X}_t$, that is, $\bar{x} = \frac{1}{L} \sum_{l=1}^{L} x^l$.

The number of parameters involved in (14) can easily be controlled by the number of model variables, in particular – the number of factors and the number of credit ratings, as well as structure of the transition matrix (see Section 7.2

below). In addition, the reduction of the number of parameters can be obtained if the pool of all L obligors is partitioned into a (small) number of homogeneous sub-pools. All of this is a matter of practical implementation of the model. Assume, for instance, that there are $\tilde{L} \ll L$ homogeneous sub-pools of obligors, and the parameters α, β, $\tilde{\beta}$ and $\hat{\beta}$ in (14) do not depend on x^l, x'^l. Then the migration intensities (14) are parameterized by $\tilde{L}(n + \hat{L} + 4)$ parameters.

3.2 *Conditionally independent migrations*

Suppose that the intensities $\lambda^l(x, x'_l; y)$ do not depend on $x^{(l)} = (x^1, x^2, \ldots, x^{l-1}, x^{l+1}, \ldots, x^L)$ for every $l = 1, 2, \ldots, L$. In addition, assume that the dynamics of the factor process Y do not depend on the migration process X. It turns out that in this case, given the structure of our generator as in (13), the migration processes X^l, $l = 1, 2, \ldots, L$, are conditionally independent given the sample path of the process Y.

We shall illustrate this point in the case of only two credit names in the pool (i.e., for $L = 2$) and assuming that there is no factor process, so that conditional independence really means independence between migration processes X^1 and X^2. For this, suppose that X^1 and X^2 are independent Markov chains, each taking values in the state space $\mathcal{K}$, with infinitesimal generator matrices Λ^1 and Λ^2, respectively. It is clear that the joint process $X = (X^1, X^2)$ is a Markov chain on $\mathcal{K} \times \mathcal{K}$. An easy calculation reveals that the infinitesimal generator of the process X is given as

$$\Lambda = \Lambda^1 \otimes \mathrm{Id}_K + \mathrm{Id}_K \otimes \Lambda^2,$$

where Id_K is the identity matrix of order K and $\otimes$ denotes the matrix tensor product. This agrees with the structure (13) in the present case.

4 Changes of measures and Markovian numeraires

In financial applications, one frequently deals with various absolutely continuous probability measures. In order to exploit – for pricing applications – the Markovian structure of the market model introduced above, we need that the model is Markovian under a particular pricing measure corresponding to some particular numeraire process β that is convenient to use for some reasons. The model does not have to be Markovian under some other equivalent probability measures, such as the statistical probability, say $\mathbb{Q}$, or the spot martingale measure, say $\mathbb{Q}^*$. Nevertheless, it may be sometimes desirable that the Markovian structure of the market model is preserved under an equivalent change of a probability measure, for instance, when we change a numeraire from β to β'. In this section, we shall provide some discussion of the issue of preservation of the Markov property of the process M.

Let $T^* > 0$ be a fixed horizon date, and let η be a strictly positive, $\mathcal{G}_{T^*}$-measurable random variable such that $\mathbb{E}_{\mathbb{P}}\eta = 1$. We define an equivalent probability measure $\mathbb{P}^\eta$ on $(\Omega, \mathcal{G}_{T^*})$ by the equality

$$\frac{\mathrm{d}\mathbb{P}^\eta}{\mathrm{d}\mathbb{P}} = \eta, \quad \mathbb{P}\text{-a.s.}$$

4.1 Markovian change of a probability measure

We place ourselves in the setup of Section 2.1.1, and we follow Palmowski and Rolski (2002) in the presentation below. The standing assumption is that the process $M = (X, Y)$ has the Markov property under $\mathbb{P}$ with respect to the filtration $\mathbb{G}$. Let $(\mathbf{A}, \mathcal{D}(\mathbf{A}))$ be the *extended generator* of M. This means that the process

$$M_t^f = f(M_t) - \int_0^t \mathbf{A}f(M_s)\,\mathrm{d}s \tag{15}$$

is a local $\mathbb{G}$-martingale for any function f in $\mathcal{D}(\mathbf{A})$.

For any strictly positive function $h \in \mathcal{D}(\mathbf{A})$, we define an auxiliary process η^h by setting

$$\eta_t^h = \frac{h(M_t)}{h(M_0)} \exp\left(- \int_0^t \frac{(\mathbf{A}h)(M_s)}{h(M_s)}\,\mathrm{d}s \right), \quad t \in [0, T^*]. \tag{16}$$

Any function h for which the process η^h is a martingale is called a *good* function for $\mathbf{A}$. Observe that for any such function h, the equality $\mathbb{E}_{\mathbb{P}}(\eta_t^h) = 1$ holds for every $t \in [0, T^*]$. Note also that any constant function h is a good function for $\mathbf{A}$; in this case we have, of course, that $\eta^h \equiv 1$. The next lemma follows from results of Palmowski and Rolski (2002) (see Lemma 3.1 therein).

Lemma 4.1. *Let h be a good function for $\mathbf{A}$. Then the process η^h is given as Doléans exponential martingale*

$$\eta_t^h = \mathcal{E}_t\big(N^h\big),$$

where the local martingale N^h is given as

$$N_t^h = \int_0^t \kappa_{s-}^h\,\mathrm{d}M_s^h$$

with $\kappa_t^h = 1/h(M_t)$. In other words, the process η^h satisfies the SDE

$$\mathrm{d}\eta_t^h = \eta_{t-}^h \kappa_{t-}^h\,\mathrm{d}M_t^h, \quad \eta_0^h = 1. \tag{17}$$

Proof. An application of Itô's formula yields

$$\mathrm{d}\eta_t^h = \frac{1}{h(M_0)} \exp\left(- \int_0^t \frac{(\mathbf{A}h)(M_s)}{h(M_s)} \,\mathrm{d}s \right) \mathrm{d}M_t^h,$$

where the local martingale M^h is given by (15). This proves formula (17). □

For any good function h for $\mathbf{A}$, we define an equivalent probability measure $\mathbb{P}^h$ on $(\Omega, \mathcal{G}_{T^*})$ by setting

$$\frac{\mathrm{d}\mathbb{P}^h}{\mathrm{d}\mathbb{P}} = \eta_{T^*}^h, \quad \mathbb{P}\text{-a.s.} \tag{18}$$

From Kunita and Watanabe (1963), we deduce that the process M preserves its Markov property with respect to the filtration $\mathbb{G}$ when the probability measure $\mathbb{P}$ is replaced by $\mathbb{P}^h$. In order to find the extended generator of M under $\mathbb{P}^h$, we set

$$\mathbf{A}^h f = \frac{1}{h}\big[\mathbf{A}(fh) - f\mathbf{A}(h)\big],$$

and we define the following two sets:

$$\mathcal{D}_{\mathbf{A}}^h = \left\{ f \in \mathcal{D}(\mathbf{A}) \colon\ fh \in \mathcal{D}(\mathbf{A}) \quad \text{and} \quad \int_0^{T^*} \big|\mathbf{A}^h f(M_s)\big| \,\mathrm{d}s < \infty,\ \mathbb{P}^h\text{-a.s.} \right\}$$

and

$$\mathcal{D}_{\mathbf{A}^h}^{h^{-1}} = \left\{ f \in \mathcal{D}\big(\mathbf{A}^h\big) \colon\ fh^{-1} \in \mathcal{D}\big(\mathbf{A}^h\big) \quad \text{and} \quad \int_0^{T^*} \big|\mathbf{A}f(M_s)\big| \,\mathrm{d}s < \infty,\ \mathbb{P}\text{-a.s.} \right\}.$$

Then the following result holds (see Palmowski and Rolski 2002, Theorem 4.2).

Theorem 4.1. *Suppose that* $\mathcal{D}_{\mathbf{A}}^h = \mathcal{D}(\mathbf{A})$ *and* $\mathcal{D}_{\mathbf{A}^h}^{h^{-1}} = \mathcal{D}(\mathbf{A}^h)$. *Then the process* M *is Markovian under* $\mathbb{P}^h$ *with the extended generator* $\mathbf{A}^h$ *and* $\mathcal{D}(\mathbf{A}^h) = \mathcal{D}(\mathbf{A})$.

We now apply the above theorem to our model. The domain of $\mathcal{D}(\mathbf{A})$ contains all functions $f(x, y)$ with compact support that are twice continuously differentiable with respect to y. Let h be a good function. Under mild assumptions on the coefficients of $\mathbf{A}$, the assumptions of Theorem 4.1 are satisfied. It follows that the generator of M under $\mathbb{P}^h$ is given as

$$
\begin{aligned}
&\mathbf{A}^h f(x,y)\\
&\quad = (1/2)\sum_{i,j=1}^{n} a_{ij}(x,y)\partial_i\partial_j f(x,y) + \sum_{i=1}^{n} b_i^h(x,y)\partial_i f(x,y)\\
&\quad + \gamma(x,y)\int_{\mathbb{R}^n}\big(f\big(x,y+g(x,y,y')\big) - f(x,y)\big)\Pi^h(x,y;\mathrm{d}y')\\
&\quad + \sum_{x'\in\mathcal{X}} \lambda^h(x,x';y)f(x',y),
\end{aligned}
$$

where

$$
\begin{aligned}
b_i^h(x,y) &= b_i(x,y) + \frac{1}{h(x,y)}\sum_{i,j=1}^{n} a_{ij}(x,y)\partial_j h(x,y),\\
\Pi^h(x,y;dy') &= \frac{h(x,y+g(x,y,y'))}{h(x,y)}\Pi(x,y;dy'),\\
\lambda^h(x,x';y) &= \lambda(x,x';y)\frac{h(x',y)}{h(x,y)}, \quad x\neq x',\\
\lambda^h(x,x;y) &= -\sum_{x'\neq x}\lambda^h(x,x';y).
\end{aligned}
\tag{19}
$$

Before we proceed to the issue of valuation of credit derivatives, we state the following useful result, whose easy proof is omitted.

Lemma 4.2. *Let h and h' be two good functions for $\mathbf{A}$. Then $\phi(h,h') := h'/h$ is a good function for $\mathbf{A}^h$. Moreover, we have that*

$$
\frac{\mathrm{d}\mathbb{P}^{h'}}{\mathrm{d}\mathbb{P}^h} = \eta_{T^*}^{\phi(h,h')}, \quad \mathbb{P}^h\text{-a.s.} \tag{20}
$$

where the process $\eta^{\phi(h,h')}$ is defined in analogy to (16) with $\mathbf{A}$ replaced with $\mathbf{A}^h$.

4.2 Markovian numeraires and valuation measures

Let us first consider a general set-up. We use here the notation and terminology of Jamshidian (2004). We fix the horizon date T^*, and we assume that $\mathcal{G} = \mathcal{G}_{T^*}$. Let us fix some ($\mathbb{G}$-adapted) *deflator process* ξ, that is, a strictly positive, integrable semimartingale, with $\xi_0 = 1$. Any $\mathcal{G}$-measurable random variable C such that $\xi_{T^*}C$ is integrable under $\mathbb{P}$ is called a *claim*. The *price process* C_t, $t\in[0,T^*]$, of a claim C is formally defined as

$$
C_t = \xi_t^{-1}\mathbb{E}_{\mathbb{P}}(\xi_{T^*}C \mid \mathcal{G}_t),
$$

so that, in particular, $C_{T^*} = C$. It is implicitly assumed here that the information carried by the filtration $\mathbb{G}$ is available to all trading agents.

Suppose that we are interested in providing valuation formulae for some financial products, and suppose that we find it convenient to use a particular *numeraire* (that is, a strictly positive claim), denoted by β. Let $\mathbb{P}^\beta$ be the corresponding *valuation measure*, defined on $(\Omega, \mathcal{G})$ as

$$\frac{d\mathbb{P}^\beta}{d\mathbb{P}} = \frac{\xi_{T^*}\beta}{\beta_0} = \frac{\xi_{T^*}\beta_{T^*}}{\beta_0}, \quad \mathbb{P}\text{-a.s.} \tag{21}$$

From the abstract Bayes rule, it follows that the price process C can be expressed as follows:

$$C_t = \beta_t \mathbb{E}_{\mathbb{P}^\beta}\big(\beta_{T^*}^{-1} C \mid \mathcal{G}_t\big).$$

As before, we assume that our market model M is Markovian under $\mathbb{P}$, where $\mathbb{P}$ might be the statistical probability measure $\mathbb{Q}$, the spot martingale measure $\mathbb{Q}^*$, or some other martingale measure. We want the process M to remain a time-homogeneous Markov process under $\mathbb{P}^\beta$.

Definition 4.1. A valuation measure $\mathbb{P}^\beta$ is said to be a *Markovian* if the process M remains a time-homogeneous Markov process under $\mathbb{P}^\beta$. Any numeraire process β such that the valuation measure $\mathbb{P}^\beta$ is Markovian is called a *Markovian numeraire*.

In view of results of Section 4.1, for a valuation measure $\mathbb{P}^\beta$ to be Markovian, it suffices that the Radon–Nikodým derivative process $\eta_t^\beta = \frac{\mathrm{d}\mathbb{P}^\beta}{\mathrm{d}\mathbb{P}}|_{\mathcal{G}_t}$ satisfies

$$\eta_t^\beta = \frac{h^\beta(M_t)}{h^\beta(M_0)} \exp\left(-\int_0^t \frac{(\mathbf{A}h^\beta)(M_s)}{h^\beta(M_s)}\,\mathrm{d}s\right), \quad t \in [0, T^*],$$

for some good function h^β for $\mathbf{A}$. The corresponding deflator process is then given as $\xi_t^\beta = \beta_0 \beta_t^{-1} \eta_t^\beta$, that is, for any claim C we have that

$$C_t = \beta_t \mathbb{E}_{\mathbb{P}^\beta}\big(\beta_{T^*}^{-1} C \mid \mathcal{G}_t\big) = \big(\xi_t^\beta\big)^{-1} \mathbb{E}_{\mathbb{P}}\big(\xi_{T^*}^\beta C \mid \mathcal{G}_t\big).$$

If β and β' are two such numeraires, and h^β and $h^{\beta'}$ are the corresponding good functions then, in view of Lemma 4.2, we have

$$\frac{\mathrm{d}\mathbb{P}^{\beta'}}{\mathrm{d}\mathbb{P}^\beta} = \eta_{T^*}^{\phi(h^\beta, h^{\beta'})}, \quad \mathbb{P}^\beta\text{-a.s.} \tag{22}$$

An interesting question arises: under what conditions on ξ and β the probability measure $\mathbb{P}^\beta$ is a Markovian valuation measure? In order to partially address this question, we shall consider the case where the valuation measure $\mathbb{P}^\beta$ is a Markovian for any constant numeraire β, that is, for any $\beta \equiv \text{const} > 0$.

Proposition 4.1. *Assume that the deflator process satisfies $\xi = \eta^h$ for some good function h for $\mathbf{A}$. Then the following statements are true*:

(i) *for any constant numeraire* β*, the valuation measure* $\mathbb{P}^\beta$ *is Markovian,*
(ii) *if a numeraire* β *is such that* $\beta = \beta_0 \eta^\chi_{T^*}/\eta^h_{T^*}$ *for some good function* χ *for* **A***, then the valuation measure* $\mathbb{P}^\beta$ *is Markovian,*
(iii) *if numeraires* β *and* β' *are such that* $\beta = \beta_0 \eta^\chi_{T^*}/\eta^h_{T^*}$ *and* $\beta' = \beta'_0 \eta^{\chi'}_{T^*}/\eta^h_{T^*}$ *for some good functions* χ *and* χ' *for* **A***, then*

$$\frac{\mathrm{d}\mathbb{P}^{\beta'}}{\mathrm{d}\mathbb{P}^\beta} = \frac{\beta'/\beta'_0}{\beta/\beta_0} = \frac{\eta^{\chi'}_{T^*}}{\eta^\chi_{T^*}}, \quad \mathbb{P}^{\xi,\beta'}\text{-a.s.} \tag{23}$$

Proof. Let $\xi = \eta^h$ for some good function h, where η^h is given by (16). Then for any constant numeraire β we get $\mathbb{P}^\beta = \mathbb{P}^h$ and thus, by results of Kunita and Watanabe (1963), the process M is Markovian under the valuation measure $\mathbb{P}^\beta$. This proves part (i). To establish the second part, it suffices to note that

$$\frac{\mathrm{d}\mathbb{P}^\beta}{\mathrm{d}\mathbb{P}} = \frac{\xi_{T^*}\beta}{\beta_0} = \eta^\chi_{T^*},$$

and to use again the result of Kunita and Watanabe (1963). Formula (23) follows easily from (21) (it can also be seen as a special case of (22)). This completes the proof. □

4.3 Examples of Markov market models

We shall now present three pertinent examples of Markov market models. We assume here that a numeraire β is given; the choice of β depends on the problem at hand.

4.3.1 Markov chain migration process

We assume here that there is no factor process Y. Thus, we only deal with a single migration process X. In this case, an attractive and efficient way to model credit migrations is to postulate that X is a *birth-and-death process* with absorption at state K. In this case, the intensity matrix Λ is tri-diagonal. To simplify the notation, we shall write $p_t(k,k') = \mathbb{P}^\beta(X_{s+t} = k' \mid X_s = k)$. The transition probabilities $p_t(k,k')$ satisfy the following system of ODEs, for $t \geqslant 0$ and $k' \in \{1,2,\ldots,K\}$,

$$\begin{aligned}
\frac{\mathrm{d}p_t(1,k')}{\mathrm{d}t} &= -\lambda(1,2)\,p_t(1,k') + \lambda(1,2)\,p_t(2,k'),\\
\frac{\mathrm{d}p_t(k,k')}{\mathrm{d}t} &= \lambda(k,k-1)\,p_t(k-1,k')\\
&\quad - \big(\lambda(k,k-1) + \lambda(k,k+1)\big)\,p_t(k,k')\\
&\quad + \lambda(k,k+1)\,p_t(k+1,k')
\end{aligned}$$

for $k = 2, 3, \ldots, K-1$, and

$$\frac{\mathrm{d}p_t(K,k')}{\mathrm{d}t} = 0,$$

with initial conditions $p_0(k,k') = \mathbb{1}_{\{k=k'\}}$. Once the transition intensities $\lambda(k,k')$ are specified, the above system can be easily solved. Note, in particular, that $p_t(K,k') = 0$ for every t if $k' \neq K$. The advantage of this representation is that the number of parameters can be kept small.

A slightly more flexible model is produced if we allow for jumps to the default state K from any other state. In this case, the master ODEs take the following form, for $t \geqslant 0$ and $k' \in \{1, 2, \ldots, K\}$,

$$\begin{aligned}\frac{\mathrm{d}p_t(1,k')}{\mathrm{d}t} &= -\big(\lambda(1,2) + \lambda(1,K)\big)p_t(1,k') + \lambda(1,2)\,p_t(2,k') \\ &\quad + \lambda(1,K)\,p_t(K,k'), \\ \frac{\mathrm{d}p_t(k,k')}{\mathrm{d}t} &= \lambda(k,k-1)\,p_t(k-1,k') - \big(\lambda(k,k-1) + \lambda(k,k+1) \\ &\quad + \lambda(k,K)\big)p_t(k,k') + \lambda(k,k+1)\,p_t(k+1,k') \\ &\quad + \lambda(k,K)\,p_t(K,k')\end{aligned}$$

for $k = 2, 3, \ldots, K-1$, and

$$\frac{\mathrm{d}p_t(K,k')}{\mathrm{d}t} = 0,$$

with initial conditions $p_0(k,k') = \mathbb{1}_{\{k=k'\}}$. Some authors model migrations of credit ratings using a (proxy) diffusion, possibly with jumps to default. The birth-and-death process with jumps to default furnishes a Markov chain counterpart of such proxy diffusion models. The nice feature of the Markov chain model is that the credit ratings are (in principle) observable state variables – whereas in case of the proxy diffusion models they are not.

4.3.2 Diffusion-type factor process

We now add a factor process Y to the model. We postulate that the factor process is a diffusion process and that the generator of the process $M = (X, Y)$ takes the form

$$\begin{aligned}\mathbf{A}f(x,y) &= (1/2)\sum_{i,j=1}^{n} a_{ij}(x,y)\partial_i\partial_j f(x,y) + \sum_{i=1}^{n} b_i(x,y)\partial_i f(x,y) \\ &\quad + \sum_{x'\in\mathcal{K}, x'\neq x} \lambda(x,x';y)\big(f(x',y) - f(x,y)\big).\end{aligned}$$

Let $\phi(t,x,y,x',y')$ be the transition probability of M. Formally,

$$\phi(t,x,y,x',y')\,\mathrm{d}y' = \mathbb{P}^{\beta}(X_{s+t} = x',\, Y_{s+t} \in \mathrm{d}y' \mid X_s = x,\, Y_s = y).$$

In order to determine the function ϕ, we need to study the following Kolmogorov equation

$$\frac{\mathrm{d}v(s,x,y)}{\mathrm{d}s} + \mathbf{A}v(s,x,y) = 0. \tag{24}$$

For the generator $\mathbf{A}$ of the present form, Eq. (24) is commonly known as the *reaction-diffusion equation*. Existence and uniqueness of classical solutions for such equations were recently studied by Becherer and Schweizer (2003). It is worth mentioning that a reaction-diffusion equation is a special case of a more general integro-partial-differential equation (IPDE). In a future work, we shall deal with issue of practical solving of equations of this kind.

4.3.3 CDS spread factor model

Suppose now that the factor process $Y_t = \kappa^{(1)}(t, T^S, T^M)$ is the forward CDS spread (for the definition of $\kappa^{(1)}(t, T^S, T^M)$, see Section 5.3 below), and that the generator for (X, Y) is

$$\mathbf{A}f(x,y) = (1/2)y^2 a(x)\frac{\mathrm{d}^2 f(x,y)}{\mathrm{d}y^2} + \sum_{x' \in \mathcal{K}, x' \neq x} \lambda(x,x')\big(f(x',y) - f(x,y)\big).$$

Thus, the credit spread satisfies the following SDE

$$\mathrm{d}\kappa^{(1)}\big(t, T^S, T^M\big) = \kappa^{(1)}\big(t, T^S, T^M\big)\sigma(X_t)\,\mathrm{d}W_t$$

for some Brownian motion process W, where $\sigma(x) = \sqrt{a(x)}$. Note that in this example $\kappa^{(1)}(t, T^S, T^M)$ is a conditionally log-Gaussian process given a sample path of the migration process X, so that we are in the position to make use of Proposition 5.1 below.

5 Valuation of single name credit derivatives

We maintain the Markovian set-up, so that $M = (X, Y)$ follows a Markov process with respect to $\mathbb{G}$ under $\mathbb{P}$. In this section, we only consider one underlying credit name, that is, we set $L = 1$. Basket credit derivatives will be studied in Section 6 below.

5.1 Survival claims

Suppose that β is a Markovian numeraire, in the sense of Definition 4.1. Let us fix $t \in [0, T]$, and let us assume that a claim C and the random variable β_t/β_{T^*} are measurable with respect to $\mathcal{G}^t = \mathcal{F}^{X,t} \vee \mathcal{F}^{Y,t}$. Then we deduce

easily that $C_t = \mathcal{V}_t^{\xi,\beta}(C)$, where

$$\mathcal{V}_t^{\xi,\beta}(C) = \mathbb{E}_{\mathbb{P}^\beta}\big(\beta_t \beta_{T^*}^{-1} C \mid M_t\big).$$

A claim C such that $C = 0$ on the set $\{\tau \leqslant T\}$, so that

$$C = \mathbb{1}_{\{\tau > T\}} C = \mathbb{1}_{\{X_T \neq K\}} C = \mathbb{1}_{\{H_T < 1\}} C,$$

is termed a *T-survival claim*. For a survival claim, a more explicit expression for the price can be established. Since most standard credit derivatives can be seen as survival claims, the following simple result will prove useful in what follows.

Lemma 5.1. *Assume that a claim C and that the random variable β_t/β_{T^*} are measurable with respect to $\mathcal{G}^t$. If C is a T-survival claim then we have*

$$C_t = \mathbb{1}_{\{X_t \neq K\}} \mathcal{V}_t^{\xi,\beta}(C) = \sum_{x=1}^{K-1} \mathbb{1}_{\{X_t = x\}} \frac{\mathbb{E}_{\mathbb{P}^\beta}(\mathbb{1}_{\{X_t = x\}} \beta_t \beta_{T^*}^{-1} C | Y_t)}{\mathbb{P}^\beta(X_t = x | Y_t)}.$$

Proof. The first equality is clear. To derive the second equality above, it suffices to apply formula (11) with $L = 1$, $s = T$, $k = 1$ and $Z = C$. □

Remark 5.1. Assume that $K = 2$, so that only the pre-default state ($x = 1$) and the default state ($x = 2$) are recognized. Then we have, for any T-survival claim C,

$$C_t = \mathbb{1}_{\{X_t \neq 2\}} \mathcal{V}_t^{\xi,\beta}(C) = \frac{\mathbb{1}_{\{X_t = 1\}} V_t^{\xi,\beta}(C)}{\mathbb{P}^\beta(X_t = 1 \mid Y_t)},$$

where $V_t^{\xi,\beta}(C) = \mathbb{E}_{\mathbb{P}^\beta}(\beta_t \beta_{T^*}^{-1} C \mid Y_t)$. In the paper by Jamshidian (2004), the process $V_t^{\xi,\beta}(C)$ is termed the *pre-price* of C.

5.2 Credit default swaps

The standing assumption is that β is a Markovian numeraire and β_t/β_s is $\mathcal{G}^t$-measurable for any $t \leqslant s$. For simplicity, we shall discuss a vanilla *credit default swap* (CDS, for short) written on a corporate discount bond under the fractional recovery of par covenant. We suppose that the maturity of the reference bond is U, and the maturity of the swap is $T < U$.

5.2.1 Default payment leg

Let $N = 1$ be a notional amount of the bond, and let δ be a deterministic recovery rate in case of default. The recovery is paid at default, so that the cash flow associated with the *default payment leg* – also known as the *reference leg* – is given by $(1 - \delta)\mathbb{1}_{\{\tau \leqslant T\}} \mathbb{1}_\tau(t)$ per unit of a notional amount, where τ is the default time of a reference credit name. Consequently, the time-t value of the

default payment leg is equal to

$$A_t^{(1)} = (1-\delta)\mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{t<\tau\leqslant T\}}\beta_t\beta_\tau^{-1} \mid M_t\big) = (1-\delta)\sum_{x=1}^{K-1}\mathbb{1}_{\{X_t=x\}}\frac{\mathbb{E}_{\mathbb{P}^\beta}(\mathbb{1}_{\{X_t=x,X_T=K\}}\beta_t\beta_\tau^{-1} \mid Y_t)}{\mathbb{P}^\beta(X_t=x \mid Y_t)}.$$

The notation $A^{(1)}$ refers to the first default, which is formally the case here, since we currently deal with one name only. Since $L=1$, the cumulative default process H takes values in the set $\{0,1\}$, and we have that $\{H_t=1\}=\{X_t=K\}$.

Since the process $S=(H,X,Y)$ is a Markov process under $\mathbb{P}^\beta$, and the transition intensity at time t of a jump from $H_t=0$ to H_t+1 is $\lambda(X_t,K;Y_t)$. Hence, it is easy to write down the form of the generator of the process S. Using the Chapman–Kolmogorov equation, we can thus compute the conditional probability (recall that conditioning on S_t is equivalent to conditioning on M_t)

$$\mathbb{P}^\beta(\tau\leqslant s \mid S_t)=\mathbb{P}^\beta(\tau\leqslant s \mid M_t).$$

Knowing the conditional density $\mathbb{P}^\beta(\tau\in \mathrm{d}s \mid M_t)$, we can evaluate the conditional expectation

$$\mathbb{E}_{\mathbb{P}^\beta}(\mathbb{1}_{\{t<\tau\leqslant T\}}\beta_t\beta_\tau^{-1} \mid M_t).$$

For example, if β is a deterministic function of time then we have

$$\mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{t<\tau\leqslant T\}}\beta_t\beta_\tau^{-1} \mid M_t\big)=\beta_t\int_t^T\beta_s^{-1}\mathbb{P}^\beta(\tau\in \mathrm{d}s \mid M_t).$$

5.2.2 Premium payment leg

Let $\mathcal{T}=\{T_1,T_2,\ldots,T_J\}$ be the tenor of the *premium payment*, where $0=T_0<T_1<\cdots<T_J<T$. If the premium accrual covenant is in force, then the cash flows associated with the premium payment leg are

$$\kappa\left(\sum_{j=1}^J\mathbb{1}_{\{T_j<\tau\}}\mathbb{1}_{T_j}(t)+\sum_{j=1}^J\mathbb{1}_{\{T_{j-1}<\tau\leqslant T_j\}}\mathbb{1}_\tau(t)\frac{t-T_{j-1}}{T_j-T_{j-1}}\right),$$

where κ is the *CDS premium* (also known as the *CDS spread*). Thus, the time-t value of the premium payment leg equals $\kappa B_t^{(1)}$, where

$$B_t^{(1)}=\mathbb{E}_{\mathbb{P}^\beta}\left(\mathbb{1}_{\{t<\tau\}}\left[\sum_{j=j(t)}^J\frac{\beta_t}{\beta_{T_j}}\mathbb{1}_{\{T_j<\tau\}}+\sum_{j=j(t)}^J\frac{\beta_t}{\beta_\tau}\mathbb{1}_{\{T_{j-1}<\tau\leqslant T_j\}}\frac{\tau-T_{j-1}}{T_j-T_{j-1}}\right]\Big|\,M_t\right),$$

where $j(t)$ is the smallest integer such that $T_{j(t)} > t$. Again, since we know the conditional density $\mathbb{P}^\beta(\tau \in ds \mid M_t)$, this expectation can be computed given our assumption about the numeraire β.

5.3 Forward CDS

As before, the reference claim is a defaultable bond maturing at time U. We now consider a *forward* (*start*) *CDS* with the maturity date $T^M < U$ and the start date $T^S < T^M$. If default occurs prior to or at time T^S the contract is terminated with no exchange of payments. Therefore, the two legs of this CDS are manifestly T^S-survival claims, and the valuation of a forward CDS is not much different from valuation a straight CDS discussed above.

5.3.1 Default payment leg

As before, we let $N = 1$ be the notional amount of the bond, and we let δ be a deterministic recovery rate in case of default. The recovery is paid at default, so that the cash flow associated with the default payment leg of the forward CDS can be represented as follows

$$(1-\delta)\mathbb{1}_{\{T^S<\tau\leqslant T^M\}}\mathbb{1}_\tau(t).$$

For any $t \leqslant T^S$, the time-t value of the default payment leg is equal to

$$A_t^{(1),T^S} = (1-\delta)\mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{T^S<\tau\leqslant T^M\}}\beta_t\beta_\tau^{-1} \mid M_t\big).$$

As explained above, we can compute this conditional expectation. If β is a deterministic function of time then simply

$$\mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{T^S<\tau\leqslant T^M\}}\beta_t\beta_\tau^{-1} \mid M_t\big) = \beta_t \int_{T^S}^{T^M} \beta_s^{-1}\mathbb{P}^\beta(\tau \in \mathrm{d}s \mid M_t).$$

5.3.2 Premium payment leg

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_J\}$ be the tenor of premium payment, where $T^S < T_1 < \cdots < T_J < T^M$. As before, we assume that the premium accrual covenant is in force, so that the cash flows associated with the premium payment leg are

$$\kappa\left(\sum_{j=1}^J \mathbb{1}_{\{T_j<\tau\}}\mathbb{1}_{T_j}(t) + \sum_{j=1}^J \mathbb{1}_{\{T_{j-1}<\tau\leqslant T_j\}}\mathbb{1}_\tau(t)\frac{t-T_{j-1}}{T_j-T_{j-1}}\right).$$

Thus, for any $t \leqslant T^S$ the time-t value of the premium payment leg is $\kappa B_t^{(1),T^S}$, where

$$B_t^{(1),T^S} = \mathbb{E}_{\mathbb{P}^\beta}\Bigg(\mathbb{1}_{\{T_S<\tau\}}\Bigg[\sum_{j=1}^{J}\frac{\beta_t}{\beta_{T_j}}\mathbb{1}_{\{T_j<\tau\}} + \sum_{j=1}^{J}\frac{\beta_t}{\beta_\tau}\mathbb{1}_{\{T_{j-1}<\tau\leqslant T_j\}}\frac{\tau-T_{j-1}}{T_j-T_{j-1}}\Bigg] \,\Big|\, M_t\Bigg).$$

Again, knowing the conditional density $\mathbb{P}^\beta(\tau \in ds \mid M_t)$, we can evaluate this conditional expectation.

5.4 CDS swaptions

We consider a forward CDS swap starting at T^S and maturing at $T^M > T^S$, as described in the previous section. We shall now value the corresponding *CDS swaption* with expiry date $T < T^S$. Let K be the strike CDS rate of the swaption. Then the swaption cash flow at expiry date T equals

$$\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+,$$

so that the price of the swaption equals, for any $t \leqslant T$,

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}^\beta}\big(\beta_t\beta_T^{-1}\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+ \mid M_t\big)\\ &\quad= \mathbb{E}_{\mathbb{P}^\beta}\big(\beta_t\beta_T^{-1}B_T^{(1),T^S}\big(\kappa^{(1)}\big(t,T^S,T^M\big)-K\big)^+ \mid M_t\big),\end{aligned}$$

where $\kappa^{(1)}(t,T^S,T^M) := A_t^{(1),T^S}/B_t^{(1),T^S}$ is the *forward CDS rate*. Note that the random variables $A_t^{(1),T^S}$ and $B_t^{(1),T^S}$ are strictly positive on the set $\{\tau > T\}$ for $t \leqslant T < T^S$, so that $\kappa^{(1)}(t,T^S,T^M)$ enjoys the same property.

5.4.1 Conditionally Gaussian case

We shall now provide a more explicit representation for the value of a CDS swaption. To this end, we fix a Markovian numeraire β and we assume that the forward CDS swap rates $\kappa^{(1)}(t,T^S,T^M)$ are conditionally log-Gaussian under $\mathbb{P}^\beta$ for $t \leqslant T$ (for an example of such a model, see Section 4.3.3). Then we have the following result.

Proposition 5.1. *Suppose that, on the set $\{\tau > T\}$ and for arbitrary $t < t_1 < \cdots < t_n \leqslant T$, the conditional distribution*

$$\begin{aligned}\mathbb{P}^\beta\big(\kappa^{(1)}\big(t_1,T^S,T^M\big) \leqslant k_1, \kappa^{(1)}\big(t_2,T^S,T^M\big)\\ \leqslant k_2,\ldots,\kappa^{(1)}\big(t_n,T^S,T^M\big) \leqslant k_n \,\big|\, \sigma(M_t)\vee\mathcal{F}_T^X\big)\end{aligned}$$

is $\mathbb{P}^\beta$-a.s. log-Gaussian. Let $\sigma(s,T^S,T^M)$, $s \in [t,T]$, denote the conditional volatility of the process $\kappa^{(1)}(s,T^S,T^M)$, $s \in [t,T]$, given the σ-field $\sigma(M_t)\vee\mathcal{F}_T^X$.

Then the price of a CDS swaption equals, for $t < T$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^\beta}\big(\beta_t\beta_T^{-1}\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+ \mid M_t\big)\\
&\quad= \mathbb{E}_{\mathbb{P}^\beta}\Bigg(\mathbb{1}_{\{\tau>T\}}\beta_t\beta_T^{-1}B_T^{(1),T^S}\Bigg[\kappa^{(1)}\big(t, T^S, T^M\big)\\
&\qquad\times N\Bigg(\frac{\log\frac{\kappa^{(1)}(t,T^S,T^M)}{K}}{\upsilon_{t,T}} + \frac{\upsilon_{t,T}}{2}\Bigg)\\
&\qquad- KN\Bigg(\frac{\log\frac{\kappa^{(1)}(t,T^S,T^M)}{K}}{\upsilon_{t,T}} - \frac{\upsilon_{t,T}}{2}\Bigg)\Bigg]\;\Big|\; M_t\Bigg),
\end{aligned}$$

where

$$\upsilon_{t,T}^2 = \upsilon\big(t, T, T^S, T^M\big)^2 := \int_t^T \sigma\big(s, T^S, T^M\big)^2\,\mathrm{d}s.$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^\beta}\big(\beta_t\beta_T^{-1}\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+ \mid M_t\big)\\
&\quad= \mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{\tau>T\}}\beta_t\beta_T^{-1}\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+ \mid M_t\big)\\
&\quad= \mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{\tau>T\}}\beta_t\beta_T^{-1}\mathbb{E}_{\mathbb{P}^\beta}\big(\big(A_T^{(1),T^S} - KB_T^{(1),T^S}\big)^+\big|\sigma(M_t)\vee\mathcal{F}_T^X\big)\big|M_t\big)\\
&\quad= \mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{\tau>T\}}\beta_t\beta_T^{-1}B_T^{(1),T^S}\mathbb{E}_{\mathbb{P}^\beta}\big(\big(\kappa^{(1)}\big(T, T^S, T^M\big) - K\big)^+\\
&\qquad\times\big|\sigma(M_t)\vee\mathcal{F}_T^X\big)\big|M_t\big).
\end{aligned}$$

In view of our assumptions, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^\beta}\big(\big(\kappa^{(1)}\big(T, T^S, T^M\big) - K\big)^+ \mid \sigma(M_t)\vee\mathcal{F}_T^X\big)\\
&\quad= \kappa^{(1)}\big(t, T^S, T^M\big)N\Bigg(\frac{\log\frac{\kappa^{(1)}(t,T^S,T^M)}{K}}{\upsilon_{t,T}} + \frac{\upsilon_{t,T}}{2}\Bigg)\\
&\qquad- KN\Bigg(\frac{\log\frac{\kappa^{(1)}(t,T^S,T^M)}{K}}{\upsilon_{t,T}} - \frac{\upsilon_{t,T}}{2}\Bigg).
\end{aligned}$$

By combining the above equalities, we arrive at the stated formula. □

6 Valuation of basket credit derivatives

In this section, we shall discuss the case of credit derivatives with several underlying credit names. Feasibility of closed-form calculations, such as analytic computation of relevant conditional expected values, depends to a great extent on the type and amount of information one wants to utilize. Typically, in

order to efficiently deal with exact calculations of conditional expectations, one will need to amend specifications of the underlying model so that information used in calculations is given by a coarser filtration, or perhaps by some proxy filtration.

6.1 k th-to-default CDS

We shall now discuss the valuation of a generic *kth-to-default credit default swap* relative to a portfolio of L reference defaultable bonds. The deterministic notional amount of the ith bond is denoted as N_i, and the corresponding deterministic recovery rate equals δ_i. We suppose that the maturities of the bonds are $U_1, U_2, \ldots, U_L$, and the maturity of the swap is $T < \min\{U_1, U_2, \ldots, U_L\}$.

As before, we shall only discuss a vanilla basket CDS written on such a portfolio of corporate bonds under the fractional recovery of par covenant. Thus, in the event that $\tau^{(k)} < T$, the buyer of the protection is paid at time $\tau^{(k)}$ a cumulative compensation

$$\sum_{i\in\mathcal{L}_k}(1-\delta_i)N_i,$$

where $\mathcal{L}_k$ is the (random) set of all reference credit names that defaulted in the time interval $]0, \tau^{(k)}]$. This means that the protection buyer is protected against the cumulative effect of the first k defaults. Recall that, in view of our model assumptions, the possibility of simultaneous defaults is excluded.

6.1.1 Default payment leg

The cash flow associated with the default payment leg is given by the expression

$$\sum_{i\in\mathcal{L}_k}(1-\delta_i)N_i\mathbb{1}_{\{\tau^{(k)}\leqslant T\}}\mathbb{1}_{\tau^{(k)}}(t),$$

so that the time-t value of the default payment leg is equal to

$$A_t^{(k)} = \mathbb{E}_{\mathbb{P}^\beta}\left(\mathbb{1}_{\{t<\tau^{(k)}\leqslant T\}}\beta_t\beta^{-1}_{\tau^{(k)}}\sum_{i\in\mathcal{L}_k}(1-\delta_i)N_i \mid M_t\right).$$

In general, this expectation will need to be evaluated numerically by means of simulations.

A special case of a kth-to-default-swap is when the protection buyer is protected against losses associated with the last default only. In the case of a *last-to-default credit default swap*, the cash flow associated with the default payment leg is given by the expression

$$\begin{aligned}&(1-\delta_{\iota^{(k)}})N_{\iota^{(k)}}\mathbb{1}_{\{\tau^{(k)}\leqslant T\}}\mathbb{1}_{\tau^{(k)}}(t)\\&\quad=\sum_{i=1}^{L}(1-\delta_i)N_i\mathbb{1}_{\{H_{\tau_i}=k\}}\mathbb{1}_{\{\tau^{(i)}\leqslant T\}}\mathbb{1}_{\tau^{(i)}}(t),\end{aligned}$$

where $\iota^{(k)}$ stands for the identity of the kth defaulting credit name. Assuming that the numeraire process β is deterministic, we can represent the value at time t of the default payment leg as follows:

$$\begin{aligned} A_t^{(k)} &= \sum_{i=1}^{L} \mathbb{E}_{\mathbb{P}^\beta}\big(\mathbb{1}_{\{t<\tau_i\leqslant T\}}\mathbb{1}_{\{H_{\tau_i}=k\}}\beta_t\beta_{\tau_i}^{-1}(1-\delta_i)N_i \mid M_t\big) \\ &= \sum_{i=1}^{L} \beta_t(1-\delta_i)N_i \\ &\quad \times \int_t^T \beta_s^{-1}\mathbb{P}^\beta(H_s = k \mid \tau_i = s, M_t)\mathbb{P}^\beta(\tau_i \in \mathrm{d}s \mid M_t). \end{aligned}$$

Note that the conditional probability $\mathbb{P}^\beta(H_s = k \mid \tau_i = s, M_t)$ can be approximated as

$$\mathbb{P}^\beta(H_s = k \mid \tau_i = s, M_t) \approx \frac{\mathbb{P}^\beta(H_s = k, X^i_{s-\epsilon} \neq K, X^i_s = K \mid M_t)}{\mathbb{P}^\beta(X^i_{s-\epsilon} \neq K, X^i_s = K \mid M_t)}.$$

Hence, if the number L of credit names is small, so that the Kolmogorov equations for the conditional distribution of the process (H, X, Y) can be solved, the value of $A_t^{(k)}$ can be approximated analytically.

6.1.2 Premium payment leg

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_J\}$ denote the tenor of the premium payment, where $0 = T_0 < T_1 < \cdots < T_J < T$. If the premium accrual covenant is in force, then the cash flows associated with the premium payment leg admit the following representation:

$$\kappa^{(k)}\left(\sum_{j=1}^{J}\mathbb{1}_{\{T_j<\tau^{(k)}\}}\mathbb{1}_{T_j}(t) + \sum_{j=1}^{J}\mathbb{1}_{\{T_{j-1}<\tau^{(k)}\leqslant T_j\}}\mathbb{1}_{\tau^{(k)}}(t)\frac{t-T_{j-1}}{T_j-T_{j-1}}\right),$$

where $\kappa^{(k)}$ is the CDS premium. Thus, the time-t value of the premium payment leg is $\kappa^{(k)}B_t^{(k)}$, where

$$\begin{aligned} B_t^{(k)} &= \mathbb{E}_{\mathbb{P}^\beta}\left(\mathbb{1}_{\{t<\tau^{(k)}\}}\sum_{j=j(t)}^{N}\frac{\beta_t}{\beta_{T_j}}\mathbb{1}_{\{T_j<\tau^{(k)}\}} \;\Big|\; M_t\right) \\ &\quad + \mathbb{E}_{\mathbb{P}^\beta}\left(\mathbb{1}_{\{t<\tau^{(k)}\}}\sum_{j=j(t)}^{J}\frac{\beta_t}{\beta_{\tau^{(k)}}}\mathbb{1}_{\{T_{j-1}<\tau^{(k)}\leqslant T_j\}}\frac{\tau^{(k)}-T_{j-1}}{T_j-T_{j-1}}|M_t\right), \end{aligned}$$

where $j(t)$ is the smallest integer such that $T_{j(t)} > t$. Again, in general, the above conditional expectation will need to be approximated by simulation. And

again, for a small portfolio size L, if either exact or numerical solution of relevant Kolmogorov equations can be derived, then an analytical computation of the expectation can be done. This is left for a future study.

6.2 Forward kth-to-default CDS

Forward kth-to-default CDS is an analogous structure to the forward CDS. The notation used here is consistent with the notation used previously in Sections 5.3 and 6.1.

6.2.1 Default payment leg

The cash flow associated with the default payment leg can be expressed as follows

$$\sum_{i\in\mathcal{L}_k}(1-\delta_i)N_i\mathbb{1}_{\{T^S<\tau^{(k)}\leqslant T^M\}}\mathbb{1}_{\tau^{(k)}}(t).$$

Consequently, the time-t value of the default payment leg equals, for every $t\leqslant T^S$,

$$A_t^{(k),T^S}=\mathbb{E}_{\mathbb{P}^\beta}\left(\mathbb{1}_{\{T^S<\tau^{(k)}\leqslant T^M\}}\beta_t\beta^{-1}_{\tau^{(k)}}\sum_{i\in\mathcal{L}_k}(1-\delta_i)N_i\;\middle|\;M_t\right).$$

6.2.2 Premium payment leg

As before, let $\mathcal{T}=\{T_1,T_2,\ldots,T_J\}$ be the tenor of a generic premium payment leg, where $T^S<T_1<\cdots<T_J<T^M$. Under the premium accrual covenant, the cash flows associated with the premium payment leg are

$$\kappa^{(k)}\left(\sum_{j=1}^{J}\mathbb{1}_{\{T_j<\tau^{(k)}\}}\mathbb{1}_{T_j}(t)+\sum_{j=1}^{J}\mathbb{1}_{\{T_{j-1}<\tau^{(k)}\leqslant T_j\}}\mathbb{1}_{\tau^{(k)}}(t)\frac{t-T_{j-1}}{T_j-T_{j-1}}\right),$$

where $\kappa^{(k)}$ is the CDS premium. Thus, the time-t value of the premium payment leg is $\kappa^{(k)}B_t^{(k),T^S}$, where

$$\begin{aligned}B_t^{(k),T^S}=\mathbb{E}_{\mathbb{P}^\beta}\Bigg(&\mathbb{1}_{\{t<\tau^{(k)}\}}\Bigg[\sum_{j=1}^{N}\frac{\beta_t}{\beta_{T_j}}\mathbb{1}_{\{T_j<\tau\}}\\&+\sum_{j=1}^{J}\frac{\beta_t}{\beta_\tau}\mathbb{1}_{\{T_{j-1}<\tau^{(k)}\leqslant T_j\}}\frac{\tau-T_{j-1}}{T_j-T_{j-1}}\Bigg]\;\Bigg|\;M_t\Bigg).\end{aligned}$$

7 Model implementation

The last section is devoted to a brief discussion of issues related to the model implementation.

7.1 Curse of dimensionality

When one deals with basket products involving multiple credit names, direct computations may not be feasible. The cardinality of the state space $\mathbf{K}$ for the migration process X is equal to K^L. Thus, for example, in case of $K = 18$ rating categories, as in Moody's ratings,[1] and in case of a portfolio of $L = 100$ credit names, the state space $\mathbf{K}$ has 18^{100} elements.[2] If one aims at closed-form expressions for conditional expectations, but K is large, then it will typically be infeasible to work directly with information provided by the state vector $(X, Y) = (X^1, X^2, \ldots, X^L, Y)$ and with the corresponding generator $\mathbf{A}$. A reduction in the amount of information that can be effectively used for analytical computations will be needed. Such reduction may be achieved by reducing the number of distinguished rating categories – this is typically done by considering only two categories: pre-default and default. However, this reduction may still not be sufficient enough, and further simplifying structural modifications to the model may need to be called for. Some types of additional modifications, such as *homogeneous grouping* of credit names and the *mean-field interactions* between credit names, are discussed in Frey and Backhaus (2004).[3]

7.2 Recursive simulation procedure

When closed-form computations are not feasible, but one does not want to give up on potentially available information, an alternative may be to carry approximate calculations by means of either approximating some involved formulae and/or by simulating sample paths of underlying random processes. This is the approach that we opt for.

In general, a simulation of the evolution of the process X will be infeasible, due to the curse of dimensionality. However, the structure of the generator $\mathbf{A}$ that we postulate (cf. (13)) makes it so that simulation of the evolution of process X reduces to recursive simulation of the evolution of processes X^l whose state spaces are only of size K each. In order to facilitate simulations even further, we also postulate that each migration process X^l behaves like a birth-and-death process with absorption at default, and with possible jumps to default from every intermediate state (cf. Section 4.3.1). Recall that $X_t^{(l)} = (X_t^1, \ldots, X_t^{l-1}, X_t^{l+1}, \ldots, X_t^L)$. Given the state $(x^{(l)}, y)$ of the process $(X^{(l)}, Y)$, the intensity matrix of the lth migration process is sub-stochastic and

[1] We think here of the following Moody's rating categories: Aaa, Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, Baa3, Ba1, Ba2, Ba3, B1, B2, B3, Caa, D(efault).

[2] The number known as *Googol* is equal to 10^{100}. It is believed that this number is greater than the number of atoms in the entire observed Universe.

[3] Homogeneous grouping was also introduced in Bielecki (2003).

is given as:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \\ K \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & \cdots & K-1 & K \end{array} \\ \begin{pmatrix} \lambda^l(1,1) & \lambda^l(1,2) & 0 & \cdots & 0 & \lambda^l(1,K) \\ \lambda^l(2,1) & \lambda^l(2,2) & \lambda^l(2,3) & \cdots & 0 & \lambda^l(2,K) \\ 0 & \lambda^l(3,2) & \lambda^l(3,3) & \cdots & 0 & \lambda^l(3,K) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda^l(K-1,K-1) & \lambda^l(K-1,K) \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \end{array},$$

where we set $\lambda^l(x^l, x'^l) = \lambda^l(x, x'_l; y)$. Also, we find it convenient to write $\lambda^l(x^l, x'^l; x^{(l)}, y) = \lambda^l(x, x'_l; y)$ in what follows.

Then the diagonal elements are specified as follows, for $x^l \neq K$,

$$\begin{aligned} \lambda^l(x, x; y) = &-\lambda^l\big(x^l, x^l - 1; x^{(l)}, y\big) - \lambda^l\big(x^l, x^l + 1; x^{(l)}, y\big) \\ &- \lambda^l\big(x^l, K; x^{(l)}, y\big) \\ &- \sum_{i \neq l} \big(\lambda^i\big(x^i, x^i - 1; x^{(i)}, y\big) + \lambda^i\big(x^i, x^i + 1; x^{(i)}, y\big) \\ &+ \lambda^i\big(x^i, K; x^{(i)}, y\big)\big) \end{aligned}$$

with the convention that $\lambda^l(1, 0; x^{(l)}, y) = 0$ for every $l = 1, 2, \ldots, L$.

It is implicit in the above description that $\lambda^l(K, x^l; x^{(l)}, y) = 0$ for any $l = 1, 2, \ldots, L$ and $x^l = 1, 2, \ldots, K$. Suppose now that the current state of the process (X, Y) is (x, y). Then the intensity of a jump of the process X equals

$$\lambda(x, y) := -\sum_{l=1}^{L} \lambda^l(x, x; y).$$

Conditional on the occurrence of a jump of X, the probability distribution of a jump for the component X^l, $l = 1, 2, \ldots, L$, is given as follows:

- probability of a jump from x^l to $x^l - 1$ equals $p^l(x^l, x^l - 1; x^{(l)}, y) := \frac{\lambda^l(x^l, x^l - 1; x^{(l)}, y)}{\lambda(x, y)}$,
- probability of a jump from x^l to $x^l + 1$ equals $p^l(x^l, x^l + 1; x^{(l)}, y) := \frac{\lambda^l(x^l, x^l + 1; x^{(l)}, y)}{\lambda(x, y)}$,
- probability of a jump from x^l to K equals $p^l(x^l, K; x^{(l)}, y) := \frac{\lambda^l(x^l, K; x^{(l)}, y)}{\lambda(x, y)}$.

As expected, we have that

$$\sum_{l=1}^{L}\big(p^l\big(x^l, x^l-1; x^{(l)}, y\big) + p^l\big(x^l, x^l+1; x^{(l)}, y\big) + p^l\big(x^l, K; x^{(l)}, y\big)\big) = 1.$$

For a generic state $x = (x^1, x^2, \ldots, x^L)$ of the migration process X, we define the *jump space* $\mathcal{J}(x) = \bigcup_{l=1}^{L}\{(x^l - 1, l), (x^l + 1, l), (K, l)\}$ with the convention that $(K+1, l) = (K, l)$. The notation (a, l) refers to the lth component of X. Given that the process (X, Y) is in the state (x, y), and conditional on the occurrence of a jump of X, the process X jumps to a point in the jump space $\mathcal{J}(x)$ according to the probability distribution denoted by $p(x, y)$ and determined by the probabilities p^l described above. Thus, if a random variable J has the distribution given by $p(x, y)$ then, for any $(x'^l, l) \in \mathcal{J}(x)$, we have that $Prob(J = (x'^l, l)) = p^l(x^l, x'^l; x^{(l)}, y)$.

7.2.1 Simulation algorithm: special case

We shall now present in detail the case when the dynamics of the factor process Y do not depend on the credit migrations process X. The general case appears to be much harder.

Under the assumption that the dynamics of the factor process Y do not depend on the process X, the simulation procedure splits into two steps. In Step 1, a sample path of the process Y is simulated; then, in Step 2, for a given sample path Y, a sample path of the process X is simulated. We consider here simulations of sample paths over some generic time interval, say $[t_1, t_2]$, where $0 \leqslant t_1 < t_2$. We assume that the number of defaulted names at time t_1 is less than k, that is $H_{t_1} < k$. We conduct the simulation until the kth default occurs or until time t_2, whichever occurs first.

Step 1: The dynamics of the factor process are now given by the SDE

$$\mathrm{d}Y_t = b(Y_t)\,\mathrm{d}t + \sigma(Y_t)\,\mathrm{d}W_t + \int_{\mathbb{R}^n} g(Y_{t-}, y)\pi(Y_{t-}; \mathrm{d}y, \mathrm{d}t),$$

$$t \in [t_1, t_2].$$

Any standard procedure can be used to simulate a sample path of Y (the reader is referred, for example, to Kloeden and Platen, 1995). Let us denote by $\widehat{Y}$ the simulated sample path of Y.

Step 2: Once a sample path of Y has been simulated, simulate a sample path of X on the interval $[t_1, t_2]$ until the kth default time.

We exploit the fact that, according to our assumptions about the infinitesimal generator **A**, the components of the process X do not jump simultaneously. Thus, the following algorithm for simulating the evolution of X appears to be feasible:

Step 2.1: Set the counter $n = 1$ and simulate the first jump time of the process X in the time interval $[t_1, t_2]$. Towards this end, simulate first a value, say $\hat{\eta}_1$, of a unit exponential random variable η_1. The simulated value of the first jump time, τ_1^X, is then given as

$$\hat{\tau}_1^X = \inf\left\{ t \in [t_1, t_2]: \int_{t_1}^{t} \lambda\big(X_{t_1}, \widehat{Y}_u\big)\, \mathrm{d}u \geqslant \hat{\eta}_1 \right\},$$

where by convention the infimum over an empty set is $+\infty$. If $\hat{\tau}_1^X = +\infty$, set the simulated value of the kth default time to be $\hat{\tau}^{(k)} = +\infty$, stop the current run of the simulation procedure and go to Step 3. Otherwise, go to Step 2.2.

Step 2.2: Simulate the jump of X at time $\hat{\tau}_1^X$ by drawing from the distribution $p(X_{t_1}, \widehat{Y}_{\hat{\tau}_1^X -})$ (cf. discussion in Section 7.2). In this way, one obtains a simulated value $\widehat{X}_{\hat{\tau}_1^X}$, as well as the simulated value of the number of defaults $\widehat{H}_{\hat{\tau}_1^X}$. If $\widehat{H}_{\hat{\tau}_1^X} < k$ then let $n := n + 1$ and go to Step 2.3; otherwise, set $\hat{\tau}^{(k)} = \hat{\tau}_1^X$ and go to Step 3.

Step 2.3: Simulate the nth jump of process X. Towards this end, simulate a value, say $\hat{\eta}_n$, of a unit exponential random variable η_n. The simulated value of the nth jump time τ_n^X is obtained from the formula

$$\hat{\tau}_n^X = \inf\left\{ t \in [\hat{\tau}_{n-1}^X, t_2]: \int_{\hat{\tau}_{n-1}^X}^{t} \lambda(X_{\hat{\tau}_{n-1}^X}, \widehat{Y}_u)\, \mathrm{d}u \geqslant \hat{\eta}_n \right\}.$$

In case $\hat{\tau}_n^X = +\infty$, let the simulated value of the kth default time to be $\hat{\tau}^{(k)} = +\infty$; stop the current run of the simulation procedure, and go to Step 3. Otherwise, go to Step 2.4.

Step 2.4: Simulate the jump of X at time $\hat{\tau}_n^X$ by drawing from the distribution $p(X_{\hat{\tau}_{n-1}^X}, \widehat{Y}_{\hat{\tau}_n^X -})$. In this way, produce a simulated value $\widehat{X}_{\hat{\tau}_n^X}$, as well as the simulated value of the number of defaults $\widehat{H}_{\hat{\tau}_n^X}$. If $\widehat{H}_{\hat{\tau}_n^X} < k$, let $n := n + 1$ and go to Step 2.3; otherwise, set $\hat{\tau}^{(k)} = \hat{\tau}_n^X$ and go to Step 3.

Step 3: Calculate a simulated value of a relevant functional. For example, in case of the kth-to-default CDS, compute

$$\widehat{A}_{t_1}^{(k)} = \mathbb{1}_{\{t_1 < \hat{\tau}^{(k)} \leqslant T\}} \hat{\beta}_{t_1} \hat{\beta}_{\hat{\tau}^{(k)}}^{-1} \sum_{i \in \widehat{\mathcal{L}}_k} (1 - \delta_i) N_i \tag{25}$$

and

$$\hat{\beta}_{t_1}^{(k)} = \sum_{j=j(t_1)}^{N} \frac{\hat{\beta}_{t_1}}{\hat{\beta}_{T_j}} \mathbb{1}_{\{T_j < \hat{\tau}^{(k)}\}} + \sum_{j=j(t_1)}^{J} \frac{\hat{\beta}_{t_1}}{\hat{\beta}_{\hat{\tau}^{(k)}}} \mathbb{1}_{\{T_{j-1} < \hat{\tau}^{(k)} \leqslant T_j\}} \frac{\hat{\tau}^{(k)} - T_{j-1}}{T_j - T_{j-1}}, \tag{26}$$

where, as usual, the 'hat' indicates that we deal with simulated values.

7.3 Estimation and calibration of the model

Our market model (13) has the same structure under either the pricing probability measure or the statistical measure. The model parameters corresponding to the two measures (or any other two measures for that matter) are related via (19).

Estimation of the statistical parameters of the model, that is, the parameters corresponding to the statistical measure, can be split into two separate problems – the estimation of the dynamics of the factor process Y, and the estimation of the transition intensities of the process X. With regard to the former: typically, the estimation of parameters of the drift function and the estimation of parameters of the Poisson measure is not easy; the estimation of parameters of the volatility function $\sigma(x, y)$ is rather straightforward, as it can be done via estimation of the quadratic variation process of the diffusion component. Estimates of parameters involved in the transition intensities can, in principle, be obtained from the statistical estimates of transition probability matrices that are produced by major rating agencies.

Calibration of the pricing parameters of the model, that is, the parameters corresponding to the pricing measure, depends on the types of the market quotes data used for calibration. Since, in case of basket credit derivatives, we typically will not have access to closed-form pricing formulae, the calibration of the model parameters will need to be done via simulation. For example, if the model is calibrated to market quotes for the kth-to-default basket swaps, in order to select the best fitted model, we shall use simulated averages of expressions (25) and (26) obtained for various parametric settings. Then, the market prices of credit risk can be obtained from estimated and calibrated values of the parameters and from formula (19). We shall deal with the issues of model estimation and calibration in a future work, which will be devoted to model implementation. Some significant progress in this direction has already been made in Bielecki et al. (2006).

7.4 Portfolio credit risk

The issue of evaluating functionals associated with multiple credit migrations, defaults in particular, is also prominent with regard to portfolio credit risk. In some segments of the credit markets, only the deterioration of the value of a portfolio of debts (bonds or loans) due to defaults is typically considered. In fact, such is the situation regarding various tranches of (cash or synthetic) collateralized debt obligations (CDOs), as well as with various tranches of recently introduced CDS indices, such as, DJ CDX NA IG or DJ iTraxx Europe.[4]

[4] See http://www.creditflux.com/public/publications/0409CFindexGuide.pdf.

Nevertheless, it is rather apparent that a valuation model reflecting the possibility of intermediate credit migrations, and not only defaults, is called for in order to better account for changes in creditworthiness of the reference credit names. Likewise, for the purpose of managing risks of a debt portfolio, it is necessary to account for changes in value of the portfolio due to changes in credit ratings of the components of the portfolio.

The problem of valuation of tranches of a CDO (or tranches of a CDS index) is closely related to the problem of valuation of the kth-to-default swap. In a future work, we shall focus on implementation of our model to all these problems. It is perhaps worth mentioning though that we have already done some numerical tests of our model so to see whether the model can reproduce so called market correlation skews. Figure 1 shows that the model performs very well in this regard.[5]

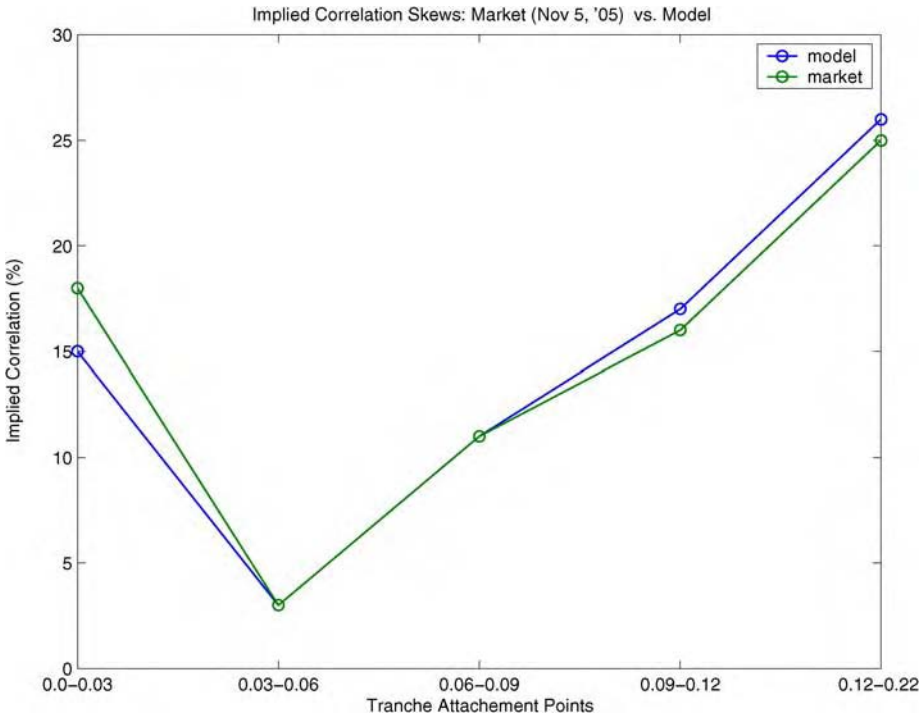


Fig. 1.

[5] We thank Andrea and Luca Vidozzi from Applied Mathematics Department at the Illinois Institute of Technology for numerical implementation of the model and, in particular, for generating the picture.

References

Albanese, C., Campolieti, J., Chen, O., Zavidonov, A. (2003). Credit barrier model. *Risk* 16 (6), 109–113.

Albanese, C., Chen, O. (2004a). Discrete credit barrier models. *Quantitative Finance* 5, 247–256.

Albanese, C., Chen, O. (2004b). Pricing equity default swaps. *Risk* 18 (6), 83–87.

Becherer, D., Schweizer, M. (2003). Classical solutions to reaction-diffusion systems for hedging problems with interacting Itô and point processes. *Annals of Applied Probability* 15, 1111–1144.

Bielecki, T.R. (2003). A multivariate Markov model for simulating dependent credit migrations. *Working paper*.

Bielecki, T.R., Rutkowski, M. (2002a). *Credit Risk: Modeling, Valuation and Hedging*. Springer-Verlag, Berlin/Heidelberg/New York.

Bielecki, T.R., Rutkowski, M. (2002b). Intensity-based valuation of basket credit derivatives. In: Yong, J. (Ed.), *Mathematical Finance*. World Scientific, Singapore, pp. 12–27.

Bielecki, T.R., Rutkowski, M. (2003). Dependent defaults and credit migrations. *Applicationes Mathematicae* 30, 121–145.

Bielecki, T.R., Vidozzi, A., Vidozzi, L. (2006). An efficient approach to valuation of basket credit products and options on ratings triggered step-up bonds. *Working paper*. IIT.

Chen, L., Filipović, D. (2003). Pricing credit default swaps with default correlation and counterparty risk. *Working paper*.

Chen, L., Filipović, D. (2005). Simple model for credit migration and spread curves. *Finance and Stochastics*, in press.

Davis, M., Esparragoza, J.C. (2004). A queueing network approach to portfolio credit risk. *Working paper*. Imperial College, London.

Douady, R., Jeanblanc, M. (2002). A rating-based model for credit derivatives. *European Investment Review* 1, 17–29.

Ethier, H.J., Kurtz, T.G. (1986). *Markov Processes. Characterization and Convergence*. Wiley, New York.

Frey, R., Backhaus, J. (2004). Portfolio credit risk models with interacting default intensities: A Markovian approach. *Working paper*.

Frey, R., McNeal, A. (2003). Dependent defaults in models of portfolio credit risk. *Journal of Risk* 6 (1), 59–92.

Giesecke, K. (2004). Correlated default with incomplete information. *Journal of Banking and Finance* 28, 1521–1545.

Giesecke, K., Weber, S. (2002). Credit contagion and aggregate losses. *Journal of Economic Dynamics and Control* 30, 741–767.

Gregory, J., Laurent, J.-P. (2004). Analytical approaches to the pricing and risk management of basket credit derivatives and CDOs. *Journal of Risk* 7 (4), 103–122.

Hull, J.C., White, A. (2001). Valuing credit default swaps II. Modeling default correlations. *Journal of Derivatives* 8 (3), 12–22.

Jamshidian, F. (2004). Valuation of credit default swap and swaptions. *Finance and Stochastics* 8, 343–371.

Jarrow, R.A., Yu, F. (2001). Counterparty risk and the pricing of defaultable securities. *Journal of Finance* 56, 1765–1799.

Kloeden, P.E., Platen, E. (1995). *Numerical Solution of Stochastic Differential Equations*, *second ed*. Springer, Berlin/Heidelberg/New York.

Kunita, H., Watanabe, S. (1963). Notes on transformations of Markov processes connected with multiplicative functionals. *Memoirs of the Faculty of Science, Kyushu University A* 17, 181–191.

Laurent, J.-P., Gregory, J. (2003). Basket default swaps, CDOs and factor copulas. *Working paper*.

Li, D. (2000). On default correlation: A copula function approach. *Journal of Fixed Income* 9, 43–54.

Palmowski, Z., Rolski, T. (2002). A technique for exponential change of measure for Markov processes. *Bernoulli* 8 (6), 767–785.

Schönbucher, P.J., Schubert, D. (2001). Copula-dependent default risk in intensity models. *Working paper*. University of Bonn.

Schmock, U., Seiler, D. (2002). Modeling dependent credit risks with mixture models. *Working paper*.

Yu, F. (2007). Correlated defaults in intensity-based models. *Mathematical Finance* 17, 155–173.

This page intentionally left blank

PART IV

Incomplete Markets

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15012-X

Chapter 12
Incomplete Markets

Jeremy Staum

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3119, USA
E-mail: j-staum@northwestern.edu
url: http://users.iems.northwestern.edu/∼staum

Abstract

In reality, markets are *incomplete*, meaning that some payoffs cannot be replicated by trading in marketed securities. The classic no-arbitrage theory of valuation in a complete market, based on the unique price of a self-financing replicating portfolio, is not adequate for nonreplicable payoffs in incomplete markets. We focus on pricing over-the-counter derivative securities, surveying many proposed methodologies, drawing relationships between them, and evaluating their promise.

1 Introduction

Incomplete markets are those in which perfect risk transfer is not possible. Despite the ever-increasing sophistication of financial and insurance markets, markets remain significantly incomplete, with important consequences for their participants: workers and homeowners remain exposed to risks involving labor income, property value, and taxes, investors and portfolio managers have limited choices, and traders of derivative securities must bear residual risks. From a theoretical perspective, incomplete markets complicate the study of financial market equilibrium, portfolio optimization, and derivative securities.

Although the theory of derivative securities in complete markets is understood very well, and is the subject of numerous textbook accounts, there is as yet no fully developed, sound theoretical framework for pricing derivative securities in incomplete markets. This has profound consequences for the practice of trading, speculating, and hedging with derivative securities. This chapter surveys the topic of incomplete markets, with an emphasis on pricing and hedging derivative securities.

Other surveys have treated different aspects of incomplete markets. For portfolio optimization in incomplete markets, see Skiadas (2006). The finance

literature emphasizes the existence and characteristics of equilibria, including market efficiency. Magill and Quinzii (1996) offer a book-length exposition, and Hens (1998) provides an overview with a low level of technicalities. Appendix B presents perspectives from the finance literature, not usually addressed in financial engineering, on the degree to which markets are actually incomplete, and the implications for welfare.

Surveys of derivative security pricing in incomplete markets include Jouini (2001), who covers no-arbitrage bounds, utility maximization, and equilibrium valuation, as an introduction to a special journal issue on these topics. Cont and Tankov (2004, Chapter 10) cover these approaches and others, including quadratic and entropy criteria, as well as calibration. Another survey is by Davis (2004b), whose "intention is not to aim at a maximum level of generality but, on the contrary, to concentrate on specific cases and solved problems which give insight into the nature of optimal strategies for hedging and investment." In contrast, we will cover all major approaches to pricing derivative securities in incomplete markets, as well as providing enough background to evaluate them and understand them in relation to one another.

Thus, due to limitation of space, we will not be able to concentrate on specific derivative securities or models of markets, although we will give simple examples that illustrate major ideas. Likewise, we can neither recount the development of each method nor provide an exhaustive list of references, so many significant papers will not be mentioned. Instead, we will merely provide references to the literature as a substitute for an exposition of the technical details of all the methods we survey, for which there is also not space. However, we address the technicalities of defining incompleteness in Appendix A.

We begin with background for the problem of incomplete markets. In Section 2, there is a description of the over-the-counter market for derivative securities and the financial engineering problems we will address. The causes of incomplete markets are addressed in Section 3. Next, we turn to general theoretical considerations about pricing in incomplete markets. The connections between pricing and optimization occupy Section 4, which covers no-arbitrage bounds, indifference prices, good deal bounds, and minimum-distance pricing measures. In Section 5, simple examples based on expected utility illustrate issues in pricing and optimization. Subsequent sections are devoted to various particular methods. The quadratic approach to hedging occupies Section 6. Exponential utility, with its connection to relative entropy, is the topic of Section 7. Several methods based on considering only losses, not gains, appear in Section 8: these include partial replication schemes such as quantile hedging. Restrictions on pricing kernels, including methods based on low-distance pricing kernels, are covered in Section 9. Ambiguity and robustness to model risk is the topic of Section 10. The standard practice of calibrating a model to market prices occupies Section 11. In Section 12, we offer some conclusions, evaluation, and directions for future research.

2 The over-the-counter market

Let us imagine ourselves in the position of a market-maker in an over-the-counter (OTC) derivatives market. Throughout this survey, we will consider incomplete markets from the market-maker's perspective, focusing on the financial engineering of solving the problems of pricing and risk management. The same considerations apply to customers in the OTC market.

2.1 The workings of OTC markets

Although some derivative securities, including some stock options and commodity or currency futures, are listed on exchanges and traded in the same manner as the underlying securities, many are not. A hedger or speculator who wishes to trade them must participate in the OTC market by calling OTC market-makers, usually at investment banks, and requesting a quote for bid and ask prices at which the market-makers are willing to buy or sell, respectively. Duffie et al. (2006) address the relationship of frictions and liquidity in OTC markets to valuation. We will focus on the process by which market-makers prepare these prices. Through an analogous process, the potential customer must then decide whether to sell at the highest of the quoted bid prices, buy at the lowest of the quoted ask prices, or do nothing.

If the customer indeed transacts a deal with a market-maker, the market-maker must bear risk associated with this trade, because markets are incomplete. In order to measure the risk of his portfolio and manage it through hedging, he needs to model the future value of the OTC derivatives he has traded. As time passes, he must track the profit or loss generated by his hedged portfolio, based on values of OTC derivatives updated in light of current market prices, a process known as *marking to market*. It is a matter of debate among practitioners whether and when it is appropriate to mark to market using a bid price, a "mid-market price" between the bid and ask prices, or an "unwind price" at which the derivative might be sold. Marking to market is not studied enough relative to pricing, but the risk-adjusted value processes of Artzner et al. (2007) may be useful in this regard.

Establishing bid and ask prices for an OTC derivative security is not the same as determining the equilibrium price for a new security if it were to be listed on an exchange, which is another goal often considered in the literature on incomplete markets. Determining the equilibrium price is more difficult than it might seem, because introducing a new security could alter the existing security prices (Boyle and Wang, 2001). In financial engineering, although equilibrium concepts may be useful in pricing, it is too ambitious to attempt to construct an entire equilibrium, based on the preferences and endowments of all participants. This is more appropriate in finance, where one may use a simplified model to formulate a hypothesis or explain some phenomenon.

2.2 Standard practice

Whether using a model in which markets are complete or incomplete, derivatives traders know that markets are actually incomplete, and that after trading, they will not be able to hedge away all the risk, to which they are averse. Nonetheless, their standard practice is to assign prices to OTC derivative securities primarily on the basis of consistency with the market prices of underlying and other derivative securities. We will further discuss and evaluate this standard practice in Section 11.

According to the classic theory of financial engineering, in a complete market, the unique no-arbitrage price of a derivative security whose payoff is X is the expected discounted payoff $\mathrm{E}_{\mathbf{Q}}[DX]$ under the risk-neutral probability measure $\mathbf{Q}$, under which the marketed securities' expected returns equal the risk-free rate of interest. Traders calibrate the parameters of $\mathbf{Q}$ to prices of marketed securities so as to minimize the discrepancy between these market prices and the prices given by the model, i.e. the expected discounted payoffs. To recoup their business expenses and to earn compensation for bearing the risks that they will not be able to hedge, traders establish a bid–ask interval around the expected discounted payoff. The exact level of the bid and ask depend on informal consideration of several factors, such as how the trade will affect the portfolio's Greeks, the trader's outlook on likely market events, what the competition is charging, and the relationship with the customer. One of the major challenges facing financial engineering in the area of derivative securities is to establish a sound basis for this pricing decision, based on quantitative risk assessment using models of incomplete markets.

If the market is incomplete, then pricing by calibration of a complete-market model does not systematically account for the costs of hedging or the risks that remain after hedging. This approach wrongly prices the unhedgeable part of the risk as though it too could be hedged away; it assigns to a derivative security the same price as a fictitious replicating portfolio strategy, when this strategy will not actually succeed in replicating the target payoff. As Foldes (2000) says,

> Enthusiasm for methods of hedging and valuation of derivatives in complete markets, and for associated methods of computation, seems often to obscure the fact that these techniques do not provide a general theory of valuation and that they are liable to give at best only imprecise results when applied beyond their proper domain.

The need to quantify and value residual risks motivates the search for a practical method of pricing with incomplete-markets models.

2.3 The apparent and real problems

The apparent problem of pricing in incomplete markets is mathematical: given the statistical probability measure $\mathbf{P}$, there is a set $\mathcal{Q}$ of *equivalent martingale measures* (EMMs) such that the expected discounted payoff $\mathrm{E}_{\mathbf{Q}}[DX]$ is

an arbitrage-free price for X.[1] There is an interval

$$\left(\inf_{\mathbf{Q}\in\mathcal{Q}} \mathrm{E}_{\mathbf{Q}}[DX],\ \sup_{\mathbf{Q}\in\mathcal{Q}} \mathrm{E}_{\mathbf{Q}}[DX]\right) \tag{1}$$

of arbitrage-free prices for X, and it is usually too wide for these *no-arbitrage bounds* (Section 4.2.1) to serve as useful bid and ask prices (see e.g. Eberlein and Jacod, 1997). The problem may appear to be that we want a way of choosing one of the pricing measures $\mathbf{Q} \in \mathcal{Q}$, so that we may then assign the unique price $\mathrm{E}_{\mathbf{Q}}[DX]$ to each payoff X.

Another way to view the situation is that the no-arbitrage criterion allows a multiplicity of possible *pricing kernels* Π. A pricing kernel $\Pi = D\,\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$ where $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$ is the likelihood ratio, i.e. Radon–Nikodym derivative, between some $\mathbf{Q} \in \mathcal{Q}$ and $\mathbf{P}$. The value $\Pi(\omega)$ of the pricing kernel in state ω can be interpreted as the price now for \$1 to be paid if state ω occurs. With no restriction on the pricing kernel, the price can be anywhere within the no-arbitrage price bounds. However, some of these pricing kernels may seem implausible from an economic perspective. See Section 9 for methodologies that work by eliminating implausible pricing kernels.

The real problem of pricing in incomplete markets depends on the objective. For example, a goal in setting bid and ask prices is to ensure that any trade undertaken at these prices is advantageous to the firm. A grounding of the pricing scheme in financial economics would be desirable. It is not clear how selecting a single pricing measure $\mathbf{Q} \in \mathcal{Q}$ will accomplish this goal; indeed, given a unique price $\mathrm{E}_{\mathbf{Q}}[DX]$, further considerations would be required to generate distinct bid and ask prices. Another objective is marking to market, in which the goal is to assign to the firm's portfolio of derivative securities a value, not a price, that is accurate from an accounting or actuarial perspective. Again, for risk management, there may be different goals that involve assessing the future value of derivative securities. However, in all cases, we want a methodology that respects the no-arbitrage bounds, is computationally efficient, and is robust to those errors that are likely in specifying its inputs, e.g. to stale prices of marketed securities, or to estimation error of statistical probabilities.

In constructing bid and ask prices, the difficulty posed by incomplete markets is more significant than it might at first seem, because of adverse selection. If the ask price is too high, few potential customers will be willing to pay so much, and the result is forgone profits. If the ask price is too low, the resulting trade is bad for the firm and good for the customer, which entices many customers to make trades that entail likely loss for the firm. For example, Dunbar (2005) describes an incident in which it was thought that a large portion of a \$200 million loss by JP Morgan could be attributed to this adverse selection:

> ... by selling a swaption straddle that expired the day before a non-farm payroll announcement and buying one that expired immediately after, a hedge fund

[1] For precise details, see Appendix A. We assume an arbitrage-free market.

> could profit from [the] potential volatility. However, a dealer on the other side of this one-day calendar spread trade might find it difficult to hedge its position over such a short interval of time, and ought to price this risk into the trade, or not undertake the trade at all. But JP Morgan seemed to lack such caution, say market sources, and in effect offered 'lottery tickets' to the market.

As we see from this example, a calibrated model can assign a wrong price. The price for the calendar spread was consistent with market prices, but did not account for an unusual feature of the statistical probability measure **P**: interest rate volatility is concentrated at the date of an important news announcement. See Section 11 for further discussion.

3 Causes of incompleteness

Several phenomena cause incompleteness. One is an insufficiency of marketed assets relative to the class of risks that one wishes to hedge, which may involve jumps or volatility of asset prices, or variables that are not derived from market prices. Market frictions, such as transaction costs and constraints on portfolios, may also cause incompleteness. A source of effective incompleteness is ambiguity, i.e. ignorance of the true stochastic model for market prices: it is effectively the same if it is impossible to transfer risk perfectly or if one merely does not know how to do so.

3.1 Insufficient span of marketed assets

Markets are incomplete with respect to payoffs that are not entirely determined by market prices: examples include weather derivatives, catastrophe bonds, and derivatives written on economic variables such as gross domestic product. Corporate investment projects provide another example; real options analysis applies to a valuation problem in an incomplete market.

Features such as jumps and stochastic volatility of marketed asset prices may also cause incompleteness, depending on the available trading opportunities. For example, in the Heston (1993) model of a stock with stochastic volatility and a bond with constant interest rate, the market is incomplete because it is not possible to hedge the risk factor associated with stochastic volatility. However, if an option on the stock were also to be marketed, both risk factors could be hedged by trading in stock and option, and the market would be complete. Jumps tend to cause incompleteness except in very simple or unusual models (see e.g. Dritschel and Protter, 1999). Whereas in the Black–Scholes model, delta is the hedge ratio that matches the locally linear dependence of an option's value on infinitesimal changes in the stock price, it is not so easy to hedge against potential jumps of various sizes, because value is not linear. To complete a market in which jumps of all sizes are possible might require many more marketed securities, for example, vanilla European options of all strikes and maturities.

Jumps and stochastic volatility are important as ways to model volatility smiles. A primary alternative is a local volatility model, in which the market is complete. However, the local volatility model is criticized (e.g. by Davis, 2004a, §2a) for the crucial, counterfactual assumption that an asset's volatility is a function of its price: precisely the absence of a second risk factor, which makes the model complete, prevents it from saying anything about volatility risk and vega hedging.

For evidence that it may be necessary to model jumps, or jumps and stochastic volatility, in describing equity or equity index returns adequately, see Andersen et al. (2002), Carr et al. (2002). The most realistic models imply incomplete markets.

3.2 Market frictions

Constraints produce incompleteness by forbidding portfolio strategies that replicate some payoffs. For example, an executive who is granted stock options is not supposed to hedge them by selling stock in the company. Different interest rates for borrowing and lending may be modeled by constraints: where $r_b > r_\ell$ are the rates for borrowing and lending respectively, only positive shares of a money market account paying rate r_ℓ and negative shares of one paying rate r_b are allowed.

Transaction costs produce incompleteness less straightforwardly. Continuous-time portfolio strategies accrue transaction costs at every instant the portfolio is rebalanced. These strategies are effectively forbidden if their costs are infinite, which can happen, for instance, in the Black–Scholes model because of the infinite first variation of geometric Brownian motion. Fixed and proportional transaction costs are the most frequently studied; the latter are equivalent to bid–ask spreads for marketed securities. There is a substantial literature on the topic, looking back to Hodges and Neuberger (1989). More recent work on the topic includes Clewlow and Hodges (1997).

Rather than model transaction costs explicitly, one might use a model in which trading is allowed only at a fixed, discrete set of times. This also eliminates continuous-time strategies that would incur infinite costs, and it can be more tractable; however, rebalancing the portfolio at fixed times is typically not as good as rebalancing at a finite number of random times.

3.3 Ambiguity

Suppose a stock index follows a geometric Brownian motion whose volatility is known to be 20%. How many years' data are required to construct a 95% two-sided confidence interval of width 1% for the drift? The answer is 6,147: this yields a width of approximately $2 \times 1.96 \times 20\%/\sqrt{6147} = 1\%$. On the other hand, according to this Black–Scholes model, knowledge of the drift is unnecessary for option pricing, and the volatility can be estimated perfectly by

observing any time interval, no matter how short. This has to do with the non-equivalence of Black–Scholes models with different volatilities, but it is merely an artifact of the continuous-time model. In reality, estimating volatility from high-frequency data is quite difficult (Zhang et al., 2005). Moreover, a cursory examination of financial time series shows that, for instance, daily historical volatility has varied dramatically from year to year. Ambiguity about volatility is so important that, according to Carr (2002), a frequently asked question in option pricing is whether one should hedge at historical or implied volatility. Carr (2002, §IX) provides a formula for the error resulting from hedging a derivative security at the wrong volatility, given a diffusion model. The hedging error can be quite substantial.

4 Pricing and optimization

Pricing can be grounded in portfolio optimization (Sections 4.1–4.2) or in an optimization over pricing measures (Section 4.4).

4.1 Portfolio optimization

Conditions for the existence of optimal portfolio strategies and related probability measures have attracted much attention. There may be no optimal strategy or measure if there is a sequence of them converging to a limit point that is excluded from the feasible set, or if the optimization problem is unbounded. If the limiting strategy is infeasible, one may be satisfied to choose a nearly optimal strategy. When the problem is unbounded, usually something is wrong with the way it has been posed. For example, if there is no bound on the expected utility one can attain by investing, it may be that the set of allowed strategies is unrealistically large, the utility function is unsuitable, or the probability measure is erroneous.

In the interests of simplicity, we will not treat the question of the existence of an optimal solution: the interested reader can find precise results in the literature cited in the sections on specific methodologies. We will also speak primarily of optimizing random wealth at a fixed future date, and the connected problem of pricing payoffs at that date, although the same ideas apply to continuous consumption streams, American options, etc. We ignore the structure of the portfolio strategies, which could be a single vector of weights determining a static portfolio in a one-period problem, or a continuous-time vector stochastic process, or something in between, focusing instead on the payoffs they provide. Expository treatments of portfolio optimization include Karatzas and Shreve (1998), Schachermayer (2002), Skiadas (2006).

However, we will now consider briefly two issues in formulating an optimization underlying a pricing scheme: whether the optimization takes into account only the market risk of the OTC trade itself or also accounts for the opportunities for future trades, and whether the portfolio strategy is instantaneously, myopically optimal or optimal over an entire time interval.

4.1.1 Opportunity

Accounting for changing investment opportunities leads to better portfolio strategies. The optimal portfolio can be decomposed into a term that would be optimal if asset returns were independent, plus a term that corrects for the dependence of current asset returns and the conditional distribution of all future asset returns. For example, suppose that there is a riskless asset and one risky stock whose log price follows a diffusion with stochastic drift and volatility. A state with a higher ratio of drift to volatility constitutes a more favorable investment opportunity (cf. the Sharpe ratio) and thus a greater certainty equivalent for wealth. Suppose further that the change in this drift-volatility or *mean–variance ratio* is negatively correlated with the asset return. The optimal allocation to the stock is greater than it would be if the mean–variance ratio were deterministic: a loss from investing in the stock is cushioned by an increased certainty equivalent for each dollar of wealth. This increased demand for stock in the optimal portfolio is *hedging demand*. For a very lucid theoretical account of this phenomenon in the context of quadratic hedging (Section 6), see Schweizer (1995, especially p. 16). Extensive numerical results for hedging options occupy Brandt (2003); Example 4.1 is related.

Analogously, for an OTC market-maker, there is a stochastic process of OTC trade opportunities, i.e. requests for a quote of bid and ask prices by a potential customer, where each customer has reservation prices below or above which he is willing to buy or sell. Routledge and Zin (2004) take a step in this direction, which merits greater attention. The methods of pricing covered in this survey all focus on whether an individual OTC trade is attractive to the market-maker without considering its effect on future trades. However, a trade done now affects the portfolio the trader will have in the future, and in light of which he will evaluate future trades. For example, if there is a risk constraint, doing an OTC trade now might prevent the trader from doing a more attractive trade in the future. Therefore, each opportunity should be evaluated in light of the stochastic process of future opportunities: the compensation for doing a trade should reflect the direct cost of possible losses and also the indirect cost of lost opportunity for profit on future trades that may be passed up due to the risk associated with this trade.

4.1.2 Local vs. global

In pricing an OTC security, a *global* optimization optimizes over portfolio strategies that cover an entire time interval. This may be difficult to solve, whether numerically or analytically, or even to set up. A simpler alternative is a *local* optimization, in which the objective and constraints contain only static criteria, changes over a single time step, or instantaneous rates of change. A local optimization optimizes over the current portfolio weights only: whether one intends to hedge dynamically or not, a local optimization is a static problem, in a sense.

Global criteria include terminal wealth, total utility from consumption over an entire time interval, value at risk, and squared hedging error. The global

criteria can be constraints as well as objectives, for example, the constraint that the wealth process never be negative. Local criteria include Greeks and often form pairs with global criteria. For example, in quadratic hedging there is a locally risk-minimizing strategy and a global variant, the variance-optimal hedge: see Example 6.1. Analogous to the usual expected utility maximization (Section 4.2.3) is the local utility maximization Kallsen (2002a), discussed in Section 9.2 similar to the following example based on Schweizer (1995, §5), but in continuous time.

Example 4.1. There is a riskless bank account whose value is always \$1, and a risky asset whose prices are given by Table 1. An investor has \$100 of initial wealth and utility function $u(W) = -(W/100)^{-4}$. The investor maximizes the expected utility of wealth at time 2 over self-financing strategies: the decision variables are ξ_1, the number of shares of the risky asset to hold over the first step, and $\xi_2^{(+)}$, $\xi_2^{(0)}$, and $\xi_2^{(-)}$, the number of shares to hold over the second step, respectively if the risky asset price at time 1 is 1, 0, or -1.

Local optimization of one-step expected utility in each of the four scenarios yields $\xi_1 = \xi_2^{(0)} = \xi_2^{(-)} = 0$ and $\xi_2^{(+)} = 16.13$: only when the risky asset's price is \$1 at time 1 is its one-step expected return positive, so that it is worth investing in it, from a local perspective. A global optimization of two-step expected utility yields $\xi_1 = -3.27$, $\xi_2^{(0)} = \xi_2^{(-)} = 0$, and $\xi_2^{(+)} = 15.60$: the negative position in the risky asset over the first step hedges the increase in the derived utility of wealth at time 1 if the asset's price should rise. See Example 6.1 for a continuation.

That is, local optimization ignores hedging demand, while global optimization captures it. Intermediate wealth is worth more in states with better investment opportunities, and the global optimization yields greater expected utility from terminal wealth by producing more wealth in the intermediate states with poorer investment opportunities.

Table 1.
Risky Asset Prices.

State	Probability	Time 0	Time 1	Time 2
1	1/9	\$0	\$1	\$3
2	1/6	\$0	\$1	\$2
3	1/18	\$0	\$1	\$0
4	1/6	\$0	\$0	\$1
5	1/6	\$0	\$0	−\$1
6	1/6	\$0	−\$1	\$0
7	1/6	\$0	−\$1	−\$2

4.2 Pricing via portfolio optimization

No-arbitrage bounds (Section 4.2.1) and indifference prices (Section 4.2.2) are special cases of the mathematical structure of *good deal bounds* (Section 4.2.4). Let R be the set of replicable payoffs, $\pi(Y)$ be the market price to replicate a payoff $Y \in R$, and A be an *acceptance set* of payoffs that are acceptable compared to the status quo. The lower good deal bound for a payoff X, which might be interpreted as a bid price, is

$$b(X) = \sup_{Y \in R} \{-\pi(Y) \mid Y + X \in A\}. \tag{2}$$

If we can buy X over the counter for less than $b(X)$ then there is a Y that we can buy in the market for $\pi(Y)$, such that in total we get $X + Y$, which is acceptable, for a cost $b(X) + \pi(Y) < 0$. The upper good deal bound or ask price for X is

$$a(X) = -b(-X) = \inf_{Y \in R} \{\pi(Y) \mid Y - X \in A\}. \tag{3}$$

To sell X or to buy $-X$ has the same effect. The other minus sign in $a(X) = -b(-X)$ reflects the convention that the buyer pays the price to the seller. Because of the relationship $a(X) = -b(-X)$, one may specify only b (or a), getting distinct price bounds unless b is antisymmetric.

The interpretation of $-b(X)$ is the cost of rendering X acceptable, and this can be thought of as a risk measure. As Jaschke and Küchler (2001, n. 6) say, "any valuation principle that yields price bounds also induces a risk measure and vice versa." Indeed, under some conditions, $-b$ is a coherent or convex risk measure (Artzner et al., 1999; Föllmer and Schied, 2002). The no-arbitrage bounds provide an example. For generalities, see Jaschke and Küchler (2001, Prop. 7) and Staum (2004, Prop. 4.2).

The acceptance set A must include $\{Z \mid Z \geqslant 0\}$, the set of riskless payoffs, which is the acceptance set that generates no-arbitrage bounds. It must not intersect the set $\{Z \mid Z < 0\}$ of pure losses with no chance of gain. Finally, $Z \in A$ and $Z' \geqslant Z$ must imply $Z' \in A$. These three properties correspond to a subset of the axioms defining coherent risk measures (Artzner et al., 1999).

The acceptance set A must also be consistent with market prices π, or arbitrage will result. For example, if there is an acceptable payoff $Y \in A$ with negative cost $\pi(Y) < 0$, then $b(0) > 0$, and the trader is thus expressing willingness to give money away in exchange for nothing. For a concrete example using expected utility indifference pricing, see Section 5.2.1. For general remarks, related to duality, see Section 4.2.5.

4.2.1 No-arbitrage pricing

The *no-arbitrage price bounds* are given by Eqs. (2) and (3) with the acceptance set $A = \{Z \mid Z \geqslant 0\} = \{Z \mid \operatorname{ess\,inf} Z \geqslant 0\}$:

$$b_{NA}(X) := \sup_{Y \in R} \{-\pi(Y) \mid Y + X \leqslant 0\} = -\inf_{Y \in R} \{\pi(Y) \mid Y \geqslant -X\} \tag{4}$$

and

$$a_{NA}(X) := \inf_{Y \in R} \{\pi(Y) \mid Y - X \leqslant 0\} = \inf_{Y \in R} \{\pi(Y) \mid Y \geqslant X\}. \tag{5}$$

That is, a payoff is acceptable if and only if it has no risk of loss under the statistical probability measure $\mathbf{P}$. Buying X for less than $b_{NA}(X)$ or selling it for more than $a_{NA}(X)$ admits arbitrage. For instance, for any $\epsilon > 0$, there is a $Y_\epsilon \in R$ such that $\pi(Y) < \epsilon - b_{NA}(X)$ and $Y_\epsilon \geqslant -X$. Thus, if we buy X for $b_{NA}(X) - \epsilon$ and also buy Y_ϵ, we acquire $Y_\epsilon + X \geqslant 0$ (this is *super-replication* of $-X$) for a negative total cost: we get paid now and assume no risk of loss. El Karoui and Quenez (1995) give a dynamic programming algorithm for computing the no-arbitrage bounds.

While $-\operatorname{ess\,inf} X$ measures the worst possible loss X can yield, $-b_{NA}$ is also a risk measure, measuring the cost of hedging to prevent the worst possible loss. The solution Y^* to Problem (4) is an optimal hedge for X: it is the cheapest payoff that combines with X to produce a portfolio with zero probability of loss. The typical result for a complete market is that $Y^* = -X$, X solves Problem (5), and $b_{NA}(X) = a_{NA}(X) = \pi(X)$, the cost of replicating X. In an incomplete market, $b_{NA}(X)$ and $a_{NA}(X)$ are usually too low and too high, respectively, to be of use to an OTC market-maker: few customers would be willing to trade at such prices (see e.g. Eberlein and Jacod, 1997).

4.2.2 Indifference pricing

Indifference prices are good deal bounds with acceptance set $A = \{Z \mid P(Z) \geqslant P(0)\}$ in Eqs. (2) and (3), where P is a *preference function* specifying *complete preferences*. Completeness of preferences is different from completeness of markets: it means that for any pair of payoffs X and Y, either one prefers X to Y, is indifferent between X and Y, or prefers Y to X. With a preference function, these three cases correspond to $P(X) > P(Y)$, $P(X) = P(Y)$, and $P(X) < P(Y)$ respectively. Buying X for less than $b(X)$ results in a nonnegative cost to acquire a total payoff $X + Y$ that is at least as good as the status quo, i.e. $P(X + Y) \geqslant 0$.

The main point of indifference pricing is not the mathematics of a preference function versus an acceptance set; it is possible to convert between them as for risk measures and acceptance sets (Jaschke and Küchler, 2001). The point is the interpretation of the set A as the set of *all* payoffs that are at least as good as the status quo. The no-arbitrage bounds are not to be interpreted as indifference prices. They have the form of indifference prices with P equal to the essential infimum, which is far too conservative: it says that zero is preferable to any payoff with a positive probability of loss.

Indifference pricing takes place against the background of the portfolio optimization problem

$$\sup_{Y \in R} \{\tilde{P}(W + Y) \mid \pi(Y) \leqslant c\} \tag{6}$$

where the initial endowment consists of c dollars and the random wealth W, and $\tilde{P}$ is a preference function over the total random wealth. Then $V^* = W + Y^*$ is the total random wealth produced by the trader's optimal portfolio strategy. A trader who has the opportunity to purchase X over the counter formulates the problem

$$b(X) = \sup_{Y \in R} \{-\pi(Y) \mid \tilde{P}(V^* + X + Y) \geqslant \tilde{P}(V)\} \tag{7}$$

to find the indifference bid price. If Y^* solves Problem (7) with the constraint tight, then the trader is indeed indifferent between $V^* + X + Y^*$ and V^*, i.e. between doing and not doing the trade at $b(X)$.

Problem (7) coincides with Problem (2) when $P(Z) = \tilde{P}(V^* + Z) - \tilde{P}(Z)$. That is, preferences over payoffs, which are changes in wealth, used in constructing indifference prices, are derived from more fundamental preferences over total wealth. Therefore, preferences over payoffs depend on the optimal total random wealth V^* in Problem (6). For various reasons, e.g. that the procedure takes too long or that one does not trust its results, one may wish to avoid solving Problem (6) first, instead simply solving Problem (7) with V, determined by the status quo portfolio strategy, replacing the optimal V^*. However, Problem (7) can be quite sensitive to V, and if $V \neq V^*$, the indifference price can violate the no-arbitrage principle. There is an example and further discussion in Section 5.2.1. Another way of dealing with this situation is to formulate indifference prices by incorporating the portfolio optimization problem (6):

$$b(X) = \sup_{Y \in R} \left\{ c - \pi(Y) \mid \tilde{P}(X + Y) \geqslant \sup_{V \in R} \{\tilde{P}(V) \mid \pi(V) \leqslant c\} \right\} \tag{8}$$

based on an initial budget of c.

4.2.3 *Expected utility*

Expected utility theory specifies the preference function as $\tilde{P}(W) = \mathrm{E}[u(W)]$, where the *utility function* u is increasing because more money is better and concave because of risk aversion. It is characteristic of expected utility indifference pricing that $a(X) \neq b(X)$ for a typical nonreplicable payoff X: it leads to price bounds, not a unique price, because of aversion to risk that cannot be hedged. As Musiela and Zariphopoulou (2004b) emphasize, "no linear pricing mechanism can be compatible with the concept of utility based valuation," so we should not expect to have the ask price $a(X) = -b(-X)$ equal to the bid price $b(X)$. Marginal indifference pricing (Section 4.3) delivers a unique price based on expected utility.

Expected utility indifference pricing is difficult to implement in the context of derivative security pricing. The key inputs to expected utility maximization are the endowment V, the statistical probability measure $\mathbf{P}$, and the utility function u. As Carr et al. (2001, §1) observe,

> Unfortunately, the maximization is notoriously sensitive to these inputs, whose formulation is suspect at the outset. This shortcoming renders the methodology potentially useless ...

OTC traders prefer calibration (Section 11), which does not require them to specify the endowment, the utility function, or the parameters of **P**, but only the form of the pricing measure **Q**. In particular, it is not required to estimate the expected return of marketed assets under **P**, which is difficult (Section 3.3), but of the utmost importance for expected utility maximization. It is likewise difficult to determine an appropriate utility function in the corporate setting of making a market in OTC derivatives. What is the basis for corporate risk aversion? The view of the equity of a firm with debt as a call option on the firm's value suggests that shareholders should be risk-seeking, so as to maximize the value of this call option. Does corporate risk aversion come from regulatory capital requirements, or from financial distress costs (for which see Jarrow and Purnanandam, 2004, and references therein), and if so, how is it to be quantified? To model the firm's endowment, one ought to include not only all securities, loans, and liabilities currently on the books, but also future business earnings as a going concern: for instance, one would want to know the dependence between portfolio returns and earnings from doing advisory work on mergers and acquisitions.

Aside from these perplexities in modeling, continuous-time expected utility maximization also involves difficult technicalities. For example, it is not easy to pick a suitable set of portfolio strategies over which to optimize (Delbaen et al., 2002; Kabanov and Stricker, 2002; Schachermayer, 2003). Work has continued in this area, to clarify the conditions that are necessary for existence of optimal portfolios and unique prices (Hugonnier and Kramkov, 2004; Hugonnier et al., 2005; Karatzas and Žitković, 2003). Schachermayer (2002) and Skiadas (2006) give expository treatments of the problem of expected utility maximization in a continuous-time incomplete market, providing a basis for indifference pricing.

4.2.4 Good deal bounds

The acceptance set A for use in the good deal bounds (2) and (3) includes only payoffs that are preferable to the status quo, but possibly not all of them. At prices below $b(X)$, it is preferable to buy; at prices above $a(X)$, it is preferable to sell. In indifference pricing, A contains all payoffs preferable to the status quo, so at prices between $b(X)$ and $a(X)$ it is preferable to do nothing. Otherwise, $b(X)$ is a lower bound on the indifference bid price and $a(X)$ is an upper bound on the indifference ask price, and the best policy at prices between $b(X)$ and $a(X)$ is unknown. This is the difference of interpretation between good deal bounds and indifference prices, which are a mathematical special case of the former.

There are two alternative interpretations of good deal bounds. One treats good deal bounds as possible bid and ask prices for a market-maker, much like indifference prices: see e.g. Cochrane and Saá-Requejo (2000, p. 86), Carr et al. (2001, §7), Staum (2004), and Larsen (2005, §5). This is financial engineering, with the goal of making only subjectively good deals, by trading outside the *subjective* price bounds, buying below $b(X)$ and selling above $a(X)$. The other interpretation is that A is a subset of the payoffs that many traders prefer to

the status quo, as in Section 9. It treats good deal bounds like no-arbitrage bounds, asserting that good deals should not be available, because almost everyone would be willing to take them: see e.g. Cochrane and Saá-Requejo (2000, p. 82), Carr et al. (2001, §1), and Černý and Hodges (2002). This may be mathematical finance, with the goal of making a more precise statement about observed prices in incomplete markets than does the no-arbitrage principle. However, if these *objective* good deal bounds are narrow enough, they indeed offer market-makers useful guidance about prices: they should buy below $a(X)$ and sell above $b(X)$. In such a trade, the counterparty sells below $a(X)$ and buys above $b(X)$, not receiving a good deal from the market-maker. If the counterparty also insists on buying below $a(X)$ and selling above $b(X)$, trades take place inside the price bounds, so that neither party gets a good deal.

So far we have been discussing an abstract framework. How can it be given economic content by specifying the acceptance set A? The primary approaches include restrictions on the pricing kernel (Section 9) and robustness (Section 10). A simple version involves a convex risk measure formed by a finite number of valuation measures and stress measures with floors (Carr et al., 2001; Larsen et al., 2005), which might be specified by looking at the marginal utility and the risk management constraints of several market participants (Carr et al., 2001, §2). In Section 8, we consider methods that yield price bounds and have the same mathematical form as good deal bounds, except that they use acceptance sets A that violate the axioms in Section 4.2. This causes them to be unsuitable for OTC pricing, although they have other uses.

4.2.5 *Duality*

Duality provides Formula (1) for no-arbitrage bounds and related expressions for a good deal bound or indifference price as in Eq. (2): see Jaschke and Küchler (2001, §4) and Staum (2004, Thm. 4.1). It yields both computational advantage and insight. For example, in pricing a path-independent European option given continuous trading, the dual optimization is taken over a set of probability measures on terminal payoffs, which is more tractable than the set of continuous-time portfolio strategies appearing in the primal problem. The two major ways of grounding pricing in optimization involve the two sides of this duality: portfolio optimization is optimization over portfolios or the payoffs they provide, while the methods of selecting minimum-distance measures or subsets of the set $\mathcal{Q}$ of EMMs involve optimization over probability measures. For more on duality in indifference pricing, see Frittelli (2000a, §3).

For an exposition of portfolio optimization in incomplete markets in terms of convex duality, including equivalent martingale measures and marginal indifference pricing, see Schachermayer (2002). Convex duality also appears in representation and optimization of risk measures (Ruszczyński and Shapiro, 2004). The conditions for the price bounds (2) and (3) to avoid arbitrage are best understood in terms of duality: for a version of the first fundamental theorem of asset pricing, see Staum (2004).

Under some conditions, including that the acceptance set A be related to a coherent risk measure, the price bounds (2), (3) have the dual representation

$$\Big(\inf_{\mathbf{Q}\in\mathcal{D}} \mathrm{E}_{\mathbf{Q}}[DX],\ \sup_{\mathbf{Q}\in\mathcal{D}} \mathrm{E}_{\mathbf{Q}}[DX]\Big), \tag{9}$$

where $\mathcal{D}$ is a subset of the set $\mathcal{Q}$ of EMMs (Jaschke and Küchler, 2001). For the no-arbitrage bounds, $\mathcal{D} = \mathcal{Q}$. When the price bounds coincide and are linear, $\mathcal{D}$ is a singleton, i.e. the method selects a single EMM (see Section 2.3 for a discussion). Marginal indifference pricing and minimum-distance measures are the principal methods of selecting a single EMM.

4.3 Marginal pricing

For any price bounds b and a, $\lim_{\gamma\downarrow 0} \gamma a(X/\gamma)$ and $\lim_{\gamma\downarrow 0} \gamma b(X/\gamma)$ may coincide and provide a unique price $\tilde{p}(X)$ suitable for small trades. For a general result on good deal bounds, see Staum (2004, Prop. 5.2). Under expected utility preferences, $\tilde{P}(W) = \mathrm{E}[u(W)]$, this suggestion corresponds to using the marginal utility u' to define a pricing measure $\mathbf{Q}$:

$$\tilde{p}(X) = \mathrm{E}_{\mathbf{P}}\big[u'(V)DX\big] = \mathrm{E}_{\mathbf{Q}}[DX], \tag{10}$$

where D is the discount factor and $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P} = u'(V)/\mathrm{E}[u'(V)]$. That is, in the most straightforward case, marginal indifference pricing results in the selection of a single EMM $\mathbf{Q}$ whose likelihood ratio with respect to the statistical probability measure $\mathbf{P}$ is proportional to the marginal utility of terminal wealth provided by an optimal portfolio.

Marginal indifference pricing is based on the idea that a single trade is small and does not need to be hedged. This argument is appropriate for finding the equilibrium price of a security that is traded and infinitely divisible, but see Section 2.1. If a small trade has negligible impact on the whole portfolio's risk profile, e.g. it has little effect on marginal utility, that is an argument for using the unique marginal indifference price. This argument is not generally appropriate for OTC market-making. A single small trade might seem to be priced adequately by marginal indifference, but many small trades cumulatively can involve large risks. Ignoring the likely cumulation of risks can cause initial, myopic underpricing of OTC securities that are in high demand, followed by a concentration of related risks and thus the need to set high prices, at which fewer trades would be made (see Section 4.1.1). The contribution of a small trade to total risk depends on the opportunities for hedging, which should therefore affect pricing.

4.4 Minimum-distance pricing measures

Marginal indifference prices based on expected utility are an example of pricing with a minimum-distance measure. The expected utility is an expectation under a statistical probability measure $\mathbf{P}$. The marginal indifference price

is an expected discounted payoff under a *minimax martingale measure* $\mathbf{Q} \in \mathcal{Q}$ that is "closest" to $\mathbf{P}$ in the sense of providing the least possible expected utility to an investor who could buy any payoff V for $\mathrm{E}_{\mathbf{Q}}[DV]$. That is, $\mathbf{Q}$ corresponds to the "least favorable market completion": in the fictitious complete market in which the price of any payoff V is $\mathrm{E}_{\mathbf{Q}}[DV]$, the utility derived from optimal investment is as low as possible (Skiadas, 2006). The minimax martingale measure $\mathbf{Q}$ is the solution to

$$\min_{\mathbf{Q}\in\mathcal{Q}} \max_{V} \{\mathrm{E}_{\mathbf{P}}[u(V)] \mid \mathrm{E}_{\mathbf{Q}}[DV] \leqslant c\}. \tag{11}$$

For a version based on local utility, see Kallsen (2002b) and references therein. Particular choices of utility yield quadratic and exponential methods in Sections 6–7; the latter distance can also be described in terms of relative entropy. The same concepts appear in Section 9, featuring not just the minimum-distance measure but a set of EMMs having low distance to $\mathbf{P}$. For more on portfolio optimization and minimum-distance measures, see Goll and Rüschendorf (2001).

Somewhat different is the case of calibration (Section 11), which is not based on a statistical probability measure $\mathbf{P}$. Instead it starts from a parametric family $\mathcal{P}$, and selects the pricing measure $\hat{\mathbf{Q}} \in \mathcal{P}$ that is closest to $\mathcal{Q}$ in the sense of having the least error in replicating the prices of marketed derivative securities; an EMM in $\mathcal{Q}$ would yield zero replication error.

Figure 1 illustrates the structure of four schemes for selecting a probability measure for pricing in an incomplete market. It uses the very simple setting of a one-period model with three states and two marketed securities: a riskless bond paying \$1 in all states and with initial price of \$1, and a stock worth \$2 in state 1, \$1 in state 2, and \$0 in state 3 and having initial price \$0.80. To simplify matters even further for purposes of two-dimensional representation, we will assume that the bond must be repriced exactly, so the price assigned to a payoff X is $\mathrm{E}_{\mathbf{Q}}[DX] = \mathrm{E}_{\mathbf{Q}}[X]$ where the pricing measure $\mathbf{Q} = (q_1, q_2, q_3)$ is a true probability measure, such that the probabilities of the three states sum to one: $q_1 + q_2 + q_3 = 1$. Thus, $q_3 = 1 - (q_1 + q_2)$, so all possible pricing measures can be parametrized by the triangle in Fig. 1: $q_1 \geqslant 0$, $q_2 \geqslant 0$, $q_1 + q_2 \leqslant 1$. The diagonal line $2q_1 + q_2 = 0.8$ represents the constraint of repricing the stock, so its line segment in the interior of the triangle is the set $\mathcal{Q}$ of EMMs for any statistical probability measure $\mathbf{P}$ that assigns positive probability to all states. The vertical line segment inside the triangle and defined by $q_1 = 0.5$ represents a set $\mathcal{P}$ of models. Of course, this example is so simple that there is no need to restrict attention to a subset of the possible pricing measures that does not include any measures that reprice the stock; also, ordinarily models include underlying securities' initial prices as parameters, so all underlying securities, as opposed to derivative securities, are repriced exactly. The point of the setup in Fig. 1 is that the resulting structure is not only very simple, but also similar to that encountered in practice, in which one works with a parametric family of models that does not include an EMM.

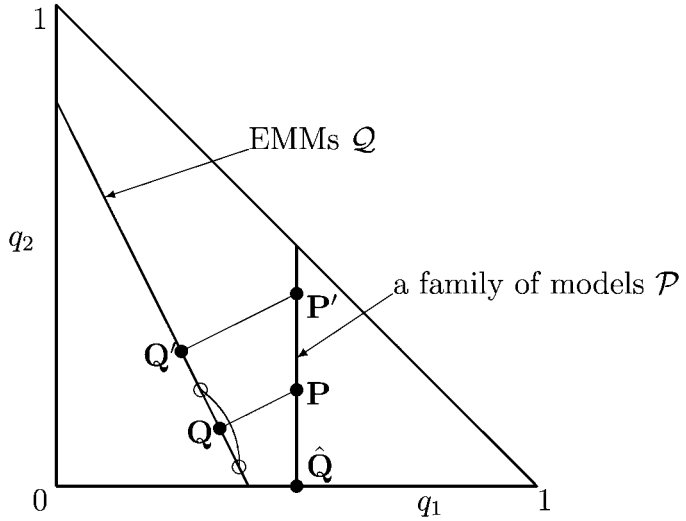


Fig. 1. Structures of schemes for selecting a pricing measure.

Calibrating the family of models $\mathcal{P}$ to the stock price selects $\hat{\mathbf{Q}} = (0.5, 0, 0.5)$ as the pricing measure, which minimizes the error in repricing the stock by assigning it the price \$1, the least possible within this family of models. Another scheme begins with a statistical probability measure $\mathbf{P}$, which may have been estimated within the family $\mathcal{P}$ by econometric inference, and then selects the EMM $\mathbf{Q}$ that is closest to $\mathbf{P}$. In Figure 1, $\mathbf{Q} = (0.34, 0.12, 0.44)$ minimizes Euclidean distance, but several distances have been proposed, relating e.g. to entropy or expected utility. Instead of selecting only $\mathbf{Q}$, which minimizes the distance to $\mathbf{P}$, to get a unique price, one may select a set of pricing measures having low distance to $\mathbf{P}$, and get an interval of prices: the empty dots connected by a curved arc around $\mathbf{Q}$ represent the extreme measures selected by this scheme. Where distance is a function of $d\mathbf{Q}/d\mathbf{P}$, this scheme includes some approaches based on pricing kernel restrictions (Section 9). The fourth scheme begins with multiple probability measures, here $\mathbf{P}$ and $\mathbf{P}'$, yielding robustness to ambiguity about the statistical probability measure (Section 10). Each of these yields a minimum-distance EMM, here $\mathbf{Q}$ and $\mathbf{Q}'$ respectively, and this resulting set of EMMs can be used to generate a unique price or a price interval.

5 Issues in pricing and expected utility examples

Our main example is adapted from Carr et al. (2001).

Example 5.1. Consider a single-period economy with five possible states and three assets: a riskless bond, a stock, and a straddle. The bond and stock are

Table 2.
Terminal Asset Values.

	State 1	State 2	State 3	State 4	State 5
Bond	\$1	\$1	\$1	\$1	\$1
Stock	\$80	\$90	\$100	\$110	\$120
Straddle	\$20	\$10	\$0	\$10	\$20

Table 3.
Expected Utility Indifference Pricing of a Straddle.

Initial wealth	Initial portfolio		Transaction type	Portfolio adjustment		Indifference price
	Bond	Stock		Bond	Stock	
\$100	77.6	0.334	buy	−25.1	0.147	\$9.92
			sell	21.3	−0.082	\$12.11
\$1000	776	3.34	buy	−22.8	0.111	\$10.94
			sell	22.5	−0.105	\$11.17

marketed, with initial prices \$0.9091 and \$88.1899 respectively. The terminal values of the three assets are given in Table 2. The no-arbitrage bounds for the straddle price are \$2.72 and \$18.18. Consider the utility function $u(W) = -(W/100)^{-4}$ for $W > 0$, and suppose the states have equal probabilities.

For any level of initial wealth, the optimal portfolio in marketed securities has 70.55% of the wealth in the bond and 29.45% in the stock.[2] Pricing by marginal utility uses the probabilities $\mathbf{Q} = (26.55\%, 22.68\%, 19.46\%, 16.78\%, 14.53\%)$, yielding a unique price of \$11.06 for the straddle. The bid and ask indifference prices when the initial wealth is \$100 or \$1000 allocated optimally are (\$9.92, \$12.11) and (\$10.94, \$11.17) respectively. Table 3 shows the corresponding portfolio adjustments providing optimal payoff Y^*.

5.1 *Dependence on trading opportunities*

The opportunities to trade in the market affect the indifference price. For example, suppose that the stock were not marketed, but the initial portfolio still had 29.45% of its wealth in the stock. Then the indifference prices based on initial wealth of \$100 would be (\$9.86, \$12.14): with fewer opportunities to rebalance the portfolio, the price interval would become wider. The marginal

[2] Power and log utilities, having constant relative risk aversion, can lead to optimal portfolios whose allocation fractions do not depend on the initial wealth: see e.g. Karatzas and Shreve (1998, Examples 3.6.6–7).

indifference price given by Eq. (10) would *not* change: it involves an infinitesimal change in the portfolio, in the direction defined by the straddle payoff, and no portfolio rebalancing. On the other hand, there is a difference in the marginal prices derived from the indifference price (8) incorporating portfolio optimization, depending on whether the optimal portfolio is allowed to contain stocks and bonds, or only bonds. In the latter case, the optimal portfolio provides the same wealth in each state, so the marginal indifference price is \$12, based on $\mathbf{Q} = (20\%, 20\%, 20\%, 20\%, 20\%)$.

5.2 Dependence on current portfolio

The indifference prices and the optimal portfolio adjustments also depend on the random wealth V provided by the initial portfolio. This is intuitively reasonable, as a trader should be less eager to acquire a payoff that exacerbates unhedgeable risk in the current portfolio than one that cancels out such risks. As Rouge and El Karoui (2000) say, "it is unrealistic that agents with different endowments should have the same attitude toward risk." Indeed, OTC market makers describe an unhedgeable risk in their portfolios as an "axe," thinking of the expression "having an axe to grind." For example, suppose a trader is long OTC options on a stock with no marketed options. It would not be easy to hedge the risk of a decline in the implied volatility (and hence value) of these options, so this long position is an "axe" which the trader would like to "grind" by selling OTC options. The trader would set low ask and bid prices, to encourage sales of options, which decrease this risk, and get adequate compensation for purchases, which increase this risk.

Table 4 illustrates this point for Example 5.1. It shows that after a trader has bought a straddle and re-optimized the portfolio, as in Table 3, the bid and ask prices decrease. The new ask price is the same as the original bid price, which makes sense: together, the two transactions return the trader to the original portfolio, so a net cost of zero produces indifference. The marginal indifference price after buying a straddle and optimally rebalancing is \$8.77 for the case of \$100 initial wealth; this too decreases because the change in the portfolio has reduced marginal utility in most of the states in which the straddle pays off.

Table 4.
Effect of Initial Portfolio on Expected Utility Indifference Pricing.

Initial wealth	Initial portfolio		Transaction type	Hedge portfolio		Indifference price
	Bond	Stock		Bond	Stock	
\$100	52.5	0.481	buy	−33.5	0.258	\$7.66
	plus 1 straddle		sell	25.1	−0.147	\$9.92
\$1000	753	3.45	buy	−23.2	0.118	\$10.72
	plus 1 straddle		sell	22.8	−0.111	\$10.94

The dependence of indifference price on initial portfolio, illustrated in Table 4 for constant relative risk aversion, occurs even with constant absolute risk aversion (exponential utility, Section 7), which is often used to obtain separation of investment and hedging decisions. In Example 5.1, if the utility function is replaced by $u(W) = -\exp(-(0.0453)W)$, the optimal allocation of \$100 initial wealth remains very nearly the same, leading to a similar bid–ask spread of (\$9.89, \$12.10). After the trader has bought a straddle and re-optimized the portfolio, the bid–ask spread becomes (\$7.64, \$9.89).

5.2.1 Optimality as prerequisite for indifference pricing

Indifference prices should fall within the no-arbitrage bounds, so as to avoid arbitrage in OTC trades. To prevent the indifference price in Eq. (7) from violating no-arbitrage bounds, the initial portfolio V must be optimal, i.e. $V = V^* = W + Y^*$ where Y^* solves the portfolio optimization problem (6). If V is suboptimal, the indifference price may exceed the market price for a replicable payoff that increases the preference index $\tilde{P}$ in Problem (6). Likewise, the indifference price for a non-replicable payoff may exceed its upper no-arbitrage bound. Example 5.2 illustrates these effects.

Example 5.2. Continuing Example 5.3, suppose the initial portfolio delivers \$100 except in state 1, in which it delivers only \$60. Consider the payoff Y provided by a portfolio long 100 shares of the bond and short one share of the stock.

Based on marginal utility, the valuation probability of state 1 is 76.28%, giving a marginal indifference price for the put of \$6.93. Its indifference price is \$8.42. These both exceed the upper no-arbitrage bound of \$5.23. Acquiring Y increases expected utility. Its market price is \$2.72, but its marginal indifference price is \$9.49 and its indifference price is \$7.28.

The need to base indifference pricing on an optimal portfolio causes a grave difficulty in using expected utility. Because expected utility maximization is not robust to ambiguity about the statistical probability measure **P** (Section 4.2.3), it is not actually a good idea to adopt the supposedly optimal portfolio. Typically, the true expectation of V^* is lower than $\mathrm{E}_{\mathbf{P}}[V^*]$, because the optimal portfolio overinvests in assets that are wrongly believed to have high expected returns. Consequently, the trader does not optimize his portfolio, $V \neq V^*$, but to avoid arbitrage, the indifference prices must be based on V^*. The result is that the trader is not indifferent between trading and not trading at these "indifference prices"; someone else with a different portfolio would be. The economic justification for expected utility indifference pricing evaporates.

5.3 Risk vs. preference

It is tempting to think of the optimal portfolio adjustment Y^* in Problem (7) as a hedge for the payoff X, but as we have seen, Y^* and the indifference price

$b(X)$ depend on the payoff V from the existing portfolio strategy as well as on the payoff X. Only in special cases such as neutralization of Greeks does hedging apply to payoffs without reference to a portfolio: delta-hedging each security in a portfolio produces the same net position as delta-hedging the whole portfolio. The hedge that minimizes a portfolio's risk does not generally coincide with the sum of such hedges for each security in the portfolio.

Another way in which the optimal Y^* in Problem (2) or (7) is not a hedge is that it need not reduce risk. To formulate a less risky alternative, suppose that market prices π are linear and there is a reference security (e.g. riskless bond) with payoff denoted $\mathbf{1}$. One can finance the purchase of a payoff X for $b(X)$ by acquiring Y^* or, more simply, by acquiring $-(b(X)/\pi(\mathbf{1}))\mathbf{1}$. By definition, it is *preferable* to acquire Y^*: $P(V+X+Y^*) \geqslant P(V+X-(b(X)/\pi(\mathbf{1}))\mathbf{1}$, but it need not be less *risky* to acquire Y^*. Unless the preference P and risk measure ρ are related as $P = -\rho$, it is possible that $\rho(V+X+Y^*) > \rho(V+X-(b(X)/\pi(\mathbf{1}))\mathbf{1}$. The following example illustrates this point.

Example 5.3. We extend Example 5.1 by including another non-traded asset, a put option on the stock with strike \$90. Its only nonzero payoff is \$10 in state 1, in which the stock is worth \$80. As a risk measure of a payoff W, we use the tail conditional expectation (see Artzner et al., 1999) of $W - 110$, the shortfall relative to investing \$100 in bonds.

The no-arbitrage bounds on the put's price are (\$0, \$5.23) and the bid and ask indifference prices when the initial wealth is \$100 allocated optimally are (\$2.22, \$2.60), with marginal indifference price \$2.41. Table 5 shows the state-by-state values of the original optimized portfolio V, of the portfolio $V + X - (b(X)/\pi(\mathbf{1}))\mathbf{1}$ after buying the put for \$2.22 by selling bonds, and of the re-optimized portfolio $V+X+Y^*$. Table 6 shows these portfolio's tail conditional expectations at several probability levels, corresponding to average values over the worst 1–5 states.

The last column, tail conditional expectation at the 100% level, is $\mathrm{E}[110 - W]$, which simply measures the portfolio's expected value. The other columns are more properly risk measurements. They each show that buying the put by selling bonds reduces risk, while re-optimizing the portfolio increases risk even beyond its original levels. Because the put adds extra wealth in state 1, the worst state for the original portfolio, it allows the re-optimized portfolio

Table 5.
Portfolio Values when Buying a Put.

Portfolio	State 1	State 2	State 3	State 4	State 5
Original Optimal	\$104.32	\$107.66	\$111.00	\$114.34	\$117.68
Buy Put, Sell Bonds	\$111.88	\$105.22	\$108.55	\$111.89	\$115.23
Re-optimized	\$107.97	\$103.61	\$109.24	\$114.88	\$120.51

Table 6.
Risk in Buying a Put.

Portfolio	Tail Conditional Expectation				
	20%	40%	60%	80%	100%
Original Optimal	5.68	4.01	2.34	0.67	−1.00
Buy Put, Sell Bonds	4.78	3.11	1.45	0.61	−0.56
Re-optimized	6.39	4.21	3.06	1.07	−1.24

to allocate a greater fraction of wealth to the stock. This maximizes expected utility, but it increases risk: for instance, the re-optimized portfolio has less wealth ($103.61) in its worst state than does the original portfolio ($104.32) in its worst state.

Even if preferences are risk-averse, preference and risk are not simply opposites, as the example shows, even though it is always preferable and less risky to have more wealth. To incorporate risk management concerns, one may add a risk constraint. We could reformulate the trader's portfolio optimization problem (6) as

$$\sup_{Y \in R} \{\tilde{P}(W + Y) \mid \pi(Y) \leqslant c, \rho(W + Y) \leqslant r\}, \tag{12}$$

where internal or external regulators impose the risk measure ρ and the limit r on the risk of the trader's portfolio. Given this formulation, one might think of the solution to Problem (6) as an optimal portfolio adjustment and of the difference between the solutions to Problems (6) and (12) as a hedge. "Hedging" is a good description of neutralizing Greeks, which is solely risk minimization, with no other preference involved; when optimizing with preferences distinct from risk, portfolio re-optimization need not be hedging i.e. risk reduction.

6 Quadratics

Quadratic hedging is a much-studied, mathematically elegant approach to incomplete markets. Surveys include Pham (2000) and Schweizer (2001). The quadratic method is a special case of expected utility indifference pricing, with quadratic utility $u(x) = -x^2$. Because it is decreasing for $x > 0$, quadratic utility is not a realistic model of preferences, as has often been pointed out, e.g. by Dybvig (1992). Quadratic utility penalizes the gain due to a hedge's excess over the liability to be covered, as well as the loss due to shortfall with respect to the liability. The same charge has been leveled against mean–variance portfolio analysis. Markowitz (2002, pp. 155–156) responds:

> ... the problem was to reconcile the use of single-period mean–variance analysis by (or on behalf of) an investor who should maximize a many-period utility function. My answer lay in the observation that for many utility functions and for probability distributions of portfolio returns "like" those observed in fact,

> one can closely approximate expected value of the (Bellman 1957 "derived") utility function knowing only the mean and variance of the distribution.

For details, see the references Markowitz (2002) cites after this quote. It would be interesting to investigate how well the mean and variance can approximate the derived utility of hedged portfolios resulting from OTC market-making.

One might try to separate the problems of hedging, to be solved with a quadratic approach for tractability, and optimal investment, to be solved with an appropriate utility function. However, Dybvig (1992) provides a negative result for the case where incompleteness is due to nonmarket risks: this separation does not occur except with constant absolute risk aversion (exponential utility) and independence of the hedging residual and the marketed risks.

Föllmer and Schweizer (1991) developed a martingale decomposition theorem that yields a *locally risk-minimizing* hedging strategy for a payoff X, where risk is instantaneous or one-step variance. The solution relates to the *minimal martingale measure* $\hat{P}$. For senses in which $\hat{P}$ is minimal, relating both to quadratic and entropy criteria, see Schweizer (1999). In local risk minimization, it is standard to optimize over hedging strategies that need not be self-financing. A non-self-financing portfolio strategy has an associated cost process C, where $C(t)$ is the cumulative cash influx required to rebalance the portfolio over the time interval $[0, t]$. At each instant t, a locally risk-minimizing strategy minimizes $\mathrm{E}[(C(T) - C(t))^2|\mathcal{F}_t]$, the conditional expectation of the squared cumulative future costs, without regard to past costs. A locally risk-minimizing strategy is "mean-self-financing" in the sense that its cost process is a martingale (Schweizer, 2001, Lem. 2.3), so $C(t) = \mathrm{E}[C(T)|\mathcal{F}_t]$, and thus local risk minimization is equivalent to minimizing the conditional variance of the cumulative cost. In discrete time, a backward recursion shows that this is equivalent to choosing the portfolio weights at time t_i to minimize $\mathrm{Var}[(C(t_{i+1}) - C(t_i))^2|\mathcal{F}_{t_i}]$, the conditional variance of the cost incurred at time t_{i+1}. This method is local in the sense that it involves one-step optimizations, and in the sense that an infinitesimal perturbation of the locally risk-minimizing strategy must increase the variance of the cost over the next step or instant. The optimal cost process is orthogonal to the gains process of the locally risk-minimizing strategy, which is a projection of the $\hat{P}$-conditional expectation process of X (Pham, 2000, Thm. 4.2).

The *mean–variance optimal* self-financing hedging strategy minimizes $\mathrm{E}[(Y - X)^2]$, the variance of the hedging residual. This global quadratic criterion relates to the *variance-optimal martingale measure* $\tilde{P}$ (Schweizer, 1996), which is a minimum-distance measure (Section 4.4) based on L^2-distance. Bertsimas et al. (2001) provide a stochastic dynamic programming algorithm for computing the mean–variance optimal hedging strategy. This hedging problem can be studied by means of martingale measures or backward stochastic differential equations: for recent work on the latter, see Lim (2004) and references therein.

Heath et al. (2001) provide a theoretical and numerical comparison of the local and global quadratic approaches. The following example illustrates the difference between a local and global approach in the quadratic setting.

Example 6.1. Continuing Example 4.1, suppose that a trader wishes to hedge the sale of a contingent claim paying \$1 in state 1.

The locally risk-minimizing hedge is $\xi_1 = 0.1$, $\xi_2^{(0)} = \xi_2^{(-)} = 0$, and $\xi_2^{(+)} = 0.2$. The variance-optimal hedge is $\xi_1 = \xi_2^{(0)} = \xi_2^{(-)} = 0$ and $\xi_2^{(+)} = 0.33$. The cost processes associated with these hedges are given in Table 7. The total cost is the hedging residual. Its variance is minimized by the variance-optimal hedge, yielding a variance of 0.037, as opposed to 0.047 for the locally risk-minimizing hedge, which does not take into account the partial cancellation of costs incurred at different times in state 2. The conditional variances at time 1 of the cost incurred at time 2 are 0 when the risky asset's price is 0 or -1, under either hedging scheme, and 0.133 or 0.222 for the locally risk-minimizing and variance-optimal hedges respectively, when the risky asset's price is 1. The unconditional variance of the cost incurred at time 1 is 0.007 or 0.037 for the locally risk-minimizing and variance-optimal hedges respectively.

Suppose that the set R of replicable payoffs is a linear space. The quadratic criteria behave linearly in the sense that, if the hedge Y is optimal for a payoff X, then for any multiple $\gamma \in \mathbb{R}$, γY is optimal for γX. Consequently, the quadratic methods result in unique prices and select a single martingale measure $\hat{P}$ or $\tilde{P}$.

However, it is not appropriate to interpret an expected discounted payoff under $\hat{P}$ or $\tilde{P}$ as a price. As suggested earlier, because quadratic utility does not model preferences well, these prices may not be compatible with the trader's preferences (Bertsimas et al., 2001). Moreover, they may violate the no-arbitrage bounds. The measures $\hat{P}$ and $\tilde{P}$ may be *signed*, that is, they may assign negative values to some events. Pricing under a signed measure

Table 7.
Quadratic Hedging Cost Processes.

State	Probability	Locally risk-minimizing			Variance-optimal		
		Time 1	Time 2	Total	Time 1	Time 2	Total
1	1/9	\$0.1	\$0.4	\$0.5	\$0.33	\$0	\$0.33
2	1/6	\$0.1	−\$0.4	−\$0.3	\$0.33	−\$0.67	−\$0.33
3	1/18	\$0.1	\$0	\$0.1	\$0.33	\$0	\$0.33
4	1/6	\$0	\$0	\$0	\$0	\$0	\$0
5	1/6	\$0	\$0	\$0	\$0	\$0	\$0
6	1/6	\$0.1	\$0	\$0.1	\$0	\$0	\$0
7	1/6	\$0.1	\$0	\$0.1	\$0	\$0	\$0

would imply willingness to pay to give away a lottery ticket, i.e. Arrow–Debreu security, for such an event (Schweizer, 1995). The reason this happens is precisely that quadratic utility penalizes gains as well as losses, so its marginal utility may be negative. For examples of arbitrage resulting from quadratic pricing, see Schweizer (1995, §5) or Frittelli (2000b, p. 50). For similar reasons, jump processes in continuous time pose difficulties for the quadratic approach: there may be negative marginal utility for wealth in a state in which a jump in marketed asset prices causes the optimal portfolio's value to exceed the liability X. An example of what can go wrong occurs in Example 6.1, where $\hat{P}(\omega_1) = \tilde{P}(\omega_1) = 0$, so the optimal initial capital for local or global quadratic hedging of the Arrow–Debreu security for state 1 is zero.

According to Biagini and Pratelli (1999), in discrete time or with jumps, the results of local risk-minimization depend on the numéraire: the hedging strategy depends on whether the costs of the portfolio, which is not self-financing, are measured in units of cash, bonds, stocks, etc. One response to this is that the trader should simply choose the numéraire such that the variance of costs as measured in this numéraire best describes his preferences. However, this observation draws attention to a theoretical shortcoming of using strategies that are not self-financing: costs which are cashflows at different times are simply added, ignoring the time value of money. This may not be a significant issue unless long time spans or high interest rates are involved.

7 Entropy and exponential utility

Another special case of expected utility indifference pricing uses *exponential utility*, also known as *negative exponential utility*, which may be conveniently expressed as $u(x) = 1 - \exp(-\alpha x)$. It has the feature of constant absolute risk aversion, which can produce theoretically elegant results, such as separation of hedging and investment decisions, and independence of the indifference price in Eq. (8) of the initial budget c. Also interesting is the relationship between maximization of exponential utility and minimization of relative entropy $\mathrm{E}_{\mathbf{Q}}[\ln(\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P})]$. The marginal exponential utility indifference price is the expected discounted payoff under a minimum-distance measure (Section 4.4), the *minimal entropy martingale measure* (MEMM) having minimal relative entropy with respect to the statistical probability measure $\mathbf{P}$ (Frittelli, 2000b; Rouge and El Karoui, 2000). Relative entropy also appears in Section 10 as a way of quantifying ambiguity.

Delbaen et al. (2002) cover the topic of exponential utility maximization and valuation via the MEMM with special attention to the set of feasible portfolio strategies over which the optimization occurs. Becherer (2003) gives a general presentation and more explicit results in a special case in which the financial market is complete, but one must value payoffs that depend also on risks independent of the financial market. Mania et al. (2003) discuss special cases in which the MEMM can be constructed explicitly. Another explicit example,

with intuition, and an algorithm for indifference pricing in a similar setting are in Musiela and Zariphopoulou (2004a, 2004b). Fujiwara and Miyahara (2003) discuss representation of the MEMM in terms of Esscher transforms when the underlying process is a geometric Lévy process, giving as examples Brownian motion plus a compound Poisson process, a stable process, and the variance gamma process.

Under some conditions, including restrictions on the form of the mean–variance ratio, the minimal martingale measure coincides with the MEMM (Mania et al., 2003, Prop. 3.2). The minimal martingale measure $\hat{P}$ (see Section 6) is the solution to the dual of the problem of maximizing exponential utility given an initial endowment equal to a multiple of the mean–variance ratio (Delbaen et al., 2002, Thm. 5.1). An alternative is to minimize the entropy-Hellinger process instead of relative entropy. Choulli and Stricker (2005) develop this approach and show that it corresponds to the neutral derivative prices of Kallsen (2002a), for which see Section 9.1, and that it selects the minimal martingale measure (Section 6) when the discounted price process is continuous. Choulli et al. (2006) provide an extension of this approach and a more general framework including it and other minimum-distance measures.

8 Loss, quantiles, and prediction

What unifies the ideas covered in this section is an emphasis on the *loss* or *shortfall* $(Y - X)^-$ associated with hedging the sale of the payoff X by acquiring the payoff Y. They ignore the positive part of the hedging residual, $(Y - X)^+$. Unfortunately, the nomenclature surrounding these methods is a bit confusing: they may also involve a *loss function* ℓ, which is another way of expressing utility: $\ell(x) = -u(-x)$. Minimizing the expected loss is then the same as maximizing expected utility, so pricing via expected loss minimization could be understood as a special case of expected utility indifference pricing. That is, the trader would be seeking the cheapest hedge Y such that $\mathrm{E}[\ell((V + Y - X - B)^-)] \leqslant \mathrm{E}[\ell((V - B)^-)]$, where V is the endowment and B is a benchmark relative to which losses are measured, possibly zero. (If $V = B = 0$, the resulting indifference prices are the no-arbitrage bounds, because gains are ignored and cannot make up for losses.) However, this is not the way that loss minimization has usually been treated.

This literature primarily addresses the problem of minimizing expected loss given a fixed initial budget with which to hedge a liability, without reference to an endowment payoff V. The focus is solely on the shortfall of an approximate hedge. This literature also addresses the very closely related problem of determining the minimal required initial budget to hedge so that expected loss does not exceed some prespecified threshold.

We will consider how the latter problem may apply to pricing in incomplete markets. The approach falls into the framework described in Section 4.2, with the acceptance set $A = \{Z \mid \mathrm{E}[\ell(Z^-)] \leqslant p\}$. Because the loss function ignores

gains, if A is nontrivial, it must include a payoff $Z < 0$. Therefore, pricing a payoff X as the minimal initial budget required to hedge so that expected loss is less than p can result in giving the counterparty an arbitrage. For example, the minimal cost of a replicable payoff Y subject to the constraint $\mathrm{E}[\ell(Y^-)] \leqslant p$ for $p > 0$ may be negative. This method is not generally sound for OTC pricing, as illustrated in Example 8.1.

The two following subsections describe two particular choices of loss function, whose original application was for hedging given a capital constraint, and show that this expected loss methodology should not be transposed directly to the application of OTC pricing.

8.1 Expected shortfall

The choice $\ell(x) = x$ is minimization of expected shortfall. For theoretical results, see Cvitanić (2000), who discusses the form of the optimal hedge in a market that is incomplete due to stochastic volatility or trading constraints.

Example 8.1. Continuing Example 5.1, consider the minimal initial capital required to hedge the straddle given a constraint on expected shortfall.

Figure 2 shows this initial capital as a function of the level p of the constraint. The lower curve in Fig. 2 has negative values for large p because the optimal "hedge" has a negative value in some states, and its cost is negative. The upper curve gives the initial capital required to attain the expected shortfall constraint given the additional constraint that the hedge must be nonnegative; this constraint is appropriate only when treating nonnegative payoffs X such as the straddle. The result is that for $p = \$12$, which is the expected shortfall of the unhedged straddle, the required initial capital is \$0. This is still below

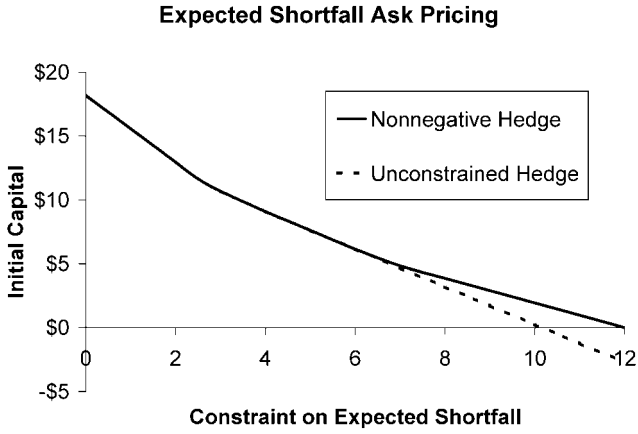


Fig. 2. Cost of hedging a straddle to achieve an expected shortfall constraint.

the lower no-arbitrage bound, which is \$2.72. For $p = 0$, the initial capital equals the no-arbitrage upper bound. The initial capital required for hedging a replicable payoff does not, in general, equal its market price. For example, the no-arbitrage price is \$0 for an equity swap replicated by a portfolio that is long 1 share of stock and short approximately 97 shares of the bond. However, with an expected shortfall constraint of $p = 0.25$, the price assigned is $-\$0.26$.

8.2 *Quantile hedging*

Another special case, $\ell(x) = \mathbf{1}\{x > 0\}$, is known as *quantile hedging* (Föllmer and Leukert, 1999). With this loss function, the hedger tries to minimize the probability of a positive shortfall, without regard to the magnitude of shortfall. Alternatively, one might try to apply quantile hedging to pricing by finding the minimal initial budget required to hedge so that the probability of a positive shortfall does not exceed p. The special case $p = 0$ results in superreplication, that is, any feasible hedge Y satisfies $Y \geqslant X$, corresponding to the no-arbitrage upper bound.

However, for $p > 0$, the method may not work: if there is an event F such that $\mathbf{P}[F] \leqslant p$ and a replicable payoff Y_F of negative price $\pi(Y_F)$ such that $Y_F \mathbf{1}_F \geqslant 0$, the optimization $\min_{Y \in R}\{\pi(Y) \mid \mathbf{P}[Y - X < 0] \leqslant p\}$ tends to be unbounded. For example, if the space of replicable payoffs R and market prices π are linear, and $Y^* \geqslant X$ is a superreplicating payoff, then $Y^* + \lambda Y_F$ is feasible for all $\lambda \in \mathbb{R}$, so the optimization is unbounded. Even if portfolio constraints and nonlinear market prices render the optimization bounded, the results are still likely to be unusable. The optimal solution will tend to involve, for any X, a large negative price to be paid to the buyer of X, funded by incurring large liabilities on some event F of sufficiently small probability. The more complete the market is, the worse this problem will be, as it becomes easier to concentrate liabilities on events of low probability but high state price. One way to ameliorate this problem is to restrict the hedge to be nonnegative (Föllmer and Leukert, 1999). This still leaves the methodology with the same deficiencies as for expected shortfall.

8.3 *Statistical prediction intervals*

Related to quantile hedging is a statistical approach based on prediction intervals for financial quantities such as cumulative interest rates and volatility over an option's life (Mykland, 2003a, 2003b). Quantile hedging looks for a hedge Y that covers the liability X on some event F_X of probability p, i.e. $\mathbf{1}_{F_X}(Y - X) \geqslant 0$ and $\mathbf{P}[F_X] = p$. By contrast, this statistical approach specifies a fixed event G, a prediction interval of probability p used for all payoffs X, and requires that Y satisfy $\mathbf{1}_G(Y - X) \geqslant 0$. This makes the bounds wider for this prediction interval approach than for quantile hedging at the same error level p, assuming the same statistical probability measure $\mathbf{P}$ in both cases. An advantage of the prediction interval approach is that it need not be based on a

single probability measure **P**. Without assuming a specific model for stochastic volatility and interest rates, much less that its parameters are known, Mykland (2003a, 2003b) works out bounds for European option prices and the related hedging strategies in a diffusion setting, given prediction intervals for cumulative volatility $\int_0^T \sigma^2(t)\,\mathrm{d}t$, or for this and cumulative interest rates $\int_0^T r(t)\,\mathrm{d}t$ together.

Because it is similar to quantile hedging, this prediction interval approach has a similar drawback as a method for OTC pricing: it assigns zero value to payoffs that are zero inside the prediction interval but positive outside it, which allows arbitrage. The prediction interval approach may be most useful in risk management, for reducing model risk (Mykland, 2003b, §1) or in formulating a liquidation strategy for a trade (Mykland, 2003a, §6).

9 Pricing kernel restrictions

One way of expressing the problem of pricing in incomplete markets is that total ignorance about the pricing kernel Π allows the price to be anywhere within the no-arbitrage price bounds (Section 2.3). This suggests that one may apply a restriction to the pricing kernel to get price bounds. The main idea is that some of the pricing kernels that are possible, in the sense of repricing all marketed securities, are economically implausible. One basis for this is to assert that some pricing kernels make some of the replicable payoffs into objective good deals (Section 4.2.4), and it is implausible that such good deals should exist. That is, one may exclude pricing kernels that would result in too good a deal for a typical investor or most investors.

An early approach, not related to good deals, is to impose restrictions on the moments of asset prices under the pricing measure, rather than directly on the moments of the pricing kernel. If the statistical probability measure **P** is known, this restriction on the pricing measure **Q** is equivalent to a restriction on the pricing kernel $\Pi = D\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$. Lo (1987) applied restrictions on a stock's variance under **Q** to pricing an option on that stock. Further research has incorporated restrictions on higher moments and developed computational algorithms. It may seem advantageous that the bounds derived from **Q**-moment restrictions do not depend on the statistical probability measure **P**, and thus do not require a choice of statistical model – but from where does knowledge of **Q**-moments come? Lo (1987) showed that, for two simple models, the **Q**-variance can be computed from the **P**-variance under the statistical measure, and that method-of-moments estimation yields the same result for the two models. However, in general, the need to connect **Q**-variance to estimable quantities can introduce dependence on a model.

9.1 Low-distance pricing measures: pricing kernels and good deals

Pricing by minimum-distance measure (Section 4.4) selects the single pricing measure **Q** with the lowest distance from the statistical probability measure **P**, or equivalently, the corresponding pricing kernel or likelihood ratio d**Q**/d**P** representing the lowest distance. For example, the relative entropy distance (Section 7) is a function of the likelihood ratio. A modification of this method is to select a set of pricing measures $\{\mathbf{Q} \mid d(\mathbf{P}, \mathbf{Q}) < \epsilon\}$ with low distance d from **P**. The distance constraint is equivalent to a restriction on the pricing kernel. It may be more convenient to consider restrictions directly in terms of the pricing kernel.

One approach is to place restrictions on the moments of the pricing kernel, which can be translated into restrictions on the assets' returns. Hansen and Jagannathan (1991) discussed relations between the mean and variance of the pricing kernel, connecting this to assets' Sharpe ratios. Cochrane and Saá-Requejo (2000) adapted these results to asset pricing and initiated the phrase "good-deal bounds" for their price bounds based on a ceiling for the variance of the pricing kernel. The point is to bound the prices of payoffs based on the assumption that they should not have Sharpe ratios that are too high. Here "too high" means more than some arbitrary multiple of the highest Sharpe ratio of any replicable payoff. A similar approach to establishing bounds is taken by all papers discussed in this section.

Bernardo and Ledoit (2000) have an approach very similar to that of Cochrane and Saá-Requejo (2000), restricting not the Sharpe ratio, but the gain–loss ratio $\mathrm{E}_{\mathbf{Q}}[X^+]/\mathrm{E}_{\mathbf{Q}}[X^-]$ of any payoff X replicable at zero cost. Here **Q** is a benchmark pricing measure: although one might not trust it to assign unique prices to all contingent claims, it can serve as the basis for assessing whether a deal is good in the sense that gains outweigh losses. Subjective considerations might be taken into account through the choice of benchmark pricing kernel. Bernardo and Ledoit (2000) relate the gain–loss ratio restriction to a restriction not on the pricing kernel's variance, but to bounds on the ratio between the pricing kernel and the benchmark pricing kernel. However, as Černý (2003, pp. 195–196) points out, it may not be possible to find any other pricing kernels that satisfy such a bound at any finite level. For example, in the Black–Scholes model, the ratio between pricing kernels is proportional to a power of the stock price, based on the equation $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P} = \exp(-(\lambda^2/2)T - \lambda B(T))$, where λ is the market price of risk, B is Brownian motion under **P**, and T is the time horizon: this ratio is unbounded because $B(T)$ is unbounded.

A drawback of the Sharpe ratio approach is that the Sharpe ratio is a poor measure of preference, especially for derivative securities having nonlinear payoffs. As Bernardo and Ledoit (2000, p. 166) point out, this can cause the lower good-deal bound based on pricing kernel variance (Sharpe ratio) for an out-of-the-money call option to be zero, because the upside variance is too great. As Černý (2003, p. 193) illustrates, one payoff may stochastically domi-

nate another, while the latter has a higher Sharpe ratio than the other. These problems relate to the defects of quadratic utility (Section 6).

Summaries of the main points and mathematical results of Bernardo and Ledoit (2000) and Cochrane and Saá-Requejo (2000) can be found in Geman and Madan (2004). For an example of good deal bounds, involving Sharpe ratios, applied to pricing European options on a stock following the Heston stochastic volatility model, see Bondarenko and Longarela (2004, §4.2). Björk and Slinko (2006) provide a solid mathematical foundation for good deal bounds based on Sharpe ratios in the case of a continuous-time underlying price process that has jumps.

Černý (2003) proposes to adapt the approach of Cochrane and Saá-Requejo (2000) by replacing the Sharpe ratio, with its connection to quadratic utility, with a *generalized Sharpe ratio* based on a more suitable utility function. For example, using exponential utility corresponds to a bound on relative entropy (Section 7), power utility corresponds to a bound on the expectation of a negative power of the pricing kernel, and log utility corresponds to a bound on the expected log of the pricing kernel. A 6-period example of a call option shows that the good deal bounds do not depend very much on the choice of utility function, i.e. of which generalized Sharpe ratio to use, but depend strongly on the level of the bound which defines the set A of good deals (Černý, 2003, §4.2). This makes sense, as changing the utility function changes the shape of A, while changing the bound changes the size of A.

In a dynamic model, one can implement the pricing kernel restrictions globally or locally (see Section 4.1.2). The local approach in continuous time rules out *instantaneous good deals*, forbidding any pricing kernel such that, if one could trade all claims frictionlessly at the prices it assigns, one could increase expected utility at too fast a rate at any instant. This relates to the local utility maximization of Kallsen (2002a), whose "neutral derivative pricing" assigns prices such that the opportunity to trade in derivatives does not allow for greater local utility than does trading in marketed securities alone. He derives price bounds by considering prices that are consistent with a limited nonzero position for the derivative security within an optimal portfolio. Exploring the local approach, Černý (2003, §5.1) concludes that, if the discounted gains processes from portfolio strategies are Itō processes, then ruling out instantaneous good deals imposes, for all utility functions having the same coefficient of absolute risk aversion, the same bound on the norm $\|\lambda(t)\|$ of the market price of risk vector stochastic process λ. As the minimal martingale measure of Section 6 corresponds to a pricing kernel $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P} = \exp(-\int_0^T \|\lambda(t)\|/2\,\mathrm{d}t - \int_0^T \lambda(t)\,\mathrm{d}B(t))$ where $\|\lambda(t)\|$ is minimal for each t, this means the instantaneous good deal bounds always contain the value assigned by the minimal martingale measure (Černý, 2003, §7).

9.2 Equilibrium and stochastic dominance

Other pricing kernel restrictions are related to equilibrium among expected utility maximizers. Structural considerations can impose direct restrictions on the pricing kernel, and bounds for prices of nonreplicable payoffs can be constructed by comparison to replicable payoffs.

A structural feature of equilibrium among expected utility maximizers is that the pricing kernel should be decreasing in aggregate wealth (or consumption, depending on the economic model). Usually, the pricing kernel is decreasing in the price of an asset that is held in net positive supply, unless it has a negative association with aggregate wealth. For example, if there are two assets, a stock in net positive supply and a bond in net zero supply, the set of pricing kernels can be restricted to include only those that are decreasing in the stock price. Chazal and Jouini (2004) show that this restriction can significantly tighten the option pricing bounds when added to restrictions on the first two moments in the manner of Lo (1987). This approach goes back at least to Perrakis and Ryan (1984), who also initiated a literature on option price bounds based on comparisons among portfolios.

That method might be thought of as ruling out comparative good deals. Perrakis and Ryan (1984) used the CAPM pricing rule, in which the expected return of a portfolio is an affine function of the covariance between a representative investor's marginal utility of consumption and the portfolio's final value. Although this marginal utility and the distribution of portfolio value may not be known, the comparison of three portfolios allows Perrakis and Ryan (1984) to formulate bounds for the price of a European call option in a model with one stock and bond. The lower and upper bounds involve an expectation of a function of the terminal stock price, discounted at either the risk-free rate or the stock's expected return, respectively. To use the bounds, one need not know the statistical probability measure **P**, but one must know the **P**-expectations of some functions of the terminal stock price. Various extensions have been derived, involving intermediate trading, transaction costs, and puts. An apparent limitation of the methodology is the necessity, for each new security, to identify new comparison portfolios. For a review of subsequent literature related to Perrakis and Ryan (1984), see Constantinides and Perrakis (2002, §1).

Bizid and Jouini (2005) point out that an equilibrium among various agents in an incomplete market need not coincide with the equilibrium in a completion of that market. They demonstrate that the bounds imposed by very weak equilibrium conditions in an incomplete market, without assuming that the pricing kernel is a nonincreasing function of aggregate consumption, might be wider than those that result from considering any possible completion of that market. They view the reliance of Perrakis and Ryan (1984) on the CAPM as invoking a possible, but unknown, completion of the market. On the other hand, the CAPM might be justified not by market completeness, but by invoking the approximately quadratic preferences of well-diversified investors. Moreover, Constantinides and Perrakis (2002) rederive and extend results of

Perrakis and Ryan (1984) and related work, while relying on stochastic dominance rather than the CAPM.

They use stochastic dominance considerations to rule out option prices that allow trades in the option to increase expected utility, versus a portfolio only of marketed securities, under all increasing, convex utility functions. If a derivative security were offered for less than this lower bound, any expected utility maximizer would prefer to buy some of it than to keep all wealth invested in the market. This is very much in the spirit of bounding the Sharpe ratios, etc., of all payoffs to be no more than a certain multiple of the maximum Sharpe ratio of a replicable payoff. The differences are that the multiple is fixed at 1, and no particular measure such as a generalized Sharpe ratio is used, rather, a price is excluded only if it gives rise to an increase in any expected utility. The CAPM approach is even more similar, with the excess return "alpha" in the CAPM substituting for the Sharpe ratio.

For stochastic dominance constraints in optimization, which may be applied to portfolio optimization or pricing, see Dentcheva and Ruszczyński (2003).

10 Ambiguity and robustness

Risk, as something that can be quantified by means of a probability distribution, is to be distinguished from *ambiguity* or *Knightian uncertainty*, which represent a greater degree of ignorance. (Sometimes "uncertainty" is used more broadly to include both risk and ambiguity.) When we can assign a probability distribution and compute risk, we know something. For example, suppose we can assign to a potentially infinite sequence of repeatable experiments a probability measure such that the experiments are independent, and an event F has probability 30% of occurring in any repetition. Although we do not know whether F will occur in the next repetition, we do know, by the law of large numbers, that the fraction of experiments in which F occurs will eventually be between 29.99 and 30.01%. If we do not possess such knowledge, then we cannot assign a probability measure to this phenomenon, and we may require concepts such as that of imprecise probability (Walley, 1991). We may, for example, regard all probability measures in a set $\mathcal{P}$ as plausible, and all those not in $\mathcal{P}$ as implausible, and assign $\inf_{\mathbf{P}\in\mathcal{P}} \mathbf{P}[F]$ and $\sup_{\mathbf{P}\in\mathcal{P}} \mathbf{P}[F]$ as bounds for the probability of F. There is a substantial literature devoted to the Ellsberg (1961) experiment, which showed that such considerations affect willingness to gamble: subjects prefer to bet on unambiguous gambles rather than ambiguous ones, and these preferences are not consistent with maximizing any expected utility function. This may be because subjects have no faith that they can describe an ambiguous gamble with a single probability measure.

The unreliability of our stochastic models of financial markets suggests that ambiguity should be an important consideration in financial engineering. As discussed in Section 3.3, to be able to hedge all payoffs perfectly, the hedger must be in a complete market and know the stochastic process that marketed

security prices follow. Ambiguity about this stochastic process is a source of effective incompleteness, as it becomes impossible to find the perfect hedge.

The theme of the application of ambiguity to incomplete markets is the decomposition of uncertainty about eventual outcomes into risk and ambiguity. A trader's aversions to delaying consumption (intertemporal substitution), to risk, and to ambiguity all determine the price at which he is willing to trade. Aversion to ambiguity is often described as a desire for robustness to misspecification of the stochastic model. A common approach in the financial literature (Chen and Epstein, 2002; Anderson et al., 2003; Maenhout, 2004) is to consider an equilibrium in which all traders have the same preferences, resulting in an analysis of assets' equilibrium expected returns in terms of market prices of risk and of ambiguity. Anderson et al. (2003) conclude, "Because mean returns are hard to estimate, ... there can still be sizable model uncertainty premia in security prices," and Maenhout (2004) concurs: "Empirically a 3% to 5% wedge is difficult to detect given the usual length of available time series. Given plausible values of risk aversion and uncertainty aversion, an equilibrium equity premium between 4% and 6% can then be sustained." Liu et al. (2005) proceed in similar fashion, but consider only ambiguity about rare jump events, not diffusion coefficients, and examine the impact on option prices. This main stream of financial research, discussed in Section 10.1, is an example of equilibrium marginal indifference pricing. A somewhat different, subjective approach occupies Section 10.2.

10.1 Complete preferences

In Section 4.2.3, expected utility maximization was criticized as a basis for portfolio optimization or derivative security pricing because it is too sensitive to unknown inputs, such as the probability measure. This defect has inspired work on *robust utility*. The literature looks back primarily to Gilboa and Schmeidler (1989), who considered portfolio optimization in which the expected utility $\mathrm{E}[u(V)]$ of random wealth V is replaced by

$$U(V) = \inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}_{\mathbf{P}}\big[u(V)\big], \tag{13}$$

where $\mathcal{P}$ is a set of plausible probability measures or *multiple priors*. Given $\mathcal{P}$, one can choose a portfolio to maximize the robust utility of Eq. (13), which as a preference function specifies complete preferences and is a foundation for indifference prices (Section 4.2.2). Talay and Zheng (2002) describe such an approach to derivative security pricing given considerations of model risk.

Although the form of Eq. (13) makes it look like a convex risk measure (Föllmer and Schied, 2002), not all convex risk measures have an interpretation in terms of ambiguity or robustness. For example, Schied (2004) describes the problem of maximizing robust utility functionals, equivalently, minimizing convex risk measures, subject to a capital constraint. He provides more explicit results for law-invariant risk measures ρ, where $\rho(X)$ depends only on the

law of X under $\mathbf{P}_0$, a reference probability measure. This lacks the interpretation of ambiguity, in which the law of X under other measures also counts. An example of a law-invariant risk measure is expected shortfall, defined by $\mathcal{P} = \{\mathbf{P} \mid \mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}_0 \leqslant r\}$, i.e. a pointwise (almost sure) constraint on the likelihood ratio. This considers only one probability measure $\mathbf{P}_0$, but all conditional probability measures $\mathbf{P}_0[\cdot|F]$, where F is an event such that $\mathbf{P}_0[F] < 1/r$.

The pointwise constraint on the likelihood ratio contrasts with a constraint on the relative entropy $\mathrm{E}_{\mathbf{P}}[\ln(\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}_0)]$. The set

$$\mathcal{P} = \left\{\mathbf{P} \mid \mathrm{E}_{\mathbf{P}}\left[\ln(\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}_0)\right] < \epsilon\right\} \tag{14}$$

can be interpreted as a set of probability measures that are plausible, given that econometric inference leaves $\mathbf{P}_0$ as the best estimate, but the econometrician remains uncertain as to the true probability measure. After estimation, some probability measures are more plausible, i.e. have a higher p-value or posterior likelihood, than others. An entropy criterion can be tractable, at least if one works with the intersection of $\mathcal{P}$ in Eq. (14) with a family of models, such as diffusions. However, entropy may not be a suitable way of describing which probability measures are plausible. It may be that different events have different levels of ambiguity, or that some aspects or parameters of the model are more ambiguous than others. For example, practitioners of financial engineering often have less confidence in their estimates of correlations or of means than of volatilities. Interesting effects arise when one considers that there may not simply be one correct entropy penalty or constraint for all traders to use in accounting for the ambiguity surrounding a probability measure estimated from commonly available data. Some assets may be more ambiguous than others, which can lead to under-diversification (Uppal and Wang, 2003) or cause negative skewness in short-term returns and premia for idiosyncratic volatility (Epstein and Schneider, 2005). Different traders may assign different levels of ambiguity to assets, which could explain the home-bias puzzle in investments (Epstein and Miao, 2003) and limited participation in the stock market (Cao et al., 2005).

Anderson et al. (2003) consider a portfolio-optimizing econometrician who wishes to construct a portfolio whose utility is robust with respect to the ambiguity about the true probability measure, that is, is high for all alternatives which remain plausible given the observed data. This leads them to an optimization including a penalty proportional to the relative entropy between each model under consideration and the best-fit model. Results can be computed using a worst-case model among those that are plausible. Maenhout (2004) considers a more tractable version of this methodology, in which the entropy penalty depends on wealth in a way that makes the optimal portfolio weights wealth-independent, and gives some more explicit results. An alternative to a penalty on relative entropy is a constraint on relative entropy, as in Eq. (14). Entropy penalty and entropy constraint model different preferences, but not only do they both result in the use of a worst-case model, entropy-penalty and entropy-constraint problems come in pairs sharing the same solution, i.e.

worst-case model and optimal portfolio (Hansen et al., 2006, §5). Thus, it is not possible to deduce whether a trader's portfolio is the result of solving a problem with an entropy penalty or constraint.

This issue is known as *observational equivalence*, and it appears frequently in the finance literature. Skiadas (2003) shows that, in a market driven by Brownian motion, the entropy-penalty value function coincides with that of stochastic differential utility (SDU): see also the discussion of source-dependent risk aversion in Skiadas (2006). Maenhout (2004) shows that his homothetic version of the entropy-penalty formulation is also observationally equivalent to SDU, but emphasizes that this observational equivalence is limited to portfolio choice and asset prices within a single model. The observational equivalence arises because the solution to the portfolio optimization with robust preferences reduces to the use of a worst-case model, which can then be mapped to a specific case of SDU. However, if market opportunities change, then the worst-case model will also change, becoming equivalent to a different case of SDU, so the observational equivalence breaks down in a broader context (Chen and Epstein, 2002, §1.2).

Moreover, from a financial engineering perspective, different methods that may yield the same answer given different inputs are different. As instrumental rather than descriptive devices, one method may be superior: it may be easier to specify good inputs and compute a useful result with one method than the other. For example, when preferences featuring risk aversion and ambiguity aversion are observationally equivalent to preferences featuring risk aversion only, the level of risk aversion is greater in the latter case. It would be easier to specify risk aversion and ambiguity aversion by introspection than to guess what level of risk aversion alone yields the same price. Indeed, Maenhout (2004) uses ambiguity aversion to explain the equity premium puzzle, which is that the level of risk aversion required to justify an expected return for equities matching the historical average is implausibly high when compared to the level of risk aversion that most subjects display when confronted with unambiguous gambles (Mehra, 2003; Mehra and Prescott, 2003). However, it may be that they display much greater risk aversion in financial markets, much of which is actually generated by aversion to these markets' ambiguity. Liu et al. (2005) find that aversion towards ambiguity about rare events involving jumps in the aggregate endowment can account for option pricing smirks. Routledge and Zin (2004) model fluctuations in the liquidity supplied by market-makers who have multiple priors, giving explicit, simple examples of OTC option trading, with the market-maker's optimal bid, ask, and hedges based on robust utility.

It is possible to construct a set $\mathcal{P}$ of multiple priors on principles other than entropy. One major motivation for not using the tractable entropy methodology is *dynamic consistency*. Various versions of dynamic consistency have been much discussed in the recent literature on risk measures: see Roorda and Schumacher (2005). Roughly speaking, dynamic consistency means that for payoffs X and Y occurring at time T, if X will always be preferred to Y at time $t < T$, then X must be preferred to Y at any time $s < t$. To do otherwise would create

inconsistency between choices at different times. Whether such inconsistency is unacceptable depends on the application: for example, it is more troublesome in regulation than in pricing OTC securities. Such inconsistency might even be appropriate given certain kinds of beliefs incorporating ambiguity: see Epstein and Schneider (2003, §4) and Roorda et al. (2005, §4).

Dynamic consistency requires that the set $\mathcal{P}$ of multiple priors be *rectangular*. In a discrete-time model, this means that $\mathcal{P}$ has the following property: for any event F_i that involves only step i, any event $F_{<i}$ that involves only steps $1, \ldots, i-1$, and any pair $P_1, P_2 \in \mathcal{P}$, there must exist a $P_3 \in \mathcal{P}$ such that $P_3(F_i \cap F_{<i}) = P_1(F_i)P_2(F_{<i})$. That is, any one-step conditional probability for step i must appear in combination with all probability measures for steps $1, \ldots, i-1$. For example, if $\mathcal{P}$ contains two probability measures, both of which correspond to multi-period binomial models of a log stock price with independent increments, and one of them says that the probability of an up move is 40% and the other says that it is 60%, then $\mathcal{P}$ is not rectangular. It would also have to contain, among others, a probability measure under which the probability of an up move is 40% at step 1 and 60% at step 2.

Rectangularity leads to the preference structure known as "recursive multiple priors" (Chen and Epstein, 2002; Epstein and Schneider, 2003), which can be viewed as a combination of stochastic differential utility and the robust utility of Equation (13). The set $\mathcal{P}$ of multiple priors defined by an entropy constraint is not rectangular (Epstein and Schneider, 2003; Hansen et al., 2006). Advocates of dynamic consistency suggest enlarging a non-rectangular candidate set of multiple priors until it becomes rectangular, while others (e.g. Hansen et al., 2006, § 9) object that the resulting rectangular set is too large, depriving the modeler of the ability to impose interesting restrictions on probabilities. In terms of pricing in incomplete markets, the result is price bounds that are too wide.

10.2 Incomplete preferences

Using the robust utility of Eq. (13) to define complete preferences as in Gilboa and Schmeidler (1989) is suitable for the application of a one-time portfolio optimization, in which a portfolio strategy is chosen with a pessimistic attitude in the face of ambiguity about which of the probability measures in $\mathcal{P}$ is correct. The result is the selection of a worst-case model $\mathbf{P}^* \in \mathcal{P}$, similar to the least favorable completion mentioned in Section 4.4. The methods discussed in Section 10.1 price all payoffs under an equilibrium pricing measure $\mathbf{Q}^*$ derived from $\mathbf{P}^*$. Assigning this price to all payoffs at all times would reflect an ongoing concern with maximizing expected utility under the worst-case model, and no concern for expected utility under any other plausible model in $\mathcal{P}$.

To see what might be undesirable about this, consider the difference between optimizing over random total wealth and optimizing over a payoff, which is a change in wealth, discussed in Section 4.2.2. Also, whereas indifference pricing is based on complete preferences, no-arbitrage pricing and other

good deal bounds are based on incomplete preferences. No-arbitrage pricing is based on the incomplete preference such that V is weakly preferable to W when $\operatorname{ess\,inf}(V - W) \geqslant 0$, i.e. $V \geqslant W$. When neither $V \geqslant W$ nor $W \geqslant V$, this preference structure expresses neither indifference nor preference between V and W, but rather cannot decide between them. A complete preference structure using the essential infimum as the preference function for total wealth evaluates portfolios based on the worst-case scenario: V is preferred to W if $\operatorname{ess\,inf} V > \operatorname{ess\,inf} W$. This is not a suitable preference structure for financial decisions. According to this preference function, it is better to get one cent for sure than to have a 99.99% chance of getting one million dollars and a 0.01% chance of getting nothing.

The same problem can occur with the Gilboa and Schmeidler (1989) robust utility. If the set $\mathcal{P}$ of plausible measures is large, reflecting a great degree of ambiguity, we may find that a change in the portfolio that increases expected utility under the worst-case measure decreases it under other plausible measures. Then we may lack confidence that this change is an improvement, or even suspect it of being a bad deal. In other words, the acceptance set

$$A_{GS} = \Big\{ Z \mid \inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}\big[u(V+Z)\big] \geqslant \inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}\big[u(V)\big] \Big\} \tag{15}$$

defined by robust utility for use in subjective good deal bounds (see Section 4.2) may not be suitable as a set of good deals.

An alternative is robust evaluation not of total wealth but of changes in it, or equivalently, incomplete preferences over portfolios, as in no-arbitrage price bounds. This corresponds to the incomplete preference scheme of Bewley (2002), in which the acceptance set is

$$A_B = \Big\{ Z \mid \inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}\big[u(V+Z)\big] - \mathrm{E}\big[u(V)\big] \geqslant 0 \Big\} \subseteq A_{GS}. \tag{16}$$

That is, a change is considered a good deal if it increases expected utility under *every* plausible probability measure, not if it merely increases expected utility under the worst-case measure. This smaller acceptance set is more conservative in that it recognizes fewer good deals and thus leads to wider good deal bounds. This Bewley (2002) approach also responds better to an error of wrongly including an implausible measure $\mathbf{P}_x$ in $\mathcal{P} = \mathcal{P}' \cup \{\mathbf{P}_x\}$, where $\mathcal{P}'$ is the correct set of plausible probability measures. Then there might be a payoff Z such that $\inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}[u(V+Z)] > \inf_{\mathbf{P}\in\mathcal{P}} \mathrm{E}[u(V)] \geqslant \inf_{\mathbf{P}\in\mathcal{P}'} \mathrm{E}[u(V)] > \inf_{\mathbf{P}\in\mathcal{P}'} \mathrm{E}[u(V+Z)]$, so that $\mathbf{P}_x$ is the worst-case model and the Gilboa and Schmeidler (1989) approach would have us erroneously switch from V to $V + Z$, which actually makes us worse off. The Bewley (2002) approach focusing on changes in portfolios would only cause us wrongly to reject some good deals, not wrongly accept bad deals.

The question is how aversion to ambiguity manifests itself in OTC market-making. Is one willing to pay high prices for "ambiguity hedges," that is, payoffs that reduce the ambiguity of one's expected utility? Or does one accept only

unambiguously good deals, paying a low enough price so that it is implausible that they do not improve one's portfolio?

11 Calibration

It is standard practice to price with an incomplete-markets model much as described in Section 2.2 for complete-markets models, by calibrating **Q** to marketed securities' prices and assigning the expected discounted payoff $\mathrm{E}_{\mathbf{Q}}[DX]$ as the price for a payoff X. If one calibrates to a family of complete-market models containing the true model, then **Q** must be the unique no-arbitrage pricing measure. However, if the market is incomplete, choosing **Q** such that $\mathrm{E}_{\mathbf{Q}}[DS]$ is the market price for any payoff S of a marketed security does not guarantee that $\mathrm{E}_{\mathbf{Q}}[DX]$ is an arbitrage-free price for any payoff X. This is merely a curve-fitting scheme. Arbitrage-free pricing requires that **Q** be equivalent to the statistical probability measure **P**. Moreover, it is characteristic of incomplete markets that more than one pricing measure **Q** yields arbitrage-free prices.

Researchers who propose new incomplete-markets models of underlying asset prices often provide a formula for a single "risk-neutral" price, which appeals to practitioners. A typical procedure is to posit a model for the statistical probability measure **P**, next to assume that one should look for an equivalent pricing measure **Q** of the same parametric form, and finally to relate the parameters under **P** and **Q**. The last step can be done by means of an unspecified market price of risk (e.g. Heston, 1993), or through construction of an equilibrium among expected utility maximizers, in which case unspecified parameters of the utility function are involved (e.g. Madan et al., 1998; Kou, 2002). This last step is not important in practice, because practitioners calibrate the parameters of **Q** to market prices without any regard to **P**.

What usually happens is that parsimonious models, with a small number of parameters, cannot exactly match the prices of all marketed securities: the models are not perfect.[3] Although multiple pricing measures **Q** are consistent with observed market prices, in practice, none of the probability measures within the family under consideration will be perfectly consistent. Calibration selects the member of the family that is most consistent. The rationale for using prices based on calibration in OTC trading is that because these prices are nearly consistent with market prices, they are likely to avoid arbitrage and to assign reasonable values to payoffs if market prices are reasonable.

That is, if market prices exclude arbitrage and good deals, then it seems likely that OTC prices calibrated to market prices should also exclude arbitrage and good deals. However, the plausibility of this conclusion depends on how

[3] Models with many parameters may fit the data exactly, but their calibrated parameters tend to change substantially over time, a sign that they are not perfect either; they tend to suffer from over-fitting.

similar payoffs of OTC securities are to those of marketed securities. If there are no marketed securities whose prices yield information through the model about events that are important to valuing the OTC securities, the scheme will be unreliable. An example is the calendar spread on swaption straddles discussed in Section 2.3. Good price quotes are available for swaptions expiring on one of a limited set of dates, all of which are much more than two days apart. Therefore, calibration to these swaptions' prices can only give information about the total volatility under **Q** of interest rates over the long periods between adjacent expiration dates, not about how volatility is spread between the expiration dates. Typical calibration schemes interpolate smoothly, assuming that there is no reason for volatility to be concentrated. However, interest rate volatility under **P** *is* concentrated around dates of scheduled major economic announcements. Therefore the prices at which JP Morgan sold the calendar spreads on swaption straddles, although consistent with market prices of swaptions, resulted in a good deal for the customers and a bad deal for JP Morgan. Having a better model could not solve this problem, because market prices do not contain the information required to calibrate the model. What is needed is a better method, one which is grounded in an assessment of statistical probabilities, allowing the trader to base pricing on such information as the concentration of volatility around economic announcements.

12 Conclusion

One might dream of a unified theory of contingent claim valuation in incomplete markets, covering not only derivative securities but also equilibrium pricing of underlying securities and corporate investment via the real options approach. Although these applications have much in common, we have focused entirely on making a market in OTC derivatives, in the belief that the practical settings of these applications differ so much that any valuation methodology should be evaluated differently, depending on the use to which it is to be put. For example, when making a market in derivatives, a trader is concerned that potential customers may possess superior information, while executives making corporate investment decisions are concerned that interested subordinates may have provided biased information about future cashflows; also, hedging is of paramount importance in trading derivatives, but of at most secondary importance in corporate investment. For the application of OTC derivatives market-making, we want a valuation methodology that is robust to misspecification of inputs that are hard to infer, that ensures that each trade made at the bid or ask prices is beneficial, and that is tractable, allowing rapid computation.

What is beneficial may be the subject of some debate, but a suitable valuation methodology should either have an appropriate economic grounding, as expected utility indifference pricing does, or be shown to give answers that agree with the results of a well-grounded method under specified circumstances. The economic grounding should involve the subjective situation of

the market-maker who is considering a trade, including his current portfolio, the risk management framework in which he operates, and his future opportunities; again, an objective methodology that does not take account of the individual's situation would be appropriate only in circumstances in which it could be shown to yield results that are subjectively beneficial. For example, if some objective good deal bounds were wider than a trader's subjective good deal bounds and still narrow enough to be usable, they would be appropriate. It is more helpful to a decision-maker to identify a price at which trade benefits him than to identify a "fair price"; $300,000 might be a fair price for a luxury automobile, but if one cannot resell it, it might not be beneficial to buy it at or near this price.

We conclude by assessing the extent to which various methods achieve these desiderata. Along the way, we will point out some cases in which further research is needed for such an evaluation. We focus on a few major kinds of methods. These include, first, the standard practice of calibrating to market prices without reference to a statistical probability measure. Second are methods based on expected utility maximization and indifference, including marginal indifference pricing or minimum-distance measures, whether founded on local or global criteria. Third, there are methods of pricing kernel restriction founded on constraints, such as pricing with low-distance measures rather than minimum-distance measures. Finally, there are methods that account for ambiguity, whether they deal with it by using just a worst-case model or by considering all plausible models.

The most tractable method is calibration: it prices all payoffs by taking expectations under a single probability measure calculated by a single parametric optimization. Next best are other methods that also price using just one probability measure, such as any minimum-distance method or marginal indifference pricing, whether it is founded on expected utility or robust utility yielding a worst-case model. It appears that it takes more work to identify these single measures than calibration requires. Less tractable than these is non-marginal indifference pricing, which is not simply pricing under one measure: it requires a new optimization to price each payoff. Local variants are more tractable than global variants. Pricing kernel restrictions and robust methods that use optimization over multiple probability measures look most difficult of all. These optimizations may be non-parametric, e.g. requiring computation of a pricing kernel in all states.

Robustness is the aim of the methods founded on multiple statistical probability measures, but it remains to be confirmed by extensive empirical study that the resulting prices are indeed robust to statistical sampling error. Pricing kernel restrictions featuring low-distance measures also use a single statistical probability measure $\mathbf{P}$, but use multiple pricing measures; their robustness too is an open question. Any method involving expected utility indifference or distance minimization, including the quadratic and exponential special cases, is not robust with respect to the statistical probability measure $\mathbf{P}$.

Calibration is more robust to its observed inputs, because it is easier to observe market prices than to infer the parameters of econometric models; however, because price data may be out of date, erroneous, or have noise due to market microstructure and bid–ask spreads, robustness to this data is still an issue. The more parsimonious models tend to be more robust. Calibration is not robust to the choice of the family $\mathcal{P}$ of models within which calibration takes place. The resulting risk of trading losses is known as *model risk*. It would be interesting to know whether model risk can be mitigated by using multiple calibrated models $\hat{\mathbf{Q}}$ from different families $\mathcal{P}$ in the manner of the robust methods (Section 10), or by using all the models from a single family that have sufficiently low calibration error, instead of just the one with minimal error, in the manner of low-distance measures (Section 9).

To ensure that trades made at bid and ask prices are beneficial, it helps to use a method that produces price bounds that are suitable for use as bid and ask prices. When using a method that produces unsuitable price bounds, or a single price, a trader is reduced to intuition in setting bid and ask prices, making it difficult to tell whether trades include adequate compensation for unhedgeable risk.

Expected utility indifference pricing, based on the trader's optimized portfolio, is the paradigm of a method for generating price bounds that are beneficial, but this method's fatal flaw is its lack of robustness. Either the trader must optimize his portfolio according to an unreliable expected utility maximization procedure, or the indifference prices are suited to an imagined optimal portfolio, not his actual portfolio (Section 5.2.1).

The methods founded on marginal indifference pricing and minimum-distance measures have weaker economic grounding than expected utility indifference pricing. It remains to be seen whether and under what circumstances they yield results that are approximately the same as expected utility indifference pricing, despite the apparent flaws of various of these methods, such as producing a single price, using an inappropriate utility function, and an objective orientation that disregards the trader's portfolio. This last point also affects the methods based on low-distance measures. A major issue in using them is the question of how great a distance is "low."

Calibration provides no reason to believe that trading at the resulting price is beneficial. Its successes have much to do with traders' skillful use of their experience and intuition, the ability to hedge well in very competitive OTC markets, and large bid–ask spreads in OTC markets where hedging is harder. Its failures point to its limitations. As hedge funds know, the ability to price all OTC securities well must include an assessment of the statistical probability measure $\mathbf{P}$, which calibration avoids. A synthesis with econometrics is desirable. This returns us to the problem of robustness to statistical errors in specifying $\mathbf{P}$. Methods based on robustness to subjective ambiguity try to overcome this problem while retaining the justifiability of expected utility indifference prices, for instance, by using robust utility. However, it remains to

show how to model and quantify the ambiguity left after econometric inference in a way that yields a useful method for OTC pricing.

A fundamental question is how we derive information from current market prices of derivative securities and from econometric study of underlying securities' price histories. In particular, how do we respond when derivative securities' current prices seem to be out of line with our beliefs about underlying securities' future prices, as expressed in the statistical probability measure **P**, making possible a good deal by trading in marketed securities? If we are not only making a market in OTC securities but also willing to speculate or invest in marketed securities, then this is an opportunity to trade against a perceived mispricing. This trade would generate enough risk for our indifference price bounds for marketed securities to adjust so that they contain the actual market prices. If we are not willing to speculate or invest in marketed securities, do we simply take account of their market prices when computing hedging costs, or do we infer something about **P** based on the belief that good deals should not exist? If the latter, inference about **P** might seek not just to maximize statistical likelihood, but to balance this objective with minimizing a distance to the set $\mathcal{Q}$ of EMMs.

Acknowledgements

The author is grateful for the support of the National Science Foundation under Grant No. DMS-0202958 and of the National Security Agency under Grant No. H98230-04-1-0047. He thanks Philippe Artzner, Aleš Černý, Dmitry Kramkov, and Costis Skiadas for valuable discussions and for suggesting references. He remains responsible for the views expressed and any errors.

Appendix A. Definition of incompleteness and fundamental theorems

We might like to define a complete market as one in which it is possible to replicate any cashflow. This raises several questions. What is the set C of cashflows that we hope to be able to replicate? What is the set Θ of possible portfolio strategies with which we hope to replicate them? What does it mean to replicate?

First, we must specify the set C of cashflows to be replicated. As usual, let us focus on cashflows that are simply random variables representing a payoff at a terminal time T. In assessing completeness of the market, it makes sense to consider replication only of payoffs that are functions of underlying financial variables observed over the time interval $[0, T]$. Even this set is too large for mathematical convenience, and the literature usually imposes a further restriction that the payoffs under consideration must be integrable or bounded. There are also economic reasons for a boundedness restriction.

The second question is of which portfolio strategies are allowed, and with limited credit, it is impossible to execute a portfolio strategy whose value is unbounded below. Along with this restriction of "admissibility" or "tameness," we also restrict attention to portfolio strategies that are self-financing (after any transaction costs), for the same reasons as in the study of no-arbitrage pricing and the first fundamental theorem of asset pricing. We must also consider only portfolio strategies whose initial cost is finite; this is a substantive restriction in models with an infinite number of marketed securities. We might also consider imposing other restrictions. For instance, we may consider only "stopping time simple" portfolio strategies, which (almost surely) include only a finite number of times at which the portfolio is rebalanced, because continuous-time hedging is impossible. We might also restrict the number of marketed securities in the portfolio to be finite, even if an infinite number are available in the model, for similar reasons. The result of defining the set Θ of possible portfolio strategies is a set of exactly replicable payoffs, $R := \{Y \mid \exists \theta \in \Theta \ni Y = \theta_T S_T\}$, where S is the stochastic process of marketed securities' prices and $\theta_T S_T$ is the terminal value of portfolio strategy θ.

Third, what does it mean to replicate? Exact replication led to the definition of R, and one candidate definition for completeness is $C = R$, all payoffs can be exactly replicated. Jarrow et al. (1999) refer to this property as *algebraic completeness*, saying, "This definition is too strong and would hardly ever be satisfied in practice." After a discussion of mathematics, we will argue that algebraic completeness is unnecessarily strong, and completeness should be defined differently.

The mathematical finance literature originally focused on algebraic completeness, but this created difficulties with the second fundamental theorem of asset pricing (FTAP), which relates market completeness to uniqueness of a pricing kernel. These difficulties were analogous to those that previously beset the first FTAP, which relates absence of arbitrage to existence of a pricing kernel. The difficulties with the first FTAP were solved by introducing a weaker notion than arbitrage, namely the *free lunch with vanishing risk* (Delbaen and Schachermayer, 1999; Protter, 2006). While an arbitrage is a portfolio strategy in Θ with nonpositive initial cost and terminal value that is nonnegative and nonzero, a free lunch with vanishing risk is a sequence of portfolio strategies in Θ with nonpositive initial cost and whose limiting terminal value is nonnegative and nonzero. Mathematically, the idea is to replace the no-arbitrage condition with a stronger one, which excludes even approximate arbitrages, such as a free lunch with vanishing risk. What constitutes an "approximate" arbitrage is determined by a topology on the space of payoffs i.e. terminal values: the concepts of closure and limit depend on this topology (Cherny, 2005; Staum, 2004). By analogy, for the second FTAP, it would make sense to replace algebraic completeness with a weaker, topological notion (Battig and Jarrow, 1999; Jarrow et al., 1999; Jarrow and Madan, 1999). The resulting second FTAPs connect uniqueness of a pricing kernel that is continuous with respect to some topology to approximate replicability of any payoff $Y \in C$ in

the sense that, for any neighborhood U of Y, there is a portfolio strategy in Θ whose terminal value is in U.

That is, the more successful mathematical notion of completeness relates the target payoffs C and the replicable payoffs R by means of a topology specifying what approximate replication means. This is an eminently practical notion, because we need only concern ourselves with whether a payoff can be approximately replicated. If we can find a hedging scheme that results in an arbitrarily small hedging error, we will be satisfied. An incomplete market, then, is one in which there are target payoffs that cannot even be approximately replicated, so that we must find a methodology for dealing with the resulting non-negligible residual risks after hedging.

Appendix B. Financial perspectives on incompleteness

B.1 Descriptive analysis: Are markets incomplete? How much so?

Whereas a financial engineer might directly test financial time series for features that are known to cause incompleteness (see Section 3.1), tests of market incompleteness in the financial literature often look for evidence of incompleteness in consumption data. As Saito (1999, §II) says, "When markets are complete, the intertemporal rate of substitution is equalized among agents." If so, then a calibrated representative agent model would reflect aggregate preferences, and microeconomic data would show that households are capable of fully insuring themselves against idiosyncratic risks. The approach focusing on calibration to aggregate data typically finds that calibrated parameters reflect implausible aggregate preferences. For instance, this is one guise of the equity premium puzzle. One response is to conclude that markets must not be complete after all. However, attempts to explain away the equity premium puzzle on the basis of incomplete markets have not been universally accepted (Mehra, 2003; Mehra and Prescott, 2003). An alternative conclusion is that the models being calibrated are themselves wrong, so that these tests do not correctly assess market completeness. However, Hansen and Jagannathan (1991) devised a test which is based on fewer assumptions and not specific to a particular model, and once again one is led to the conclusion that there is a puzzle if markets are complete. Another approach to testing is to use microeconomic data to show that household consumption has not been fully insured against idiosyncratic risks. A tentative conclusion is that markets are significantly incomplete, but some doubts may remain. See Saito (1999, §II) for further references.

B.2 Normative analysis: What should we do about incompleteness?

Completing the market increases welfare, but increasing the attainable span in an incomplete market without completing it may increase or decrease welfare. See Huang (2000, §III.A) for a qualitative summary and Duffie and Rahi

(1995, §2.2) for a mathematical synopsis of one result. Even something as apparently straightforward as an increase in welfare does not have clear normative implications. One way in which incomplete markets can lower welfare is by inducing agents to engage in precautionary saving as a substitute for unavailable insurance against risks. The resulting investment exceeds the level consistent with maximal welfare of agents existing today and thus produces economic growth which is in this sense excessive (Saito, 1999, §IV.B), but which may lead to greater welfare for future generations.

It is also unclear how great the welfare loss due to incompleteness is. Many factors influence the welfare loss generated within a model of an incomplete-market equilibrium: how many goods there are, whether the model describes only exchange or also production, what assets are marketed, whether there is aggregate risk or only idiosyncratic risk, and whether the time horizon and the persistence of shocks are infinite, short, or long relative to agents' patience, which has to do, for instance, with whether one can find a new job after being laid off, and with the length of business cycles. Levine and Zame (2002) ask "Does market incompleteness matter?" They answer that it does not in a model of an exchange economy with a single perishable good, agents who are patient, i.e. have a low discount rate in their intertemporal utility functions, and have an infinite time horizon, shocks that are not persistent, and only idiosyncratic risk; incompleteness matters if it prevents insuring against aggregate risks or the relative prices of multiple goods. To their question, Kübler and Schmedders (2001) respond unequivocally that "incomplete markets matter for welfares," even if agents are patient. Kim et al. (2003) study a simple international model of two countries and report that welfare loss is negligible when agents are patient and shocks are transitory, but is considerable and highly sensitive to the model's parameters in the more realistic case of patience and persistent shocks.

Equilibria in incomplete markets may even be Pareto inefficient given the constraints about contingent claims that cannot be traded, because agents make decisions based on the current equilibrium prices, whereas everyone's welfare might be increased at a different price system and allocation: see Hens (1998, §4), Huang (2000, §III.B), and Duffie and Rahi (1995, §3.3). This raises the possibility that suitably crafted regulatory intervention might increase welfare (for a simple example, see Huang, 2000, Appendix), but such a suggestion needs to be treated with the utmost caution, as the relevant central planning problem would require a tremendous amount of information: see Huang (2000, §§IV–VI) and Herings and Polemarchakis (2005).

There is a connection between Pareto inefficiency and the topic of sunspot equilibria in incomplete markets. On sunspot equilibrium see e.g. Hens (1998, §9). A sunspot equilibrium is one in which allocations of goods depend on extrinsic events, such as sunspots, having nothing to do with preferences, endowments, and production possibilities; agents may have self-fulfilling expectations associated with these extrinsic events, generating volatility in excess of what is warranted by fundamentals (Prescott and Shell, 2002). According

to Hens (1998, §9.2), "sunspots matter if and only if markets are incomplete." Sunspot equilibria generate excess uncertainty and are Pareto inefficient and dominated by nonsunspot equilibria in strictly convex economies (Prescott and Shell, 2002). Pareto efficiency of sunspot and non-sunspot equilibria remains a subject of active research, e.g. Pietra (2004). The existence of sunspot and non-sunspot equilibria has an interesting relation to options. Antinolfi and Keister (1998) report that the introduction into a market of a small number of options can render it "strongly sunspot immune," i.e. eliminate the possibility of sunspot equilibria no matter what extrinsic phenomenon constitutes the sunspots; this is in contrast to previous results they cite, asserting that options can have a destabilizing effect. For instance, according to Bowman and Faust (1997), it is possible for the addition of a market in options to introduce sunspot equilibria into an economy that previously did not have any, even if that economy's market was already complete! This has to do with the fact that options, as derivative securities, have payoffs related to underlying security prices and not directly to the state of the economy.

Public prices reveal private information and one may analyze how much private information a certain market structure reveals (Duffie and Rahi, 1995, §3.2); recent work on this topic includes Kübler et al. (2002). One is tempted to suppose that complete revelation is desirable because it increases market efficiency (in the sense of the efficient markets hypothesis, not Pareto efficiency) and thus promotes the allocation of resources to maximally productive uses. However, the normative issues surrounding information revelation are not simple. When private information is not revealed, uninformed investors may be hesitant to trade; this is the rationale behind the prohibition on insider trading, and the insight underlying an extensive economic literature spawned by the famous paper on lemons (Akerlof, 1970). Yet private information might be revealed in a way that resolves uncertainty about individuals' endowments so that they cannot well insure it. Hirshleifer (1971) describes how public information can disrupt a market's ability to provide insurance. For example, suppose that the only uncertainty about the price of corn at harvest comes from ignorance about the total number of acres planted. In this case farmers would not be able to insure themselves well against price risk by hedging in futures markets, because the very act of their attempting to hedge their crops would reveal the total crop size, thus resolving all uncertainty about the price. Marin and Rahi (2000) and Dow and Rahi (2003) study this tension; the magnitude of these opposing informational effects is unknown.

References

Akerlof, G.A. (1970). The market for lemons – Quality uncertainty and market mechanism. *Quarterly Journal of Economics* 84 (3), 488–500.

Andersen, T.G., Benzoni, L., Lund, J. (2002). An empirical investigation of continuous-time equity return models. *Journal of Finance* 57 (3), 1239–1284.

Anderson, E.W., Hansen, L.P., Sargent, T.J. (2003). A quartet of semigroups for model specification, robustness, prices of risk, and model detection. *Journal of the European Economic Association* 1 (1), 68–123.

Antinolfi, G., Keister, T. (1998). Options and sunspots in a simple monetary economy. *Economic Theory* 11, 295–315.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9, 203–228.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H. (2007). Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research* 152, 5–22.

Battig, R.J., Jarrow, R.A. (1999). The second fundamental theorem of asset pricing: A new approach. *Review of Financial Studies* 12 (5), 1219–1235.

Becherer, D. (2003). Rational hedging and valuation of integrated risks under constant absolute risk aversion. *Insurance Mathematics and Economics* 33, 1–28.

Bernardo, A.E., Ledoit, O. (2000). Gain, loss, and asset pricing. *Journal of Political Economy* 108 (1), 144–172.

Bertsimas, D., Kogan, L., Lo, A.W. (2001). Hedging derivative securities and incomplete markets: An ϵ-arbitrage approach. *Operations Research* 49 (3), 372–397.

Bewley, T.F. (2002). Knightian decision theory: Part I. *Decisions in Economics and Finance* 25, 79–110. First appeared in 1986 as Cowles Foundation Discussion Paper No. 807.

Biagini, F., Pratelli, M. (1999). Local risk minimization and numéraire. *Journal of Applied Probability* 36 (4), 1126–1139.

Bizid, A., Jouini, E. (2005). Equilibrium pricing in incomplete markets. *Journal of Financial and Quantitative Analysis* 40 (4).

Björk, T., Slinko, I. (2006). Towards a general theory of good-deal bounds. *Review of Finance* 10 (2), 221–260.

Bondarenko, O., Longarela, I.R. (2004). Benchmark good-deal bounds: An application to stochastic volatility models of option pricing. *Working paper*.

Bowman, D., Faust, J. (1997). Options, sunspots, and the creation of uncertainty. *Journal of Political Economy* 105 (5), 957–975.

Boyle, P., Wang, T. (2001). Pricing of new securities in an incomplete market: The Catch 22 of no-arbitrage pricing. *Mathematical Finance* 11 (3), 267–284.

Brandt, M.W. (2003). Hedging demands in hedging contingent claims. *Review of Economics and Statistics* 85 (1), 119–140.

Cao, H.H., Wang, T., Zhang, H.H. (2005). Model uncertainty, limited market participation and asset prices. *Review of Financial Studies* 18 (4), 1219–1251.

Carr, P. (2002). Frequently asked questions in option pricing theory. *Journal of Derivatives*, in press.

Carr, P., Geman, H., Madan, D.B. (2001). Pricing and hedging in incomplete markets. *Journal of Financial Economics* 62, 131–167.

Carr, P., Geman, H., Madan, D.B., Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75 (2), 305–332.

Černý, A. (2003). Generalised Sharpe ratios and asset pricing in incomplete markets. *European Finance Review* 7, 191–233.

Černý, A., Hodges, S. (2002). The theory of good-deal pricing in financial markets. In: Geman, H., Madan, D., Pliska, S., Vorst, T. (Eds.), *Mathematical Finance – Bachelier Congress 2000*. Springer-Verlag, Berlin, pp. 175–202.

Chazal, M., Jouini, E. (2004). Equilibrium pricing bounds on option prices. *Working paper*.

Chen, Z., Epstein, L. (2002). Ambiguity, risk, and asset returns in continuous time. *Econometrica* 70 (4), 1403–1443.

Cherny, A.S. (2005). General arbitrage pricing model: probability approach. *Working paper*.

Choulli, T., Stricker, C. (2005). Minimal entropy-Hellinger martingale measure in incomplete markets. *Mathematical Finance* 15 (3), 465–490.

Choulli, T., Stricker, C., Li, J. (2006). Minimal Hellinger martingale measures of order q. *Working paper*.

Clewlow, L., Hodges, S. (1997). Optimal delta-hedging under transactions costs. *Journal of Economic Dynamics and Control* 21, 1353–1376.

Cochrane, J.H., Saá-Requejo, J. (2000). Beyond arbitrage: Good-deal asset price bounds in incomplete markets. *Journal of Political Economy* 108 (1), 79–119.

Constantinides, G.M., Perrakis, S. (2002). Stochastic dominance bounds on derivatives prices in a multiperiod economy with proportional transaction costs. *Journal of Economic Dynamics and Control* 26, 1323–1352.

Cont, R., Tankov, P. (2004). *Financial Modelling with Jump Processes*. *Financial Mathematics Series*. Chapman & Hall/CRC, London.

Cvitanić, J. (2000). Minimizing expected loss of hedging in incomplete and constrained markets. *SIAM Journal on Control and Optimization* 38 (4), 1050–1066.

Davis, M.H.A. (2004a). Complete-market models of stochastic volatility. *Proceedings of the Royal Society of London (A)* 460, 11–26.

Davis, M.H.A. (2004b). Valuation, hedging and investment in incomplete financial markets. In: Hill, J.M., Moore, R. (Eds.), *Applied Mathematics Entering the 21st Century: Invited Talks from the ICIAM 2003 Congress*. In: *Proceedings in Applied Mathematics*, vol. 116. Society for Industrial and Applied Mathematics, Philadelphia. Chapter 4.

Delbaen, F., Schachermayer, W. (1999). Non-arbitrage and the fundamental theorem of asset pricing: Summary of main results. In: Heath, D.C., Swindle, G. (Eds.), *Introduction to Mathematical Finance*. In: *Proceedings of Symposia in Applied Mathematics*, vol. 57. American Mathematical Society, Providence, RI, pp. 49–58.

Delbaen, F., Grandits, P., Rheinländer, T., Samperi, D., Schweizer, M., Stricker, C. (2002). Exponential hedging and entropic penalties. *Mathematical Finance* 12 (2), 99–123.

Dentcheva, D., Ruszczyński, A. (2003). Optimization with stochastic dominance constraints. *SIAM Journal on Optimization* 14 (2), 548–566.

Dow, J., Rahi, R. (2003). Informed trading, investment, and welfare. *Journal of Business* 76 (3), 439–454.

Dritschel, M., Protter, P. (1999). Complete markets with discontinuous security price. *Finance and Stochastics* 3 (2), 203–214.

Duffie, D., Rahi, R. (1995). Financial market innovation and security design: An introduction. *Journal of Economic Theory* 65, 1–42.

Duffie, D., Gârleanu, N., Pedersen, L.H. (2006). Valuation in over-the-counter markets. *Working paper*.

Dunbar, N. (2005). JP Morgan reorganises US operations after trading losses. *RISK Magazine* 0, 12–13. January.

Dybvig, P.H. (1992). Hedging non-traded wealth: When is there separation of hedging and investment? In: Hodges, S. (Ed.), *Options: recent advances in theory and practice, vol. 2*. Manchester Univ. Press, NY, pp. 13–24. Chapter 2.

Eberlein, E., Jacod, J. (1997). On the range of options prices. *Finance and Stochastics* 1, 131–140.

El Karoui, N., Quenez, M.-C. (1995). Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal on Control and Optimization* 33 (1), 29–66.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–669.

Epstein, L.G., Miao, J. (2003). A two-person dynamic equilibrium under ambiguity. *Journal of Economic Dynamics and Control* 27, 1253–1288.

Epstein, L.G., Schneider, M. (2003). Recursive multiple-priors. *Journal of Economic Theory* 113, 1–31.

Epstein, L.G., Schneider, M. (2005). Ambiguity, information quality and asset pricing. *Working paper*.

Foldes, L. (2000). Valuation and martingale properties of shadow prices: An exposition. *Journal of Economic Dynamics and Control* 24, 1641–1701.

Föllmer, H., Leukert, P. (1999). Quantile hedging. *Finance and Stochastics* 3, 251–273.

Föllmer, H., Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics* 6 (4), 429–447.

Föllmer, H., Schweizer, M. (1991). Hedging of contingent claims under incomplete information. In: Davis, M.H.A., Elliott, R.J. (Eds.), *Applied Stochastic Analysis*. In: *Stochastics Monographs*, vol. 5. Gordon and Breach, New York, pp. 389–414.

Frittelli, M. (2000a). Introduction to a theory of value coherent with the no-arbitrage principle. *Finance and Stochastics* 4, 275–297.

Frittelli, M. (2000b). The minimal entropy martingale measure and the valuation problem in incomplete markets. *Mathematical Finance* 10, 39–52.
Fujiwara, T., Miyahara, Y. (2003). The minimal entropy martingale measures for geometric Lévy processes. *Finance and Stochastics* 7, 509–531.
Geman, H., Madan, D.B. (2004). Pricing in incomplete markets: From absence of good deals to acceptable risk. In: Szegö, G. (Ed.), *Risk Measures for the 21st Century*. Wiley, Hoboken, NJ, pp. 451–474. Chapter 21.
Gilboa, I., Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–153.
Goll, T., Rüschendorf, L. (2001). Minimax and minimal distance martingale measures and their relationship to portfolio optimization. *Finance and Stochastics* 5, 557–581.
Hansen, L.P., Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99 (2), 225–262.
Hansen, L.P., Sargent, T.J., Turmuhambetova, G.A., Williams, N. (2006). Robust control and model misspecification. *Journal of Economic Theory* 128 (1), 45–90.
Heath, D., Platen, E., Schweizer, M. (2001). A comparison of two quadratic approaches to hedging in incomplete markets. *Mathematical Finance* 11 (4), 385–413.
Hens, T. (1998). Incomplete markets. In: Kirman, A. (Ed.), *Elements of General Equilibrium Analysis*. Blackwell, Oxford, pp. 139–210. Chapter 5.
Herings, P.J.-J., Polemarchakis, H. (2005). Pareto improving price regulation when the asset market is incomplete. *Economic Theory* 25, 135–154.
Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6 (2), 327–343.
Hirshleifer, J. (1971). Private and social value of information and reward to inventive activity. *American Economic Review* 61 (4), 561–574.
Hodges, S.D., Neuberger, A. (1989). Optimal replication of contingent claims under transactions costs. *Review of Futures Markets* 8, 222–239.
Huang, P.H. (2000). A normative analysis of new financially engineered derivatives. *Southern California Law Review* 73 (471), 471–521.
Hugonnier, J., Kramkov, D. (2004). Optimal investment with random endowments in incomplete markets. *Annals of Applied Probability* 14 (2), 845–864.
Hugonnier, J., Kramkov, D., Schachermayer, W. (2005). On utility-based pricing of contingent claims in incomplete markets. *Mathematical Finance* 15 (2), 203–212.
Jarrow, R.A., Madan, D.B. (1999). Hedging contingent claims on semimartingales. *Finance and Stochastics* 3, 111–134.
Jarrow, R.A., Purnanandam, A. (2004). The valuation of a firm's investment opportunities: A reduced form credit risk perspective. *Working paper*.
Jarrow, R.A., Jin, X., Madan, D.B. (1999). The second fundamental theorem of asset pricing. *Mathematical Finance* 9 (3), 255–273.
Jaschke, S., Küchler, U. (2001). Coherent risk measures and good-deal bounds. *Finance and Stochastics* 5, 181–200.
Jouini, E. (2001). Arbitrage and control problems in finance: A presentation. *Journal of Mathematical Economics* 35, 167–183.
Kabanov, Y.M., Stricker, C. (2002). On the optimal portfolio for the exponential utility maximization: Remarks to the six-author paper. *Mathematical Finance* 12 (2), 125–134.
Kallsen, J. (2002a). Derivative pricing based on local utility maximization. *Finance and Stochastics* 6, 115–140.
Kallsen, J. (2002b). Utility-based derivative pricing in incomplete markets. In: Geman, H., Madan, D., Pliska, S., Vorst, T. (Eds.), *Mathematical Finance – Bachelier Congress 2000*. Springer-Verlag, Berlin, pp. 313–338.
Karatzas, I., Shreve, S.E. (1998). *Methods of Mathematical Finance*. *Applications of Mathematics*, vol. 39. Springer-Verlag, New York.
Karatzas, I., Žitković, G. (2003). Optimal consumption from investment and random endowment in incomplete semimartingale markets. *Annals of Probability* 31 (4), 1821–1858.

Kim, J., Kim, S.H., Levin, A. (2003). Patience, persistence, and welfare costs of incomplete markets in open economies. *Journal of International Economics* 61, 385–396.
Kou, S.G. (2002). A jump-diffusion model for option pricing. *Management Science* 48 (8), 1086–1101.
Kübler, F., Schmedders, K. (2001). Incomplete markets, transitory shocks, and welfare. *Review of Economic Dynamics* 4, 747–766.
Kübler, F., Chiappori, P.-A., Ekeland, I., Polemarchakis, H.M. (2002). The identification of preferences from equilibrium prices under uncertainty. *Journal of Economic Theory* 102, 403–420.
Larsen, K., Pirvu, T.A., Shreve, S.E., Tütüncü, R. (2005). Satisfying convex risk limits by trading. *Finance and Stochastics* 9 (2), 177–195.
Levine, D.K., Zame, W.R. (2002). Does market incompleteness matter? *Econometrica* 70 (5), 1805–1839.
Lim, A.E.B. (2004). Quadratic hedging and mean–variance portfolio selection with random parameters in an incomplete market. *Mathematical Methods of Operations Research* 29 (1), 132–161.
Liu, J., Pan, J., Wang, T. (2005). An equilibrium model of rare-event premia and its implication for option smirks. *Review of Financial Studies* 18 (1), 131–164.
Lo, A. (1987). Semiparametric upper bounds for option prices and expected payoffs. *Journal of Financial Economics* 19 (2), 373–388.
Madan, D.B., Carr, P.P., Chang, E.C. (1998). The variance gamma process and option pricing. *European Finance Review* 2, 79–105.
Maenhout, P.J. (2004). Robust portfolio rules and asset pricing. *Review of Financial Studies* 17 (4), 951–983.
Magill, M., Quinzii, M. (1996). *Theory of Incomplete Markets, vol. 1*. MIT Press, Cambridge, MA.
Mania, M., Santacroce, M., Tevzadze, R. (2003). A semimartingale BSDE related to the minimal entropy martingale measure. *Finance and Stochastics* 7, 385–402.
Marin, J.M., Rahi, R. (2000). Information revelation and market incompleteness. *Review of Economic Studies* 67, 455–481.
Markowitz, H.M. (2002). Efficient portfolios, sparse matrices, and entities: A retrospective. *Operations Research* 50 (1), 154–160.
Mehra, R. (2003). The equity premium: Why is it a puzzle? *Financial Analysts Journal* 59 (1), 54–69.
Mehra, R., Prescott, E.C. (2003). The equity premium puzzle in retrospect. In: Constantinides, G., Harris, M., Stulz, R. (Eds.), *Handbook of the Economics of Finance*. Elsevier, Amsterdam, pp. 887–936. Chapter 14.
Musiela, M., Zariphopoulou, T. (2004a). An example of indifference prices under exponential preferences. *Finance and Stochastics* 8, 229–239.
Musiela, M., Zariphopoulou, T. (2004b). A valuation algorithm for indifference prices in incomplete markets. *Finance and Stochastics* 8, 399–414.
Mykland, P.A. (2003a). Financial options and statistical prediction intervals. *Annals of Statistics* 31, 1413–1438.
Mykland, P.A. (2003b). The interpolation of options. *Finance and Stochastics* 7, 417–432.
Perrakis, S., Ryan, P.J. (1984). Option pricing bounds in discrete time. *Journal of Finance* 39 (2), 519–525.
Pham, H. (2000). On quadratic hedging in continuous time. *Mathematical Methods of Operations Research* 51, 315–339.
Pietra, T. (2004). Sunspots, indeterminacy and Pareto inefficiency in economies with incomplete markets. *Economic Theory* 24, 687–699.
Prescott, E.C., Shell, K. (2002). Introduction to sunspots and lotteries. *Journal of Economic Theory* 107, 1–10.
Protter, P. (2006). A partial introduction to financial asset pricing theory. In: Birge, J.R., Linetsky, V. (Eds.), *Financial Engineering*. In: *Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam.
Roorda, B., Schumacher, H. (2005). Time consistency conditions for acceptability measures, with an application to tail value at risk. *Working paper*.
Roorda, B., Schumacher, H., Engwerda, J. (2005). Coherent acceptability measures in multiperiod models. *Mathematical Finance* 15 (4), 589–612.

Rouge, R., El Karoui, N. (2000). Pricing via utility maximization and entropy. *Mathematical Finance* 10 (2), 259–276.
Routledge, B.R., Zin, S.E. (2004). Model uncertainty and liquidity. *Working paper*.
Ruszczyński, A., Shapiro, A. (2004). Optimization of convex risk functions. *Working paper*.
Saito, M. (1999). Dynamic allocation and pricing in incomplete markets: A survey. *Monetary and Economic Studies* 17 (1), 45–75.
Schachermayer, W. (2002). Optimal investment in incomplete financial markets. In: Geman, H., Madan, D., Pliska, S., Vorst, T. (Eds.), *Mathematical Finance – Bachelier Congress 2000*. Springer-Verlag, Berlin, pp. 427–462.
Schachermayer, W. (2003). A super-martingale property of the optimal portfolio process. *Finance and Stochastics* 7, 433–456.
Schied, A. (2004). On the Neyman–Pearson problem for law-invariant risk measures and robust utility functionals. *Annals of Applied Probability* 14 (3), 1398–1423.
Schweizer, M. (1995). Variance-optimal hedging in discrete time. *Mathematics of Operations Research* 20 (1), 1–32.
Schweizer, M. (1996). Approximation pricing and the variance-optimal martingale measure. *Annals of Probability* 24, 206–236.
Schweizer, M. (1999). A minimality property of the minimal martingale measure. *Statistics and Probability Letters* 42, 27–31.
Schweizer, M. (2001). A guided tour through quadratic hedging approaches. In: Jouini, E., Cvitanić, J., Musiela, M. (Eds.), *Option Pricing, Interest Rates and Risk Management*. In: *Handbooks in Mathematical Finance*. Cambridge University Press, Cambridge, pp. 538–574. Chapter 15.
Skiadas, C. (2003). Robust control and recursive utility. *Finance and Stochastics* 7, 475–489.
Skiadas, C. (2006). Dynamic portfolio theory. In: Birge, J.R., Linetsky, V. (Eds.), *Financial Engineering*. In: *Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam.
Staum, J. (2004). Fundamental theorems of asset pricing for good deal bounds. *Mathematical Finance* 14 (2), 141–161.
Talay, D., Zheng, Z. (2002). Worst case model risk management. *Finance and Stochastics* 6, 517–537.
Uppal, R., Wang, T. (2003). Model misspecification and underdiversification. *Journal of Finance* 58 (6), 2465–2486.
Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, New York.
Zhang, L., Mykland, P.A., Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100 (472), 1394–1411.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15013-1

Chapter 13

Option Pricing: Real and Risk-Neutral Distributions

George M. Constantinides
University of Chicago and NBER
E-mail: gmc@ChicagoGSB.edu

Jens Carsten Jackwerth
University of Konstanz
E-mail: Jens.Jackwerth@uni-konstanz.de

Stylianos Perrakis
Concordia University
E-mail: SPerrakis@jmsb.concordia.ca

Abstract

The central premise of the Black and Scholes [Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–659] and Merton [Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–184] option pricing theory is that there exists a self-financing *dynamic trading policy* of the stock and risk free accounts that renders the market *dynamically complete*. This requires that the market be *complete* and *perfect*. In this essay, we are concerned with cases in which dynamic trading breaks down either because the market is incomplete or because it is imperfect due to the presence of trading costs, or both. Market incompleteness renders the risk-neutral probability measure non unique and allows us to determine the option price only within a range. Recognition of trading costs requires a refinement in the definition and usage of the concept of a risk-neutral probability measure. Under these market conditions, a replicating dynamic trading policy does not exist. Nevertheless, we are able to impose restrictions on the pricing kernel and derive testable restrictions on the prices of options. We illustrate the theory in a series of market setups, beginning with the single period model, the two-period model and, finally, the general multiperiod model, with or without transaction costs. We also review related empirical results that document widespread violations of these restrictions.

Keywords: Derivative pricing; Risk-neutral distribution; Incomplete markets; Stochastic dominance bounds; Transaction costs; Index options; Volatility smile

1 Introduction

The Nobel-winning ingenious idea behind the classic option pricing model of Black and Scholes (1973) and Merton (1973), hereafter BSM, is that, in the absence of arbitrage, the price of an option equals the cost of setting up a judiciously managed portfolio with payoff that replicates the option payoff.

The central premise of the BSM theory is that there exists a self-financing *dynamic trading policy* of the stock and risk free accounts that renders the market *dynamically complete*. This requires that the market be *complete* and *perfect*. Two assumptions of the BSM model make the market complete. First, the price of the underlying security has continuous sample paths at the exclusion of jumps. Second, the stock return volatility is constant. These assumptions essentially imply that the price of the underlying security is a geometric Brownian motion. Finally, the assumption of the BSM model that renders the market perfect is that trading is frictionless. In the BSM model, the volume of trading over any finite time interval is infinite. The transaction costs associated with the replicating dynamic trading policy would be infinite for any given positive proportional transactions cost rate.

Formally, absence of arbitrage in a frictionless market implies the existence of a *risk-neutral probability measure*, not necessarily unique, such that the price of any asset equals the expectation of its payoff under the risk-neutral measure, discounted at the risk free rate. Furthermore, if the market is complete then the risk-neutral measure is unique and the option price is unique as well. In the BSM model, the price of the underlying security follows a geometric Brownian motion which renders the market complete and the option price unique as well.

The risk-neutral probability measure is the real probability measure with the expected rate of return on the underlying security replaced by the risk free rate. The real probability distribution of stock returns can be estimated from the time series of past returns. The risk-neutral probability distribution of stock returns can be estimated from the cross section of option prices. As discussed in detail in the empirical Section 10, this prediction of the BSM theory does not fare well and provides the motivation to reexamine the premises of the theory.

In this essay, we are concerned with cases in which dynamic trading breaks down either because the market is incomplete or because there are trading costs or both. Market incompleteness renders the risk-neutral probability measure non unique and allows us to determine the option price only within a range. Recognition of trading costs requires a refinement in the definition and usage of the concept of a risk-neutral probability measure.

In Section 2, we discuss the implications of the absence of arbitrage. We introduce the concept of the risk-neutral probability and the closely related concept of the *state price density* or *pricing kernel*. We apply the theory to price

options under the assumption of the absence of arbitrage in complete and incomplete markets. In Section 3, we lay out the general framework for pricing options in a market that is incomplete and also imperfect due to trading costs. Under these market conditions, a replicating dynamic trading policy does not exist. Nevertheless, we are able to impose further restrictions on the pricing kernel and provide testable restrictions on the prices of options. In Sections 4–9, we illustrate the theory in a series of market setups, beginning with the single period model, the two-period model and finally the general multiperiod model, with or without transaction costs. In Section 10, we review related empirical results and, in Section 11, conclude.

2 Implications of the absence of arbitrage

2.1 General theory

Absence of arbitrage in a frictionless market implies the existence of a risk-neutral probability measure, not necessarily unique, such that the price of any asset equals the expectation of its payoff under the risk-neutral measure, discounted at the risk free rate. If a risk-neutral measure exists, the ratio of the risk-neutral probability density and the real probability density, discounted at the risk free rate, is referred to as the *pricing kernel* or *stochastic discount factor* (SDF). Thus, absence of arbitrage implies the existence of a strictly positive SDF. These ideas are implicit in the option pricing theory of Black and Scholes (1973) and Merton (1973) and are further developed by Ross (1976), Cox and Ross (1976), Constantinides (1978), Harrison and Kreps (1979), Harrison and Pliska (1981), and Delbaen and Schachermayer (1994).

To fix ideas, let there be J securities. Security j, $j = 1, \ldots, J$, has price P_j at the beginning of the period and payoff X_{ij} in state i, $i = 1, \ldots, I$, at the end of the period. An investor purchases θ_j securities of type j, $j = 1, \ldots, J$, with the objective to minimize the purchase cost, subject to the constraint that the portfolio payoff is strictly positive in all states of nature. The investor solves the following LP problem:

$$\inf_{\{\theta_j\}} \sum_{j=1}^{J} \theta_j P_j \tag{2.1}$$

subject to

$$\sum_{j=1}^{J} \theta_j X_{ij} > 0, \quad \forall i. \tag{2.2}$$

If the minimum purchase cost is negative, then there is an arbitrage opportunity.

Absence of arbitrage implies that the above problem, with the added condition

$$\sum_{j=1}^{J} \theta_j P_j < 0 \tag{2.3}$$

is infeasible. Then the *dual* of this LP problem is feasible. This implies the existence of strictly positive *state prices*, $\{\pi_i\}_{i=1,\ldots,I}$, such that:

$$P_j = \sum_{i=1}^{I} \pi_i X_{ij}, \quad \forall j \tag{2.4}$$

and

$$\pi_i > 0, \quad \forall i. \tag{2.5}$$

If the number of states does not exceed the number of securities with linearly independent payoffs, the market is said to be *complete* and the state prices are unique. Otherwise, the market is *incomplete* and the state prices are not unique.

The normalized state prices $q_i \equiv \pi_i / \sum_{k=1}^{I} \pi_k$ can be thought of as probabilities because they are strictly positive and add up to one. The inverse of the sum of the state prices, $R \equiv 1/\sum_{k=1}^{I} \pi_k$, has the interpretation as one plus the risk free rate. Then we may write Eq. (2.4) as

$$P_j = R^{-1} \sum_{i=1}^{I} q_i X_{ij} = R^{-1} E^Q[X_j], \quad \forall j \tag{2.6}$$

with the interpretation that the price of security j is its expected payoff under the probability measure $Q = \{q_i\}$, discounted at the risk free rate. For this reason, the probability measure Q is referred to as a *risk-neutral* or *risk-adjusted* probability measure. Thus, absence of arbitrage implies the existence of a risk-neutral probability measure. This property of the absence of arbitrage is far more general than this simple illustration implies.

Let $P = \{p_i\}$ denote the real probability measure of the states. The ratio $m_i \equiv \pi_i / p_i$ is referred to as the *state price density* or *stochastic discount factor* or *pricing kernel* or *intertemporal marginal rate of substitution*. In terms of the pricing kernel, we may write Eq. (2.4) as

$$P_j = \sum_{i=1}^{I} p_i m_i X_{ij} = E^P[m_i X_j], \quad \forall j \tag{2.7}$$

where the expectation is with respect to the real probability measure P.

2.2 Application to the pricing of options

Let the stock market index have price S_0 at the beginning of the period; *ex dividend* price S_i with probability p_i in state i, $i = 1, \ldots, I$, at the end of the period; and *cum dividend* price $(1 + \delta)S_i$ at the end of the period. The jth derivative, $j = 1, \ldots, J$, has price P_j at the beginning period, and its cash payoff X_{ij} is $G_j(S_i)$, a given function of the terminal stock price, at the end of the period in state i. In this context, absence of arbitrage implies the existence of a strictly positive pricing kernel $m : m_i$, $i = 1, \ldots, I$, such that:

$$1 = R \sum_{i=1}^{I} p_i m_i, \tag{2.8}$$

$$S_0 = \sum_{i=1}^{I} p_i m_i (1 + \delta) S_i \tag{2.9}$$

and

$$P_j = \sum_{i=1}^{I} p_i m_i G_j(S_i), \quad j = 1, \ldots, J. \tag{2.10}$$

Non-existence of a strictly positive pricing kernel implies arbitrage such as violations of the Merton (1973) no-arbitrage restrictions on the prices of options.

In practice, it is always possible to estimate the real probability measure P from time series data on past index returns. A *derivatives pricing model* is then a theory that associates the appropriate pricing kernel $m : m_i > 0$, $i = 1, \ldots, I$, with the estimated probability measure P.

In the absence of arbitrage, a *unique* pricing kernel may be derived in terms of the prices of J securities with linearly independent payoffs, if the market is complete, $J \geqslant I$. Then any derivative is uniquely priced in terms of the prices of I securities. This is the essence of derivatives pricing when the market is complete. An example of a complete market is the binomial model, described next.

In a *single-period binomial* model, there are just two states and the pricing kernel is derived in terms of the prices of the risk free asset and the stock or index on which options are written. Then any derivative is uniquely priced in terms of the risk free rate and the stock or index price. The natural extension of the single period binomial model is the widely used *multiperiod binomial* model developed by Cox and Ross (1976), Cox et al. (1979), and Rendleman and Bartter (1979). The stock price evolves on a multi-stage binomial tree over the life of the option so that the stock price assumes a wide range of values. Yet the market is complete because in each subperiod there are only two states. An option can be hedged or replicated on the binomial tree by adjusting the amounts held in the stock and the risk free asset at each stage of the binomial process. This type of trading is called *dynamic trading* and renders the market

dynamically complete. These fundamental ideas underlie the original option pricing model of Black and Scholes (1973) and Merton (1973). The binomial model is often used as a pedagogical tool to illustrate these ideas as in the textbook treatments by Hull (2006) and McDonald (2005). The binomial model is also a powerful tool in its own right in numerically pricing American and exotic options.

In this essay, we are concerned with cases in which dynamic trading or hedging breaks down either because the market is incomplete or because there are trading costs or both. In these cases, we impose further restrictions on the pricing kernel by taking into account the economic environment in which the derivatives are traded.

3 Additional restrictions implied by utility maximization

3.1 Multiperiod investment behavior with proportional transaction costs

We consider a market with heterogeneous agents and investigate the restrictions on option prices imposed by a particular class of utility-maximizing traders that we simply refer to as *traders*. We do not make the restrictive assumption that all agents belong to the class of the utility-maximizing traders. Thus our results are unaffected by the presence of agents with beliefs, endowments, preferences, trading restrictions, and transaction cost schedules that differ from those of the utility-maximizing traders.

As in Constantinides (1979), trading occurs at a finite number of trading dates, $t = 0, 1, \ldots, T, \ldots, T'$.[1] The utility-maximizing traders are allowed to hold only two primary securities in the market, a bond and a stock. The stock has the natural interpretation as the market index. Derivatives are introduced in the next section. The bond is risk free and pays constant interest $R - 1$ each period. The traders may buy and sell the bond without incurring transaction costs. At date t, the *cum dividend* stock price is $(1 + \delta_t)S_t$, the cash dividend is $\delta_t S_t$, and the *ex dividend* stock price is S_t, where δ_t is the dividend yield. We assume that the rate of return on the stock, $(1+\delta_{t+1})S_{t+1}/S_t$, is identically and independently distributed over time.

The assumption of i.i.d. returns is not innocuous and, in particular, rules out state variables such as stochastic volatility, stochastic risk aversion, and stochastic conditional mean of the growth rate in dividends and consumption. In this essay, we deliberately rule out such state variables in order to explore the extent to which market incompleteness and market imperfections (trading costs) alone explain the prices of index options. We discuss models with such state variables in Section 10.

[1] The calendar length of the trading horizon is N years and the calendar length between trading dates is N/T' years. Later on we vary T' and consider the mispricing of options under different assumptions regarding the calendar length between trading dates.

Stock trades incur proportional transaction costs charged to the bond account as follows. At each date t, the trader pays $(1 + k)S_t$ out of the bond account to purchase one *ex dividend* share of stock and is credited $(1 - k)S_t$ in the bond account to sell (or, sell short) one *ex dividend* share of stock. We assume that the transactions cost rate satisfies the restriction $0 \leqslant k < 1$. Note that there is no presumption that all agents in the economy face the same schedule of transaction costs as the traders do.

A trader enters the market at date t with dollar holdings x_t in the bond account and y_t/S_t *ex dividend* shares of stock. The endowments are stated net of any dividend payable on the stock at time t.[2] The trader increases (or, decreases) the dollar holdings in the stock account from y_t to $y'_t = y_t + \upsilon_t$ by decreasing (or, increasing) the bond account from x_t to $x'_t = x_t - \upsilon_t - k|\upsilon_t|$. The decision variable υ_t is constrained to be measurable with respect to the information at date t. The bond account dynamics are

$$x_{t+1} = \{x_t - \upsilon_t - k|\upsilon_t|\}R + (y_t + \upsilon_t)\frac{\delta_t S_{t+1}}{S_t}, \quad t \leqslant T' - 1 \tag{3.1}$$

and the stock account dynamics are

$$y_{t+1} = (y_t + \upsilon_t)\frac{S_{t+1}}{S_t}, \quad t \leqslant T' - 1. \tag{3.2}$$

At the terminal date, the stock account is liquidated, $\upsilon_{T'} = -y_{T'}$, and the net worth is $x_{T'} + y_{T'} - k|y_{T'}|$. At each date t, the trader chooses investment υ_t to maximize the expected utility of net worth, $E[u(x_{T'} + y_{T'} - k|y_{T'}|)|S_t]$.[3] We make the plausible assumption that the utility function, $u(\cdot)$, is increasing and concave, and is defined for both positive and negative terminal net worth.[4] Note that even this weak assumption of monotonicity and concavity of preferences is not imposed on all agents in the economy but only on the subset of agents that we refer to as traders.

We recursively define the value function $V(t) \equiv V(x_t, y_t, t)$ as

$$V(x_t, y_t, t) = \max_{\upsilon} E\Big[V\Big(\{x_t - \upsilon - k|\upsilon|\}R$$

[2] We elaborate on the precise sequence of events. The trader enters the market at date t with dollar holdings $x_t - \delta_t y_t$ in the bond account and y_t/S_t *cum dividend* shares of stock. Then the stock pays cash dividend $\delta_t y_t$ and the dollar holdings in the bond account become x_t. Thus, the trader has dollar holdings x_t in the bond account and y_t/S_t *ex dividend* shares of stock.

[3] The results extend routinely to the case that consumption occurs at each trading date and utility is defined over consumption at each of the trading dates and over the net worth at the terminal date. See Constantinides (1979) for details. The model with utility defined over terminal net worth alone is a more realistic representation of the objective function of financial institutions.

[4] If utility is defined only for non-negative net worth, then the decision variable is constrained to be a member of a convex set that ensures the non-negativity of net worth. See Constantinides (1979) for details. However, the derivation of bounds on the prices of derivatives requires an entirely different approach and yields weaker bounds. This problem is studied in Constantinides and Zariphopoulou (1999, 2001).

$$+ (y_t + \upsilon)\frac{\delta_t S_{t+1}}{S_t}, (y_t + \upsilon)\frac{S_{t+1}}{S_t}, t+1\Big) \Big| S_t\Big] \tag{3.3}$$

for $t \leqslant T' - 1$, and

$$V(x_{T'}, y_{T'}, T') = u\big(x_{T'} + y_{T'} - k|y_{T'}|\big). \tag{3.4}$$

We assume that the parameters satisfy appropriate technical conditions such that the value function exists and is once differentiable.

Equations (3.1)–(3.4) define a dynamic program that can be numerically solved for given utility function and stock return distribution. We shall not solve this dynamic program because our goal is to derive restrictions on the prices of options that are independent of the specific functional form of the utility function but solely depend on the plausible assumption that the traders' utility function is *monotone increasing* and *concave* in the terminal wealth.

The value function is increasing and concave in (x_t, y_t), properties that it inherits from the assumed monotonicity and concavity of the utility function, as shown in Constantinides (1979):

$$V_x(t) > 0, \qquad V_y(t) > 0, \quad t = 0, \ldots, T, \ldots, T' \tag{3.5}$$

and

$$\begin{aligned} &V\big(\alpha x_t + (1-\alpha)x'_t, \alpha y_t + (1-\alpha)y'_t, t\big) \\ &\quad \geqslant \alpha V(x_t, y_t, t) + (1-\alpha)V(x'_t, y'_t, t), \\ &\quad 0 < \alpha < 1, \ t = 0, \ldots, T, \ldots, T'. \end{aligned} \tag{3.6}$$

On each date, the trader may transfer funds between the bond and stock accounts and incur transaction costs. Therefore, the marginal rate of substitution between the bond and stock accounts differs from unity by, at most, the transaction costs rate:

$$(1-k)V_x(t) \leqslant V_y(t) \leqslant (1+k)V_x(t), \quad t = 0, \ldots, T, \ldots, T'. \tag{3.7}$$

Marginal analysis on the bond holdings leads to the following condition on the marginal rate of substitution between the bond holdings at dates t and $t+1$:

$$V_x(t) = RE_t\big[V_x(t+1)\big], \quad t = 0, \ldots, T, \ldots, T' - 1. \tag{3.8}$$

Finally, marginal analysis on the stock holdings leads to the following condition on the marginal rate of substitution between the stock holdings at date t and the bond and stock holdings at date $t+1$:

$$\begin{aligned} V_y(t) &= E_t\left[\frac{S_{t+1}}{S_t}V_y(t+1) + \frac{\delta_t S_{t+1}}{S_t}V_x(t+1)\right], \\ &\quad t = 0, \ldots, T, \ldots, T' - 1. \end{aligned} \tag{3.9}$$

Below we employ these restrictions on the value function to derive restrictions on the prices of options.

3.2 Application to the pricing of options

We consider J European-style derivatives on the index, with random cash payoff $G_j(S_T)$, $j = 1, 2, \ldots, J$, at their common expiration date T, $T \leqslant T'$. At time zero, the trader can buy the jth derivative at price $P_j + k_j$ and sell it at price $P_j - k_j$, net of transaction costs. Thus $2k_j$ is the bid–ask spread plus the round-trip transaction costs that the trader incurs in trading the jth derivative. Note that there is no presumption that all agents in the economy face the same bid–ask spreads and transaction costs as the traders do.

We assume that the traders are marginal in all J derivatives. Furthermore, we assume that, if a trader holds a finite (positive or negative) number of derivatives, these positions are sufficiently small relative to her holdings in the bond and stock that the monotonicity and concavity conditions (3.5) and (3.6) on the value function remain valid.[5]

Marginal analysis leads to the following restrictions on the prices of options:

$$(P_j - k_j)V_x(0) \leqslant E_0\big[G_j(S_T)V_x(T)\big] \leqslant (P_j + k_j)V_x(0), \quad j = 1, 2, \ldots, J. \tag{3.10}$$

Similar restrictions apply to the prices of options at dates $t = 1, \ldots, T - 1$.

Below, we illustrate the implementation of the restrictions on the prices of options in a number of important special cases. First, we consider the case $T = 1$ which rules out trading between the bond and stock accounts over the lifetime of the options. We refer to this case as the *single-period case*. Note that the single-period case does not rule out trading over the trader's horizon after the options expire; it just rules out trading over the lifetime of the options. We discuss the single-period case both with and without transaction costs.

A useful way to identify the options that cause infeasibility or near-infeasibility of the problem is to single out a "test" option, say the Jth option, and solve the problem

$$\min_{\{V_x(t),V_y(t)\}_{t=0,\ldots,T}} E_0\left[G_J(S_T)\frac{V_x(T)}{V_x(0)}\right], \tag{3.11}$$

subject to conditions (3.5)–(3.10), where in Eq. (3.10) the subscript j runs from 1 to $J - 1$. If this problem is feasible, then the attained minimum has the following interpretation. If one can buy the test option for less than the minimum attained in this problem, then at least one investor, *but not necessarily all investors*, increases her expected utility by trading the test option.

Likewise, we may solve the problem

$$\max_{\{V_x(t),V_y(t)\}_{t=0,\ldots,T}} E_0\left[G_J(S_T)\frac{V_x(T)}{V_x(0)}\right], \tag{3.12}$$

[5] Conditions (3.7)–(3.9) remain valid even if the holdings of the derivatives are not small.

subject to conditions (3.5)–(3.10), where in Eq. (3.10) the subscript j runs from 1 to $J-1$. If this problem is feasible, then the attained maximum has the following interpretation. If one can write the test option for more than the maximum attained in this problem, then at least one investor, *but not necessarily all investors*, increases her expected utility by trading the test option.

As the number of trading dates T increases, the computational burden rapidly increases. One way to reduce computational complexity is to limit attention to the case $J = 1$ (one option) and convex payoff (as, for example, the payoff of a call or put option). In this special case, we present closed-form solutions with and without transaction costs and, in many cases, present limiting forms of the option prices, as the number of intermediate trading dates becomes infinitely large.

4 Special case: one period without transaction costs

4.1 Results for general payoffs

The stock market index has price S_0 at the beginning of the period; *ex dividend* price S_i with probability p_i in state i, $i = 1, \ldots, I$, at the end of the period; *cum dividend* price $(1+\delta)S_i$ at the end of the period; and return $(1+\delta)S_i/S_0$. We define by $z_i \equiv S_i/S_0$ the *ex dividend* price ratio. We order the states such that S_i is increasing in i. The jth derivative, $j = 1, \ldots, J$, has price P_j at the beginning period and cash payoff $G_j(z_i)$ at the end of the period in state i. We denote by $V^i(t)$ the value function at date t and state i.

Since the transaction costs rate is assumed to be zero, we have $V_x(0) = V_y(0)$ and $V_x^i(1) = V_y^i(1)$. We identify the previously defined stochastic discount factor or pricing kernel m_i with the intertemporal marginal rate of substitution in state i, $m_i \equiv V_x^i(1)/V_x(0)$. Conditions (3.8)–(3.10) become:

$$1 = R\sum_{i=1}^{I} p_i m_i, \tag{4.1}$$

$$1 = \sum_{i=1}^{I} p_i m_i (1+\delta) z_i \tag{4.2}$$

and

$$P_j = \sum_{i=1}^{I} p_i m_i G_j(z_i), \quad j = 1, \ldots, J. \tag{4.3}$$

The concavity relation (3.6) of the value function implies additional restrictions on the pricing kernel. Historically, the expected premium of the return on the stock over the bond is positive. Under the assumption of positive expected

premium, the trader is long in the stock. Since the assumption in the single-period model is that there is no trading between the bond and stock accounts over the life of the option, the trader's wealth at the end of the period is increasing in the stock return. Note that this conclusion critically depends on the assumption that there is no intermediate trading in the bond and stock. Since we employed the convention that the stock return is increasing in the state i, the trader's wealth on date T is increasing in the state i. Then the concavity of the value function implies that the marginal rate of substitution is decreasing in the state i:

$$m_1 \geqslant m_2 \geqslant \cdots \geqslant m_I > 0. \tag{4.4}$$

A pricing kernel satisfying restrictions (4.1)–(4.4) defines the intertemporal marginal rate of substitution of a trader who maximizes her increasing and concave utility and is marginal in the options, the index and the risk free rate. If there does not exist a pricing kernel satisfying restrictions (4.1)–(4.4), then any trader with increasing and concave utility can increase her expected utility by trading in the options, the index, and the risk free rate – hence equilibrium does not exist. These strategies are termed *stochastically dominant* for the purposes of this essay, insofar as they would be adopted by all traders with utility possessing the required properties, in the same way that all risk averse investors would choose a dominant portfolio over a dominated one in conventional second degree stochastic dominance comparisons. Thus, the existence of a pricing kernel that satisfies restrictions (4.1)–(4.4) is said to rule out *stochastic dominance* between the observed prices.

We emphasize that the restriction on option prices imposed by the criterion of the absence of stochastic dominance is motivated by the economically plausible assumption that there exists at least *one* agent in the economy with the properties that we assign to a trader. This is a substantially weaker assumption than requiring that *all* agents have the properties that we assign to traders. Stochastic dominance then implies that at least one agent, *but not necessarily all agents*, increases her expected utility by trading.[6]

As before, we single out a "test" option, say the Jth option, and derive bounds that signify infeasibility if the price of the test option lies outside the bounds. The general form of this problem was stated in expressions (3.11) and (3.12). In the special case of no trading over the life of the option and zero transactions costs, the bounds on the test option with payoff $G_J(z_i)$ in state i are given by

$$\max_{\{m_i\}}\Big(\text{or}, \min_{\{m_i\}}\Big) \sum_{i=1}^{I} p_i m_i G_J(z_i), \tag{4.5}$$

[6] We also emphasize that the restriction of the absence of stochastic dominance is weaker than the restriction that the capital asset pricing model (CAPM) holds. The CAPM requires that the pricing kernel be linearly decreasing in the index price. The absence of stochastic dominance merely imposes that the pricing kernel be monotone decreasing in the index price.

subject to conditions (4.1)–(4.4), where in Eq. (4.3) the subscript j runs from 1 to $J-1$.

4.2 Results for convex payoffs

The feasibility of relations (4.1)–(4.4) can be expressed in closed form in the special case where the options are puts and calls, with payoff $G_j(z_i)$ that is a convex function of the end-of-period return (or stock price). Ryan (2000, 2003) provided inequalities that define an admissible range of prices for each option by considering the prices of the two options with immediately adjacent strike prices and Huang (2005) tightened these inequalities. In practice, this means that (4.1)–(4.4) become infeasible in most realistic problems with a large enough set of traded options.

Perrakis and Ryan (1984), Levy (1985), and Ritchken (1985) expressed the upper and lower bounds in (4.5) in closed form in the special case $J=1$ (one option) where the option is a put or call, with payoff $G_1(z_i)$ that is a convex function of the end-of-period stock price. Consider a European call option with strike price K, payoff $G_1(z_i)=[S_0 z_i(1+\delta)-K]^+ \equiv c_i$ and price $P_1=c$. Define $\hat{z} \equiv \sum_{i=1}^{I} p_i z_i$ and assume $(1+\delta)\hat{z} \geqslant R$. Equations (4.1)–(4.5) become

$$\max_{\{m_i\}}(\text{or, min}) \sum_{i=1}^{I} p_i m_i c_i \tag{4.6}$$

subject to

$$\sum_{i=1}^{I} p_i m_i (1+\delta) z_i = 1, \quad \text{and}$$

$$R \sum_{i=1}^{I} p_i m_i = 1, \quad m_1 \geqslant \cdots \geqslant m_I > 0. \tag{4.7}$$

The solution to (4.6)–(4.7) crucially depends on the minimum value $z_{\min} \equiv z_1$. If $z_{\min} > 0$, the upper and lower bounds $\bar{c}_0$ and $\underline{c}_0$ on the call option price are given by

$$\begin{aligned}
\bar{c}_0 &= \frac{1}{R}\left[\frac{R-(1+\delta)z_{\min}}{(1+\delta)(\hat{z}-z_{\min})}\hat{c}_I + \frac{(1+\delta)\hat{z}-R}{(1+\delta)(\hat{z}-z_{\min})}c_1\right], \\
\underline{c}_0 &= \frac{1}{R}\left[\frac{R-(1+\delta)\hat{z}_h}{(1+\delta)(\hat{z}_{h+1}-\hat{z}_h)}\hat{c}_{h+1} + \frac{(1+\delta)\hat{z}_{h+1}-R}{(1+\delta)(\hat{z}_{h+1}-\hat{z}_h)}\hat{c}_h\right].
\end{aligned} \tag{4.8}$$

In the above equations, h is a state index such that $(1+\delta)\hat{z}_h \leqslant R \leqslant (1+\delta)\hat{z}_{h+1}$ and we have used the following notation for conditional expectations for $k = 1, \ldots, I$:

$$\hat{c}_k = \frac{\sum_{i=1}^{k} c_i p_i}{\sum_{i=1}^{k} p_i} = E\big[c_T \mid S_T \leqslant S_0(1+\delta)z_k\big],$$

$$\hat{z}_k = \frac{\sum_{i=1}^{k} z_i p_i}{\sum_{i=1}^{k} p_i} = E\big[z_T \mid z_T \leqslant z_k\big]. \tag{4.9}$$

Inspection of Eqs. (4.8) and (4.9) reveals that both the upper and lower bounds of the call option are discounted expectations with two different distributions, $U = \{u_i\}$ and $L = \{l_i\}$. These distributions are both *risk neutral*, since it can be easily verified that $R^{-1}\sum_{i=1}^{i=I} u_i(1+\delta)z_i = R^{-1}\sum_{i=1}^{i=I} l_i(1+\delta)z_i = 1$. These distributions are:

$$\begin{aligned}
u_1 &= \frac{R-(1+\delta)z_{\min}}{(1+\delta)(\hat{z}-z_{\min})}p_1 + \frac{(1+\delta)\hat{z}-R}{(1+\delta)(\hat{z}-z_{\min})},\\
u_i &= \frac{R-(1+\delta)z_{\min}}{(1+\delta)(\hat{z}-z_{\min})}p_i, \quad i = 2,\ldots,I,\\
l_i &= \frac{(1+\delta)\hat{z}_{h+1}-R}{(1+\delta)(\hat{z}_{h+1}-\hat{z}_h)}\frac{p_i}{\sum_{k=1}^{h} p_k} + \frac{R-(1+\delta)\hat{z}_h}{(1+\delta)(\hat{z}_{h+1}-\hat{z}_h)}\frac{p_i}{\sum_{k=1}^{h+1} p_k},\\
&\quad i = 1,\ldots,h,\\
l_{h+1} &= \frac{R-(1+\delta)\hat{z}_h}{(1+\delta)(\hat{z}_{h+1}-\hat{z}_h)}\frac{p_{h+1}}{\sum_{k=1}^{h+1} p_k}.
\end{aligned} \tag{4.10}$$

As the states increase, the distribution of z becomes continuous over the interval $[z_{\min}, \infty)$, with actual distribution $P(z)$ and expectation $E(z)$. Then, U and L become

$$\begin{aligned}
U(z) &= \begin{cases} P(z) & \text{with probability } \frac{R-(1+\delta)z_{\min}}{(1+\delta)(E(z)-z_{\min})},\\ 1_{z_{\min}} & \text{with probability } \frac{(1+\delta)E(z)-R}{(1+\delta)(E(z)-z_{\min})}. \end{cases}\\
L(z) &= P\big(z \mid (1+\delta)E(z) \leqslant R\big).
\end{aligned} \tag{4.11}$$

We note that the two call option bounds become two increasing and *convex* functions $\bar{c}(S_0)$ and $\underline{c}(S_0)$ given by

$$\begin{aligned}
\bar{c}(S_0) &= \frac{1}{R}E^U\big[\big(S_0(1+\delta)z - K\big)^+\big],\\
\underline{c}(S_0) &= \frac{1}{R}E^L\big[\big(S_0(1+\delta)z - K\big)^+\big].
\end{aligned} \tag{4.12}$$

In the important special case $z_{\min} = 0$, the upper bound in (4.12) becomes

$$\bar{c}(S_0) = \frac{1}{(1+\delta)E[z]}E^P\big[\big(S_0(1+\delta)z - K\big)^+\big]. \tag{4.13}$$

Similar results are available for put options. We have thus shown that under the *no intermediate trading* assumption the option price is bound by two values given as the expectation of discounted payoff under two limiting distributions. Oancea and Perrakis (2006) provided corresponding bounds when $(1+\delta)\hat{z} \leqslant R$.

5 Special case: one period with transaction costs and general payoffs

In a one-period model with transaction costs and general payoffs, conditions (3.8)–(3.10) become

$$V_x(0) = R\sum_{i=1}^{I} p_i V_x^i(1), \tag{5.1}$$

$$V_y(0) = \sum_{i=1}^{I} p_i\left[\frac{S_i}{S_0}V_y^i(1) + \frac{\delta S_i}{S_0}V_x^i(1)\right] \tag{5.2}$$

and

$$(P_j - k_j)V_x(0) \leqslant \sum_{i=1}^{I} p_i G_j(S_i)V_x^i(1) \leqslant (P_j + k_j)V_x(0),$$
$$j = 1, \ldots, J. \tag{5.3}$$

Conditions (3.5)–(3.7) become[7]

$$V_x(0) > 0, V_y(0) > 0, V_x^i(1) > 0, V_y^i(1) > 0, \quad i = 1, \ldots, I, \tag{5.4}$$

$$V_y^1(1) \geqslant V_y^2(1) \geqslant \cdots \geqslant V_y^I(1) > 0 \tag{5.5}$$

and

$$(1 - k)V_x^i(1) \leqslant V_y^i(1) \leqslant (1 + k)V_x^i(1), \quad i = 1, \ldots, I. \tag{5.6}$$

As before, a useful way to pinpoint options that cause infeasibility or near-infeasibility of the problem is to single out a "test" option and solve the problems (3.11) and (3.12) subject to restrictions (5.1)–(5.6).

In order to highlight the difference in the formulation brought about by transaction costs, we adopt a notation similar to that in (4.1)–(4.5). We define $m_i \equiv V_x^i(1)/V_x(0)$, the marginal rate of substitution between the *bond* account at time one and the bond account at time zero and state i; and $\lambda_i \equiv V_y^i(1)/V_x(0)$, the marginal rate of substitution between the *stock* account at time one and the bond account at time zero and state i. Then (5.1)–(5.6) become

$$1 = R\sum_{i=1}^{I} p_i m_i, \tag{5.7}$$

[7] Since the value of the bond account at the end of the period is independent of the state i, the concavity conditions $V_{xx}(t) < 0$ and $V_{xx}(1)V_{yy}(1) - (V_{xy}(1))^2 > 0$ cannot be imposed. Only the concavity condition $V_{yy}(t) < 0$ is imposed as in Eq. (5.5).

$$(1-k) \leqslant \sum_{i=1}^{I} p_i z_i(\lambda_i + \delta m_i) \leqslant (1+k), \tag{5.8}$$

$$(P_j - k_j) \leqslant \sum_{i=1}^{I} p_i m_i G_j(z_i) \leqslant (P_j + k_j), \quad j = 1, \ldots, J, \tag{5.9}$$

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_I > 0 \tag{5.10}$$

and

$$(1-k)m_i \leqslant \lambda_i \leqslant (1+k)m_i, \quad i = 1, \ldots, I. \tag{5.11}$$

The bounds on the nth option with payoff $G_n(z_i)$ in state i are given by

$$\max_{m_i, \lambda_i}\Big(\text{or, } \min_{m_i, \lambda_i}\Big) \sum_{i=1}^{I} p_i m_i G_n(z_i). \tag{5.12}$$

Transaction costs double the number of variables that must be determined by the solution of the program. Furthermore, transaction costs expand the feasible region of the pricing kernel for any given set of option prices. Indeed, it is easy to see that for $k = 0$, $k_j = 0$, $j = 1, \ldots, J$ the problem (5.7)–(5.12) becomes identical to (4.1)–(4.5). Therefore, if a feasible solution to (4.1)–(4.4) exists, then this solution is feasible for (5.7)–(5.11) with $m_i = \lambda_i$, $i = 1, \ldots, I$. This implies that the spread between the two objective functions of (4.5) lies within the spread of the objective functions of (5.12).

6 Special case: two periods without transaction costs and general payoffs

The single-period model without transaction costs implies that the wealth at the end of the period is an increasing function of the stock price at the end of the period and, therefore, the pricing kernel is a decreasing function of the stock price at the end of the period. Likewise, the single period model with transaction costs implies that the value of the stock account at the end of the period is an increasing function of the stock price at the end of the period and, therefore, the marginal utility of wealth out of the stock account is a decreasing function of the stock price at the end of the period.

Constantinides and Zariphopoulou (1999) pointed out that intermediate trading invalidates the above implications with or without transaction costs, because the wealth at the end of the period (or, the value of the stock account at the end of the period) becomes a function not only of the stock price at the option's expiration but also of the entire sample path of the stock price.[8]

[8] In the special case of i.i.d. returns, power utility, and zero transaction costs, the wealth at the end of the period is a function only of the stock price. However, this assumption would considerably diminish the generality of the model.

Constantinides and Perrakis (2002) recognized that it is possible to recursively apply the single-period approach with or without transaction costs and derive stochastic dominance bounds on option prices in a market with intermediate trading over the life of the options.

In this section, we study a two-period model without transaction costs and, in the next section, a two-period model with transaction costs. In the absence of transaction costs, the value function $V(t) \equiv V(x_t, y_t, t)$ defined in (3.1)–(3.4) becomes a function of the aggregate trader wealth, $V(x_t + y_t, t)$. Therefore, we have $V_x(t) = V_y(t)$, $t = 0, 1, 2$. As before, we define the first period pricing kernel as $m_{1i} \equiv V_x^i(1)/V_x(0)$. For the second period, we define the pricing kernel as $m_{2ik} \equiv V_y^{ik}(2)/V_x(0)$, $i, k = 1, \ldots, I$. Then conditions (3.5)–(3.11) become

$$1 = R\sum_{i=1}^{I} p_i m_{1i}, \qquad 1 = R\sum_{k=1}^{I} p_k \frac{m_{2ik}}{m_{1i}}, \quad i = 1, \ldots, I, \tag{6.1}$$

$$1 = \sum_{i=1}^{I} p_i m_{1i}(1+\delta)z_i, \qquad 1 = \sum_{k=1}^{I} p_k \frac{m_{2ik}}{m_{1i}}(1+\delta)z_k, \quad i = 1, \ldots, I, \tag{6.2}$$

$$P_j = \sum_{i=1}^{I}\sum_{k=1}^{I} p_i p_k m_{2ik} G_j(z_i z_k), \quad j = 1, \ldots, J \tag{6.3}$$

and

$$m_{11} \geqslant m_{12} \geqslant \cdots \geqslant m_{1I} > 0, \qquad m_{2i1} \geqslant m_{2i2} \geqslant \cdots \geqslant m_{2iI} > 0, \quad i = 1, \ldots, I. \tag{6.4}$$

We test for feasibility by solving the program

$$\max_{m_{1i}, m_{2ik}} \left(\text{or, } \min_{m_{1i}, m_{2ik}}\right) \sum_{i=1}^{I}\sum_{k=1}^{I} p_i p_k m_{2ik} G_n(z_{1i} z_{2k}). \tag{6.5}$$

The extension of the program (6.1)–(6.5) to more than two periods becomes potentially explosive. In Section 8, we present closed form expressions for the bounds on the prices of European options in the special case where the payoff $G_j(S_T)$ is convex (call or put) and $J = 1$, by using the expressions developed in Section 4.2.

7 Special case: two periods with transaction costs and general payoffs

We now allow for transaction costs in the two-period model with general payoffs. Unlike Section 6, we have $V_x(t) \neq V_y(t)$, $t = 0, 1, 2$. We define the first period marginal rates of substitution as $m_{1i} \equiv V_x^i(1)/V_x(0)$ and

$\lambda_{1i} \equiv V_y^i(1)/V_x(0)$, $i = 1, \ldots, I$. We define the two-period marginal rates of substitution as $m_{2ik} \equiv V_x^{ik}(2)/V_x(0)$ and $\lambda_{2ik} \equiv V_y^{ik}(2)/V_x(0)$, $i, k = 1, \ldots, I$. Then conditions (3.5)–(3.11) become

$$1 = R\sum_{i=1}^{I} p_i m_{1i}, \qquad 1 = R\sum_{k=1}^{I} p_k \frac{m_{2ik}}{m_{1i}}, \quad i = 1, \ldots, I, \tag{7.1}$$

$$(1-k) \leqslant \sum_{i=1}^{I} p_i z_{1i}(\lambda_{1i} + \delta m_{1i}) \leqslant (1+k),$$

$$\lambda_{1i} = \sum_{k=1}^{I} p_k z_{2k}(\lambda_{2ik} + \delta m_{2ik}), \quad i = 1, \ldots, I, \tag{7.2}$$

$$P_j - k_j \leqslant \sum_{i=1}^{I}\sum_{k=1}^{I} p_i p_k m_{2ik} G_j(z_i z_k) \leqslant P_j + k_j, \quad j = 1, \ldots, J, \tag{7.3}$$

$$\lambda_{11} \geqslant \lambda_{12} \geqslant \cdots \geqslant \lambda_{1I} > 0, \qquad \lambda_{2i1} \geqslant \lambda_{2i2} \geqslant \cdots \geqslant \lambda_{2iI} > 0,$$
$$i = 1, \ldots, I \tag{7.4}$$

and

$$(1-k)m_{1i} \leqslant \lambda_{1i} \leqslant (1+k)m_{1i},$$
$$(1-k)m_{2ik} \leqslant \lambda_{2ik} \leqslant (1+k)m_{2ik}, \quad i = 1, \ldots, I, \ k = 1, \ldots, I. \tag{7.5}$$

As before, we test for feasibility by solving the program

$$\max_{m_{1i}, \lambda_{1i}, m_{2ik}, \lambda_{2ik}} \left(\text{or}, \min_{m_{1i}, \lambda_{1i}, m_{2ik}, \lambda_{2ik}}\right) \sum_{i=1}^{I}\sum_{k=1}^{I} p_i p_k m_{2ik} G_n(z_{1i} z_{2k}) \tag{7.6}$$

subject to (7.1)–(7.5). Constantinides et al. (2007) tested for violations of the stochastic dominance conditions (7.1)–(7.6).

In Section 9, we present closed form expressions for the bounds on the prices of European options for $T \geqslant 2$ in the special case where the payoff $G_j(S_T)$ is convex (call or put) and $J = 1$, by using the expressions developed in Section 4.2.

8 Multiple periods without transaction costs and with convex payoffs

For the case $J = 1$ and with convex payoffs, it is possible to use the special structure of the closed-form solution (4.8)–(4.12), in order to decompose the general problem into a series of one-period problems for any value of T. Indeed, consider the U and L distributions defined in (4.10) or (4.11) and define

the following recursive functions:

$$\bar{c}_t(S_t) = \frac{1}{R}E^U\big[\bar{c}_{t+1}\big(S_t(1+\delta)z_{t+1}\big) \mid S_t\big],$$
$$\underline{c}_t(S_t) = \frac{1}{R}E^L\big[\underline{c}_{t+1}\big(S_t(1+\delta)z_{t+1}\big) \mid S_t\big],$$
$$\bar{c}_T(S_T) = \underline{c}_T(S_T) = \big(S_{T-1}z_T(1+\delta) - K\big)^+. \tag{8.1}$$

In (8.1), the P, U and L distributions of the successive price ratios $z_{t+1} \equiv S_{t+1}/S_t$ are allowed to depend on the current index value S_t, *provided* such dependence preserves the convexity of the option value $c_t(S_t)$ at any time t with respect to S_t.

Assume that z_{t+1} takes I ordered values $z_{t+1,i}$, $i = 1, \ldots, I$ that determine the states at time $t+1$, set $c_{t+1,i} \equiv c_t(S_t(1+\delta)z_{t+1,i})$ and define at time t the variables $m_{t+1} : m_{t+1,i} \equiv V_y^i(t+1)/V_x(t)$, $i = 1, \ldots, I$. We can then show by induction that the expressions (8.1) define upper and lower bounds on the option value $c_t(S_t)$ at any time $t < T$.[9] We clearly have[10]

$$c_t(S_t) = \sum_{i=1}^{i=I} p_{t+1,i}m_{t+1,i}c_{t+1,i} = E^P\big[m_{t+1}c_t\big(S_t(1+\delta)z_{t+1}\big) \mid S_t\big]. \tag{8.2}$$

With these definitions consider now the program

$$\min(\text{or, } \max)_{\{m_{t+1,i}\}} c_t = \sum_{i=1}^{I} c_{t+1,i}p_{t+1,i}m_{t+1,i},$$
$$\text{subject to:} \quad 1 = \sum_{i=1}^{I}(1+\delta)z_{t+1,i}p_{t+1,i}m_{t+1,i},$$
$$1 = R\sum_{i=1}^{I} p_{t+1,j}m_{t+1,i},$$
$$m_{t+1,1} \geqslant m_{t+1,2} \geqslant \cdots \geqslant m_{t+1,I} > 0. \tag{8.3}$$

Given the assumed convexity of $c_{t+1} = c_t(S_t(1+\delta)z_{t+1})$, the solution of (8.3) produces upper and lower bounds on $c_t(S_t)$ that are discounted expectations of $c_t(S_t(1+\delta)z_{t+1})$ under the U and L distributions given by (4.10) or (4.11),

[9] The multiperiod upper bound in (8.1) was initially developed in Perrakis (1986). The lower bound was derived in Ritchken and Kuo (1988).

[10] In (8.2) the expectations are conditional on the stock price at time t. In fact the model is more general and the P-distribution may be allowed to depend on other variables such as, for instance, the current volatility of the stock price *provided* convexity is preserved and these other variables do not affect independently the trader's utility function.

conditional on S_t. The bounds on c_t are still given by the recursive expressions in (8.1).

Oancea and Perrakis (2006) addressed the asymptotic behavior of the multiperiod bounds in (8.1) as the number of trading dates increases. They considered specific cases of convergence of the P distribution to a particular stochastic process at the limit of continuous time. They showed that both the U and L distributions defined in (4.11) converge to a *single* risk-neutral stochastic process whenever the P distribution converges to a generalized diffusion, possibly a two-dimensional one, that preserves convexity of the option with respect to the underlying asset price.[11] A necessary and sufficient condition for the convergence of a discrete process to a diffusion is the *Lindeberg condition*, which was used by Merton (1982) to develop criteria for the convergence of binomial and, more generally, multinomial discrete time processes. This condition is applicable to multidimensional diffusion processes.

With minor reformulation, Oancea and Perrakis (2006) extended the validity of the bounds to stochastic volatility and GARCH models of the stock price. They also demonstrated that U and L converge to distinct limits when the limit of the P distribution is a mixed jump-diffusion process. They applied the stochastic dominance bounds to a discrete time process that converges to a mixed jump-diffusion process, in which the logarithm of the jump size amplitude G converges to a distribution with support $G \in [G_{\min}, G_{\max}]$, with $G_{\min} < 0 < G_{\max}$. The fact that the two option bounds converge to two different values is not particularly surprising. Recall that the bounds derived in earlier studies are also dependent either on the special assumption of fully diversifiable jump risk as in Merton (1976), or on the risk aversion parameter of the power utility function of the representative investor, as in Bates (1991) and Amin (1993). The option prices derived in these earlier studies are special cases located within the continuous time limits of the stochastic dominance bounds derived by (8.1).

9 Multiple periods with transaction costs and with convex payoffs

Constantinides and Perrakis (2002) recognized that it is possible to recursively apply the single-period approach with transaction costs and derive stochastic dominance bounds on option prices in a market with intermediate trading over the life of the options. The task of computing these bounds is easy compared to the full-fledged investigation of the feasibility of conditions (3.5)–(3.10) for large T for two reasons. As with the no transaction costs case, the derivation of the bounds takes advantage of the special structure of the payoff

[11] The conditions for the preservation of convexity were first presented by Bergman et al. (1996). Convexity is preserved in all one-dimensional diffusions and in most two-dimensional diffusions that have been used in the option pricing literature.

of a call or put option, specifically the convexity of the payoff as a function of the stock price. Second, the set of assets is limited to three assets: the bond, stock, and one option, the test option. Below, we state these bounds without proof.

At any time t prior to expiration, the following is an upper bound on the price of a call:

$$\bar{c}(S_t, t) = \frac{(1+k)}{(1-k)\{(1+\delta)\hat{z}\}^{T-t}} E\big[\big[(1+\delta)S_T - K\big]^+ \mid S_t\big], \tag{9.1}$$

where $(1+\delta)\hat{z}$ is the expected return on the stock per unit time. Observe that (9.1) is the same as the upper bound given in (4.13) for $z_{\min} = 0$ times the roundtrip transaction cost. The tighter upper bound given in (4.8), (4.11), and (8.1) does not survive the introduction of transaction costs and is eventually dominated by (9.1).

A partition-independent lower bound for a call option can also be found, but only if it is additionally assumed that there exists at least one trader for whom the investment horizon coincides with the option expiration, $T = T'$. In such a case, transaction costs become irrelevant in the put-call parity and the following is a lower bound[12]:

$$\begin{aligned} \underline{c}(S_t, t) &= (1+\delta)^{t-T} S_t - K/R^{T-t} \\ &\quad + E\big[(K - S_T)^+ \mid S_t\big] \big/ \big\{(1+\delta)\hat{z}\big\}^{T-t}, \end{aligned} \tag{9.2}$$

where R is one plus the risk free interest rate per unit time.

Put option upper and lower bounds also exist that are independent of the frequency of trading. They are given as follows:

$$\bar{p}(S_t, t) = \frac{K}{R^{T-t}} + \frac{1-k}{1+k}\big((1+\delta)\hat{z}\big)^{t-T}\big[E\big[[K - S_T]^+\big] - K \mid S_t\big], \tag{9.3}$$

and

$$\underline{p}(S_t, t) \begin{cases} \big((1+\delta)\hat{z}\big)^{t-T} \dfrac{1-k}{1+k} E\big[[K - S_T]^+ \mid S_t\big], & t \leqslant T-1, \\ [K - S_T]^+, & t = T. \end{cases} \tag{9.4}$$

The bounds presented in (9.1)–(9.4) may not be the tightest possible bounds for any given frequency of trading. Nonetheless, they have the property that they do not depend on the frequency of trading over the life of the option. For a comprehensive discussion and derivation of these and other possibly tighter bounds that are specific to the allowed frequency of trading, see Constantinides and Perrakis (2002). See also Constantinides and Perrakis (2007) for extensions to American-style options and futures options.

[12] In the special case of zero transaction costs, the assumption $T = T'$ is redundant because the put-call parity holds.

10 Empirical results

A robust prediction of the BSM option pricing model is that the volatility implied by market prices of options is constant across strike prices. Rubinstein (1994) tested this prediction on the S&P 500 index options (SPX), traded on the Chicago Board Options Exchange, an exchange that comes close to the dynamically complete and perfect market assumptions underlying the BSM model. From the start of the exchange-based trading in April 1986 until the October 1987 stock market crash, the implied volatility is a moderately downward-sloping function of the strike price, a pattern referred to as the "volatility smile", also observed in international markets and to a lesser extent on individual-stock options. Following the crash, the volatility smile is typically more pronounced.[13]

An equivalent statement of the above prediction of the BSM model, that the volatility implied by market prices of options is constant across strike prices, is that the *risk-neutral* stock price distribution is lognormal. Aït-Sahalia and Lo (1998), Jackwerth and Rubinstein (1996), and Jackwerth (2000) estimated the risk-neutral stock price distribution from the cross section of option prices.[14] Jackwerth and Rubinstein (1996) confirmed that, prior to the October 1987 crash, the risk-neutral stock price distribution is close to lognormal, consistent with a moderate implied volatility smile. Thereafter, the distribution is systematically skewed to the left, consistent with a more pronounced smile.

Several no-arbitrage models have been proposed and tested that generalize the BSM model. These models explore the effects of generalized stock price processes including stock price jumps and stochastic volatility and typically generate a volatility smile. The textbooks by Hull (2006) and McDonald (2005) provide excellent discussions of these models.

Economic theory imposes restrictions on equilibrium models beyond merely ruling out arbitrage. As we have demonstrated in Section 3, if prices are set by a utility-maximizing trader in a frictionless market, the pricing kernel must be a monotonically decreasing function of the market index price. To see this, the pricing kernel equals the representative agent's intertemporal marginal rate of substitution over each trading period. If the representative agent has *state independent* (derived) utility of wealth, then the concavity of the utility function implies that the pricing kernel is a decreasing function of wealth. Under the two maintained hypotheses that the marginal investor's (derived) utility of wealth is state independent *and* wealth is monotone increasing in the market index level, the pricing kernel is a decreasing function of the market index level.

[13] Brown and Jackwerth (2004), Jackwerth (2004), Shefrin (2005), and Whaley (2003) review the literature and potential explanations.

[14] Jackwerth (2004) reviews the parametric and non-parametric methods for estimating the risk-neutral distribution.

In a frictionless representative-agent economy, Aït-Sahalia and Lo (2000), Jackwerth (2000), and Rosenberg and Engle (2002) estimated the pricing kernel implied by the observed cross section of prices of S&P 500 index options as a function of wealth, where wealth is proxied by the S&P 500 index level. Jackwerth (2000) reported that the pricing kernel is everywhere decreasing during the pre-crash period 1986–1987 but widespread violations occur over the post-crash period 1987–1995. Aït-Sahalia and Lo (2000) reported violations in 1993 and Rosenberg and Engle (2002) reported violations over the period 1991–1995.[15] On the other hand, Bliss and Panigirtzoglou (2004) estimated plausible values for the risk aversion coefficient of the representative agent, albeit under the assumption of power utility, thus restricting the shape of the pricing kernel to be monotone decreasing in wealth.

Several theories have been suggested to explain the inconsistencies with the BSM model and the violations of monotonicity of the pricing kernel. Bollen and Whaley (2004) suggested that *buying pressure* drives the volatility smile while Han (2004) and Shefrin (2005) provided behavioral explanations based on *sentiment*. However, most of the discussion has focused on the *risk premia* associated with *stock market crashes* and *state dependence of the pricing kernel*.

Bates (2001) introduced heterogeneous agents with utility functions that explicitly depend on the number of stock market crashes, over and above their dependence on the agent's terminal wealth. The calibrated economy exhibits the inconsistencies with the BSM model but fails to generate the non-monotonicity of the pricing kernel. Brown and Jackwerth (2004) suggested that the reported violations of the monotonicity of the pricing kernel may be an artifact of the maintained hypothesis that the pricing kernel is state independent but concluded that volatility cannot be the sole omitted state variable in the pricing kernel.

Pan (2002), Garcia et al. (2003), and Santa-Clara and Yan (2004), among others, obtained plausible parameter estimates in models in which the pricing kernel is state dependent, using panel data on S&P 500 options. Others calibrated equilibrium models that generate a volatility smile pattern observed in option prices. Liu et al. (2005) investigated rare-event premia driven by uncertainty aversion in the context of a calibrated equilibrium model and demonstrated that the model generates a volatility smile pattern observed in option prices. Benzoni et al. (2005) extended the above approach to show that uncertainty aversion is not a necessary ingredient of the model. More significantly, they demonstrated that the model can generate the stark regime shift that occurred at the time of the 1987 crash. While not all of the above papers deal explicitly with the monotonicity of the pricing kernel, they do address the problem of reconciling the option prices with the historical index record.

[15] Rosenberg and Engle (2002) found violations when they used an orthogonal polynomial pricing kernel but not when they used a power pricing kernel.

These results are encouraging but stop short of demonstrating absence of stochastic dominance violations on a month-by-month basis in the cross section of S&P 500 options. This inquiry is the focus in Constantinides et al. (2007), hereafter CJP. CJP empirically investigated whether the observed cross sections of S&P 500 index option prices are consistent with various economic models that explicitly allow for a dynamically incomplete market and also recognize trading costs and bid–ask spreads. In the first part of their paper, CJP introduced transaction costs (trading fees and bid–ask spreads) in trading the index and options and investigated the extent to which violations of stochastic dominance, gross of transaction costs, are explained by transactions costs. They found that transaction costs decrease the frequency of violations but violations persist in several months both before and after the October 1987 crash.

Then CJP explored the second maintained hypothesis that every economic agent's wealth on the expiration date of the options is monotone increasing in the S&P 500 index price on that date. This assumption is unwarranted once we recognize that trading occurs over the (one-month) life of the options. With intermediate trading, a trader's wealth on the expiration date of the options is generally a function not only of the price of the market index on that date but also of the entire path of the index level. Thus the pricing kernel is a function not only of the index level but also of the *entire path* of the index level. CJP explored the month-by-month violations of stochastic dominance while allowing the pricing kernel to depend on the path of the index level.

In estimating the real distribution of the S&P 500 index returns, CJP refrained from adopting a particular parametric form of the distribution and proceeded in four different ways. In the first approach, they estimated the *unconditional* distribution as the histograms extracted from two different *historical* index data samples covering the periods 1928–1986 and 1972–1986. In the second approach, they estimated the *unconditional* distribution as the histograms extracted from two different *forward-looking* samples, one that includes the October 1987 crash (1987–2002) and one that excludes it (1988–2002). In the third approach, CJP modeled the variance of the index return as a GARCH (1, 1) process and estimated the *conditional* variance over the period 1972–2002 by the semiparametric method of Engle and Gonzalez-Rivera (1991) that does not impose the restriction that conditional returns are normally distributed. In the fourth approach, CJP used the VIX-implied volatility as an estimate of the *conditional* variance.

Based on the index return distributions extracted in the above four approaches, CJP tested the compliance of option prices with the predictions of models that sequentially introduce market incompleteness, transactions costs, and intermediate trading over the life of the options.

CJP's empirical design allows for at least three implications associated with state dependence. First, each month they searched for a pricing kernel to price the cross section of one-month options without imposing restrictions on the time series properties of the pricing kernel month by month. Thus they allowed

the pricing kernel to be state dependent. Second, in the second part of their investigation, CJP allowed for intermediate trading; a trader's wealth on the expiration date of the options is generally a function not only of the price of the market index on that date but also of the entire path of the index level thereby rendering the pricing kernel state dependent. Third, CJP allowed the variance of the index return to be state dependent and employed the estimated conditional variance.

A novel finding is that, even though pre-crash option prices follow the BSM model reasonably well, it does not follow that these options are correctly priced. Pre-crash option prices are incorrectly priced, if index return expectations are formed based on the historical experience. Furthermore, some of these prices lie below the theoretical bounds, contrary to received wisdom that historical volatility generally underprices options in the BSM model.

Another novel finding dispels the common misconception that the observed smile is too steep after the crash. Most of the violations post-crash are due to the option smile not being steep enough relative to expectations on the index price formed post-crash. Even though the BSM model assumes that there is no smile, an investor who properly understood the post-crash distribution of index returns should have priced the options with a steeper smile than the smile reflected in the actual option prices.

In all cases, there is a higher percentage of months with stochastic dominance violations by out-of-the-money calls (or, equivalently, in-the-money puts) than by in-the-money calls, suggesting that the mispricing is caused by the right-hand tail of the index return distribution and not by the left-hand tail. This observation is novel and contradicts the common inference drawn from the observed implied volatility smile that the problem lies with the left-hand tail of the index return distribution.

Finally, CJP found that the effect of allowing for one intermediate trading date over the life of the one-month options is to uniformly decrease the number of feasible months in each subperiod. They concluded that intermediate trading strengthens the single-period evidence of systematic stochastic dominance violations.

Constantinides et al. (2007) extended the results in CJP to American options on S&P 500 index *futures*. They demonstrated corresponding violations and implemented trading strategies that exploit the violations.

11 Concluding remarks

We presented an integrated approach to the pricing of options that allows for incomplete and imperfect markets. The BSM option pricing model is the nested case of complete and perfect markets. When the market is incomplete, imperfect, or both, the principle of no-arbitrage by itself implies restrictions on the prices of options that are too weak to be useful to either price options or confront the data with a testable hypothesis.

Instead of the principle of the absence of arbitrage that underlies the BSM model, we introduced the economic restriction that at least one risk-averse trader is a marginal investor in the options and the underlying security. Given the cross section of the prices of options and the real probability distribution of the return of the underlying security, the implied restrictions may be tested by merely solving a linear program. We also showed that the economic restrictions may be expressed in the form of upper and lower bounds on the price of an option, given the prices of the stock and the other outstanding options.

By providing an integrated approach to the pricing of options that allows for incomplete and imperfect markets, we provided testable restrictions on option prices that include the BSM model as a special case. We reviewed the empirical evidence on the prices of S&P 500 index options. The economic restrictions are violated surprisingly often, suggesting that the mispricing of these options cannot be entirely attributed to the fact that the BSM model does not allow for market incompleteness and realistic transaction costs. These are indeed exciting developments and are bound to stimulate further theoretical and empirical work to address the month-by-month pattern of option price violations.

Acknowledgements

Constantinides acknowledges financial support from the Center for Research in Security Prices of the University of Chicago and Perrakis from the Social Sciences and Humanities Research Council of Canada.

References

Aït-Sahalia, Y., Lo, A. (1998). Nonparametric estimation of state price densities implicit in financial asset prices. *Journal of Finance* 53, 499–547.

Aït-Sahalia, Y., Lo, A. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics* 94, 9–51.

Amin, K.I. (1993). Jump diffusion option valuation in discrete time. *Journal of Finance* 48, 1833–1863.

Bates, D.S. (1991). The crash of '87: Was it expected? The evidence from option markets. *Journal of Finance* 46, 1009–1044.

Bates, D.S. (2001). The market for crash risk. Working paper. University of Iowa, Iowa City.

Benzoni, L., Collin-Dufresne, P., Goldstein, R.S. (2005). Can standard preferences explain the prices of out-of-the-money S&P 500 options? Working paper. University of Minnesota.

Bergman, Y.Z., Grundy, B.D., Wiener, Z. (1996). General properties of option prices. *The Journal of Finance* 51, 1573–1610.

Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–659.

Bliss, R.R., Panigirtzoglou, N. (2004). Option-implied risk aversion estimates. *Journal of Finance* 59, 407–446.

Bollen, N., Whaley, R. (2004). Does net buying pressure affect the shape of implied volatility functions? *Journal of Finance* 59, 711–753.

Brown, D.P., Jackwerth, J. (2004). The kernel puzzle: Reconciling index option data and economic theory. Working paper. University of Wisconsin, Madison.

Constantinides, G.M. (1978). Market risk adjustment in portfolio valuation. *Journal of Finance* 33, 603–616.

Constantinides, G.M. (1979). Multiperiod consumption and investment behavior with convex transactions costs. *Management Science* 25, 1127–1137.

Constantinides, G.M., Perrakis, S. (2002). Stochastic dominance bounds on derivatives prices in a multiperiod economy with proportional transaction costs. *Journal of Economic Dynamics and Control* 26, 1323–1352.

Constantinides, G.M., Perrakis, S. (2007). Stochastic dominance bounds on American option prices in markets with frictions. *Review of Finance* 11, 71–115.

Constantinides, G.M., Zariphopoulou, T. (1999). Bounds on prices of contingent claims in an intertemporal economy with proportional transaction costs and general preferences. *Finance and Stochastics* 3, 345–369.

Constantinides, G.M., Zariphopoulou, T. (2001). Bounds on derivative prices in an intertemporal setting with proportional transaction costs and multiple securities. *Mathematical Finance* 11, 331–346.

Constantinides G.M., Czerwonko M., Jackwerth J., Perrakis S. (2007). Are options on index futures profitable for risk averse investors? Empirical Evidence. *Working paper*. University of Chicago, Chicago.

Constantinides, G.M., Jackwerth, J., Perrakis, S. (2007). Mispricing of S&P 500 index options. *Review of Financial Studies*, in press.

Cox, J., Ross, S.A. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3, 145–166.

Cox, J., Ross, S.A., Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7, 229–263.

Delbaen, F., Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* 300, 463–520.

Engle, R.F., Gonzalez-Rivera, G. (1991). Semiparametric ARCH models. *Journal of Business and Economic Statistics* 9/4, 345–359.

Garcia, R., Luger, R., Renault, E. (2003). Empirical assessment of an intertemporal option pricing model with latent variables. *Journal of Econometrics* 116, 49–83.

Han, B. (2004). Limits of arbitrage, sentiment and index option smile. Working paper. Ohio State University.

Harrison, J.M., Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20, 381–408.

Harrison, J.M., Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* 11, 215–260.

Huang, J. (2005). Option bounds and second order arbitrage opportunities. Working paper. Lancaster University.

Hull, J.C. (2006). *Options, Futures, and Other Derivatives*. Prentice Hall.

Jackwerth, J. (2000). Recovering risk aversion from option prices and realized returns. *Review of Financial Studies* 13, 433–451.

Jackwerth, J. (2004). Option-implied risk-neutral distributions and risk aversion. ISBN 0-943205-66-2, Research Foundation of AIMR, Charlotteville, USA.

Jackwerth, J., Rubinstein, M. (1996). Recovering probability distributions from option prices. *Journal of Finance* 51, 1611–1631.

Levy, H. (1985). Upper and lower bounds of put and call option value: Stochastic dominance approach. *Journal of Finance* 40, 1197–1217.

Liu, J., Pan, J., Wang, T. (2005). An equilibrium model of rare-event premia and its implications for option smirks. *Review of Financial Studies* 18, 131–164.

McDonald, R.L. (2005). *Derivatives Markets*. Addison–Wesley.

Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 141–184.

Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.

Merton, R.C. (1982). On the mathematics and economics assumptions of continuous-time models. In: *Essays in Honor of Paul Cootner*. Prentice Hall, Englewood Cliffs, NJ.
Oancea, I., Perrakis, S. (2006). Stochastic dominance and option pricing: An alternative paradigm. Working paper. Concordia University.
Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* 63, 3–50.
Perrakis, S. (1986). Option bounds in discrete time: Extensions and the pricing of the American put. *Journal of Business* 59, 119–141.
Perrakis, S., Ryan, P.J. (1984). Option pricing bounds in discrete time. *Journal of Finance* 39, 519–525.
Rendleman, R., Bartter, B. (1979). Two-state option pricing. *Journal of Finance* 34, 1092–1110.
Ritchken, P.H. (1985). On option pricing bounds. *Journal of Finance* 40, 1219–1233.
Ritchken, P.H., Kuo, S. (1988). Option bounds with finite revision opportunities. *Journal of Finance* 43, 301–308.
Rosenberg, J., Engle, R. (2002). Empirical pricing kernels. *Journal of Financial Economics* 64, 341–372.
Ross, S.A. (1976). Options and efficiency. *Quarterly Journal of Economics* 90, 75–89.
Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance* 3, 771–818.
Ryan, P.J. (2000). Tighter option bounds from multiple exercise prices. *Review of Derivatives Research* 4 (2), 155–188.
Ryan, P.J. (2003). Progressive option bounds from the sequence of concurrently expiring options. *European Journal of Operational Research* 151, 193–223.
Santa-Clara, P., Yan, S. (2004). Jump and volatility risk and risk premia: A new model and lessons from S&P 500 options. Working paper. UCLA.
Shefrin, H. (2005). *A Behavioral Approach to Asset Pricing*. Elsevier/North-Holland, Amsterdam.
Whaley, R.E. (2003). Derivatives. In: Constantinides, G.M., Harris, M., Stulz, R. (Eds.), *Financial Markets and Asset Pricing: Handbook of the Economics of Finance, vol. 1B*. In: *Handbooks in Economics*, vol. 21. Elsevier/North-Holland, Amsterdam.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15014-3

Chapter 14

Total Risk Minimization Using Monte Carlo Simulations

Thomas F. Coleman

Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: tfcoleman@uwaterloo.ca

Yuying Li

School of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: yuying@uwaterloo.ca

Maria-Cristina Patron

Risk Capital, 1790 Broadway, 15th Floor, New York, NY 10019, USA

Abstract

In an incomplete market, it is generally impossible to replicate an option exactly. In this case, total risk minimization chooses an optimal self-financing strategy that best approximates the option payoff by its final value. Total risk minimization is a dynamic stochastic programming problem, which is generally very challenging to solve; a direct approach may lead to very expensive computations.

We investigate total risk minimization using a piecewise linear criterion. We describe a method for computing the optimal hedging strategies for this stochastic programming problem using Monte Carlo simulation and spline approximations. We illustrate this method in the Black–Scholes and the stochastic volatility frameworks. We also compare the hedging performance of the strategies based on piecewise linear risk minimization, the traditional, quadratic risk minimizing strategies and the shortfall risk minimizing strategies. The numerical results show that piecewise linear risk minimization may lead to smaller hedging cost and significantly different, possibly better, hedging strategies. The values of the shortfall risk for the piecewise linear total risk minimizing strategies suggest that these strategies typically under-hedge the options.

1 Introduction

Hedging is a method for reducing the sensitivity of a portfolio to market fluctuations. In particular, when hedging an option, one tries to construct a

trading strategy that replicates the option payoff with no inflow or outflow of capital besides the initial costs. In the Black–Scholes framework, an option can be hedged by using only the underlying asset and a bond. However, the investor's position must be adjusted continuously, since it is only instantaneously risk-free. In practice, however, it is impossible to hedge continuously in time. In addition, one may want to hedge as little as possible due to transaction costs. If only discrete hedging times are allowed, achieving a risk-free position at each time is no longer possible since this instantaneous hedging will not last till the next rebalancing time. Moreover, presence of additional risks, e.g., jump risks, leads to an incomplete market. Under these conditions, it is not possible to totally hedge the intrinsic risk of an option that cannot be exactly replicated. There is much uncertainty regarding the choice of an optimal hedging strategy and in defining the fair price of an option.

El Karoui and Quenez (1995) use the super-replication method for pricing and hedging in incomplete markets. The method consists in finding a self-financing strategy of minimum initial cost such that its final value is always larger than the option payoff. This minimum initial cost represents the ask price, or the seller's price of the option. Correspondingly, the method computes a bid price, or a buyer's price. However, only an interval of no-arbitrage prices is determined in this manner. Moreover, there are cases when using a super-replicating strategy for hedging an option is not interesting from a financial point of view. For example, in the Hull–White (1987) stochastic volatility model, the super-replicating strategy for a call option is to hold the underlying asset (Frey, 1997). In addition, the minimum initial cost of a super-replicating strategy may be undesirably large.

Another approach to pricing and hedging in incomplete markets is to compute an optimal strategy by minimizing a particular measure of the intrinsic risk of the option. Föllmer and Schweizer (1989), Schäl (1994), Schweizer (1995, 2001), Mercurio and Vorst (1996), Heath et al. (2001a, 2001b), Bertsimas et al. (2001) study quadratic criteria for risk minimization. We only briefly describe them here, but they are presented in more detail in Section 2.

Suppose we want to hedge an option whose payoff is denoted by H and we only have a finite number of hedging times: $t_0, t_1, \ldots, t_M$. Suppose also that the financial market is modeled by a probability space $(\Omega, \mathcal{F}, P)$, with filtration $(\mathcal{F}_k)_{k=0,1,\ldots,M}$ and the discounted underlying asset price follows a square integrable process. Denote by V_k the value of the hedging strategy at time t_k and by C_k the cumulative cost of the hedging strategy up to time t_k (this includes the initial cost for setting up the hedging portfolio and the cost for rebalancing it at the hedging times $t_0, \ldots, t_k$).

Currently, there are two main quadratic hedging approaches for choosing an optimal strategy. One possibility is to control the total risk by minimizing the L^2-norm $E((H - V_M)^2)$, where $E(\cdot)$ denotes the expected value with respect to the probability measure P. This is the total risk minimization criterion. An optimal strategy for this criterion is self-financing, that is, its cumulative cost process is constant. A total risk minimizing strategy exists under the additional

assumption that the discounted underlying asset price has a bounded mean–variance tradeoff. In this case, the strategy is given by an analytic formula. The existence and the uniqueness of a total risk minimizing strategy have been extensively studied by Schweizer (1995).

Another possibility is to control the local incremental risk, by minimizing $E((C_{k+1} - C_k)^2|\mathcal{F}_k)$ for all $0 \leqslant k \leqslant M - 1$. This is the local quadratic risk minimizing criterion. The same assumption that the discounted underlying asset price has a bounded mean–variance tradeoff is sufficient for the existence of an explicit local risk minimizing strategy (see Schäl, 1994). This strategy is no longer self-financing, but it is mean-selffinancing, i.e., the cumulative cost process is a martingale. In general, the initial costs for the local risk minimizing and total risk minimizing strategies are different. As Schäl noticed, the initial costs agree in the case when the discounted underlying asset price has a deterministic mean–variance tradeoff. He then suggests the interpretation of this initial cost as a fair hedging price for the option. However, as mentioned by Schweizer (1995), this is not always appropriate.

The quadratic total and local risk minimizing hedging strategies have many theoretical properties, their existence and uniqueness have been extensively studied and, in the case of existence, they are given by analytic formula. However, the optimal hedging strategies hinge on the criteria for measuring the risk. Therefore, it is important to answer the natural question of how different hedging strategies are under different risk measures. Moreover, how should one choose a risk measure?

In the Black–Scholes framework, an option can be hedged completely, with no risk, i.e., zero in or out cashflows, besides the initial cost. When rebalancing can only be done at discrete times, a natural optimal hedging strategy is the one which minimizes the expected magnitude of the cashflows; this leads to the optimization problems, minimize $E(|H - V_M|)$, or minimize $E(|C_{k+1} - C_k| \, |\mathcal{F}_k)$, respectively.

Coleman et al. (2003) investigate the piecewise linear criterion for local risk minimization. They illustrate the fact that piecewise linear local risk minimization may lead to very different, possibly better, hedging strategies. These strategies have a larger probability of small hedging cost and risk, although a very small probability of larger cost and risk than the traditional quadratic risk minimizing strategies. Although there is no analytic solution to the piecewise linear local risk minimization problem, the optimal hedging strategies can be computed very easily.

In this paper, we investigate hedging strategies based on piecewise linear total risk minimization. Minimizing the piecewise linear risk, $E(|H - V_M|)$, and minimizing the quadratic risk, $E((H - V_M)^2)$ are also likely to yield significantly different solutions. Assume that $p(S)$ is the conditional density function of the underlying price at time T. Minimizing $E((H - V_M))$ puts more emphasis on reducing the largest value of $\sqrt{p(S)}|H - V_M|$, whereas minimizing $E(|H - V_M|)$ attempts to reduce the density weighted incremental cashflow, $p(S)|H - V_M|$ for each underlying value S equally.

To illustrate the above discussion in more detail, consider the following comparison between the piecewise linear risk minimization with respect to the total risk measure $E(|H - V_M|)$, and the quadratic risk minimization with respect to $E((H - V_M)^2)$. Suppose the price of the underlying asset satisfies the stochastic differential equation:

$$\frac{dS_t}{S_t} = \mu\, dt + \sigma\, dZ_t$$

where Z_t is a Wiener process. Let the initial value of the asset $S_0 = 100$, the instantaneous expected return $\mu = 0.2$, the volatility $\sigma = 0.2$ and the riskless rate of return $r = 0.1$. Suppose we want to statically hedge a deep in-the-money and a deep out-of-money put option with maturity $T = 1$; we only have one hedging opportunity, at time 0. At the maturity T we compare the payoff of the options with the hedging portfolio values of the strategies obtained by the piecewise linear and quadratic local risk minimization. The payoff and the hedging portfolio values at time T are multiplied by the density function of the asset price and are discounted to time 0. The first plot in Fig. 1 shows

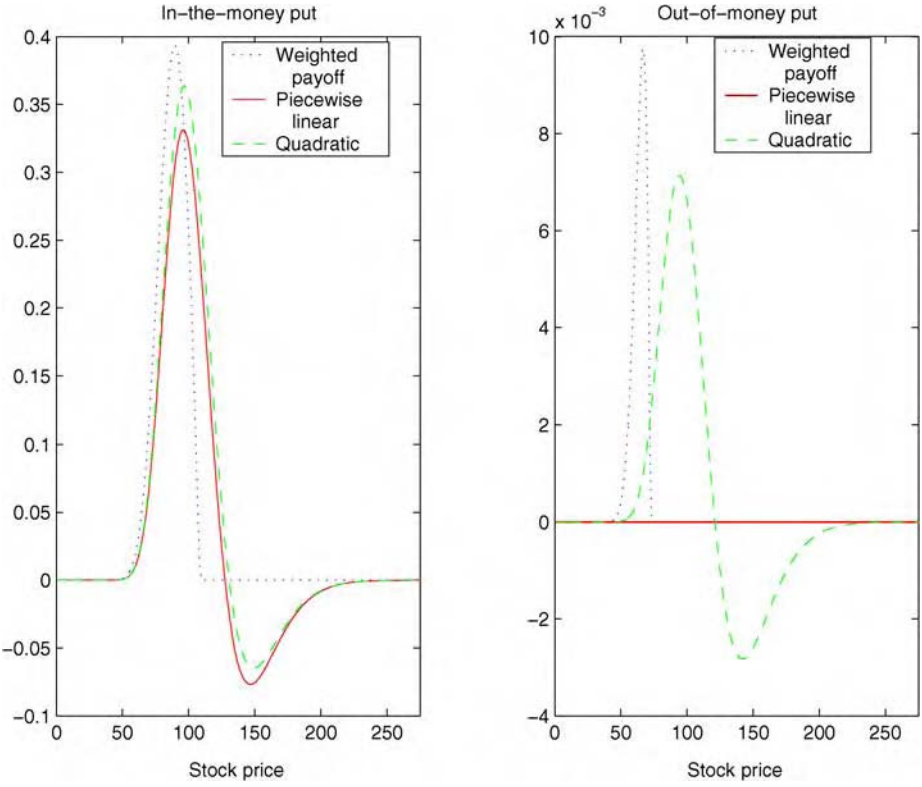


Fig. 1. Best fitting of the option payoff.

the density weighted payoff and the density weighted values of the hedging portfolios at the maturity T for the in-the-money put option. The second plot presents the corresponding data for the out-of-money put option.

In the case of the in-the-money put option, the weighted payoff, closer to lognormal, is much easier to fit. We remark that in this case both criteria generate similar plots of the hedging strategy values and they fit the option payoff relatively well. However, the weighted payoff for the out-of-money put option seems more difficult to match. Despite the small values (of order 10^{-3}), it is important to note that the relative differences between the weighted payoff and the weighted values of the hedging portfolios are large. (The cost of an out-of-money put is much smaller than the cost of the in-the-money put.) We have illustrated the hedging of only one out-of-money put option; if we want to hedge 100 put options identical to the one considered, the absolute differences between the weighted payoff and the weighted hedging portfolio values will also be significant. The hedging styles of the two strategies are very different. The L^2-norm (i.e., quadratic) attempts to penalize large residuals excessively and this actually leads to a worse fit under most scenarios. Indeed, the probability that the put option expires out of money is very large, around 0.97, but the L^2-hedging strategy either over or under replicates the option payoff. On the other hand, the L^1-strategy hedges exactly the option payoff when it expires out of money. Suppose we short the out-of-money put option. At the maturity of the option, our possible losses are never greater than the strike price. Assume now that we want to hedge our position by buying the L^2-hedging strategy. We can see from the figure that, by excessively trying to reduce the risk in the unlikely event that the option expires in the money, the L^2-strategy actually introduces the very small probability of unlimited losses. This is not the case if we try to hedge the short position using the L^1-strategy.

The main difficulty in computing the optimal strategies under the piecewise linear total risk minimization criterion is that, because these strategies are self-financing, the total risk, $H - V_M$, depends on the entire path of the stock price. Total risk minimization is a dynamic stochastic programming problem which is computationally challenging to solve. Using a tree method to model the future uncertainties may lead to very expensive computations for solving this stochastic programming problem, since the number of tree nodes increases exponentially as the number of trading opportunities increases. We propose a method for computing the piecewise linear total risk minimizing hedging strategies using Monte Carlo simulation and approximating the holdings in the hedging portfolios by unknown cubic splines which are determined as the solution to an optimization problem.

The key insight underlying our method is similar to the idea behind the Longstaff–Schwartz method for valuing American options (Longstaff and Schwartz, 2001). Essentially, the optimal exercise strategy for an American option is determined by the conditional expected value of the payoff from continuing to keep the option alive. Longstaff and Schwartz compute the optimal

exercise strategy for American options using Monte Carlo methods and approximating the conditional expected values of the payoff from continuation by functions of the state variables.

The method we propose for computing the optimal piecewise linear total risk minimizing strategies may also be useful in computing the quadratic total risk minimizing strategies, for example, in the case of the stochastic volatility models. Schweizer (1995) establishes an analytical formula for the computation of the quadratic risk minimizing strategies when the stock price has a bounded mean–variance tradeoff and Bertsimas et al. (2001) present a formula based on dynamic programming under the additional assumption of vector-Markov price processes. However, the numerical implementation of these formula may be quite involved in the stochastic volatility framework.

We illustrate our method in the Black–Scholes and stochastic volatility framework. We also investigate the differences between the hedging styles of the trading strategies based on piecewise linear and quadratic risk minimization. The behavior of the different hedging strategies for total risk minimization is similar to the one observed in the case of the local risk minimization (see Coleman et al., 2003). Piecewise linear total risk minimization generally leads to smaller hedging cost and risk than the corresponding quadratic criterion, although there is a very small probability of larger cost and risk.

Both quadratic and piecewise linear risk minimization are symmetric risk measures, since they penalize losses as well as gains. However, when hedging an option, one may be more interested in penalizing only the losses of his position. This leads to minimizing the shortfall risk, $E((H - V_M)^+)$. We remark that, while total risk minimization can be used for both hedging and pricing an option, shortfall risk minimization can only be used for hedging purposes. We investigate criteria for shortfall risk minimization and compare the optimal hedging strategies for these criteria with the quadratic and piecewise linear total risk minimizing strategies. The optimal hedging strategy performances depend on the moneyness of the options and the number of rebalancing opportunities. Analyzing the values of the shortfall risk for the optimal total risk minimizing strategies, suggests that, while quadratic total risk minimization shows no trend for either over-hedging, or under-hedging, the corresponding piecewise linear criterion typically under-hedges the options.

To summarize the main contributions of this paper, we firstly propose a computational method to approximate optimal hedging strategies for total risk minimization under the L^1-risk measure. Secondly, we compare the total risk minimizing hedging strategies for the L^1, L^2 and shortfall risk measures.

Section 2 of the paper describes the different risk minimization criteria for discrete hedging. In Section 3 we present our method for computing the piecewise linear total risk minimizing strategies. We illustrate this method in the Black–Scholes framework and compare the different criteria for total risk minimization in this framework. Section 4 has a similar analysis for a stochastic volatility framework. In Section 5 we investigate criteria for shortfall risk minimization and compare the performance of the hedging strategies for shortfall,

piecewise linear and quadratic total risk minimization. We conclude in Section 6.

2 Discrete hedging criteria

Consider a financial market where a risky asset (called stock) and a risk-free asset (called bond) are traded. Let $T > 0$ and assume we only have a finite number of hedging dates over the time horizon $[0, T]$. Let $0 = t_0 < t_1 < \cdots < t_M = T$ denote these discrete hedging times. Suppose the financial market is modeled as a filtered probability space $(\Omega, \mathcal{F}, P)$, with filtration $(\mathcal{F}_k)_{k=0,1,\ldots,M}$, where $\mathcal{F}_k$ corresponds to the hedging time t_k and w.l.o.g. $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is trivial. Suppose, moreover, that the stock price follows a stochastic process $S = (S_k)_{k=0,1,\ldots,M}$, with S_k being $\mathcal{F}_k$-measurable for all $0 \leqslant k \leqslant M$. We can set the bond price $B \equiv 1$ by assuming the discounted stock price process $X = (X_k)_{k=0,1,\ldots,M}$, where $X_k = \frac{S_k}{B_k}$, $\forall 0 \leqslant k \leqslant M$.

Assume that we want to hedge a European option with maturity T and payoff given by a $\mathcal{F}_M$-measurable random variable H. For example, $H = (K - X_M)^+$ for a European put with maturity T and discounted strike price K.

A trading strategy is given by two stochastic processes $(\xi_k)_{k=0,1,\ldots,M}$ and $(\eta_k)_{k=0,1,\ldots,M}$, where ξ_k is the number of shares held at time t_k and η_k is the amount invested in the bond at time t_k. We assume ξ_k, η_k are $\mathcal{F}_k$-measurable, for all $0 \leqslant k \leqslant M$ and $\xi_M = 0$. Consider the portfolio consisting of the combination of the stock and bond given by the trading strategy. The condition $\xi_M = 0$ corresponds to the fact that at time M we liquidate the portfolio in order to cover for the option payoff.

The value of the portfolio at any time t_k, $0 \leqslant k \leqslant M$, is given by

$$V_k = \xi_k X_k + \eta_k.$$

For all $0 \leqslant j \leqslant M - 1$, denote by $\Delta X_j = X_{j+1} - X_j$. With this notation, $\xi_j \Delta X_j$ represents the change in value due to the change in the stock price at time t_{j+1} before any changes in the portfolio. Therefore, the accumulated gain G_k is given by:

$$G_k(\xi) = \sum_{j=0}^{k-1} \xi_j \Delta X_j, \quad 1 \leqslant k \leqslant M$$

and $G_0 = 0$.

The cumulative cost at time t_k, C_k, is defined by:

$$C_k = V_k - G_k, \quad 0 \leqslant k \leqslant M.$$

A strategy is called self-financing if its cumulative cost process $(C_k)_{k=0,1,\ldots,M}$ is constant over time, i.e. $C_0 = C_1 = \cdots = C_M$. This is equivalent to $(\xi_{k+1} - \xi_k) X_{k+1} + \eta_{k+1} - \eta_k = 0$ (a.s.), for all $0 \leqslant k \leqslant M - 1$. In other

words, any fluctuations in the stock price can be neutralized by rebalancing ξ and η with no inflow or outflow of capital. The value of the portfolio for a self-financing strategy is then given by $V_k = V_0 + G_k$ at any time $0 \leqslant k \leqslant M$.

A market is complete if any claim H is attainable, that is, there exists a self-financing strategy with $V_M = H$ (a.s.). If the market is incomplete, for instance in the case of discrete hedging, a claim is, in general, non-attainable and a hedging strategy has to be chosen based on some optimality criterion.

One approach to hedging in an incomplete market is to first impose $V_M = H$. Since such a strategy cannot be self-financing, we should then choose the optimal trading strategy to minimize the incremental cost incurred from adjusting the portfolio at each hedging time. This is the *local risk minimization*. The traditional criterion for local risk minimization is the quadratic criterion, given by minimizing:

$$E\big((C_{k+1} - C_k)^2 \big| \mathcal{F}_k\big), \quad 0 \leqslant k \leqslant M-1. \tag{1}$$

This criterion is discussed in detail in Föllmer and Schweizer (1989), Schäl (1994), Schweizer (1995, 2001).

A quadratic local risk minimizing strategy is guaranteed to exist under the assumptions that H is a square integrable random variable, X is a square integrable process with bounded mean–variance tradeoff, that is:

$$\frac{(E(\Delta X_k|\mathcal{F}_k))^2}{\mathrm{Var}(\Delta X_k|\mathcal{F}_k)} \text{ is P-a.s. uniformly bounded.}$$

Moreover, this hedging strategy is given explicitly by

$$\begin{cases} \xi_M^{(l)} = 0, \quad \eta_M^{(l)} = H, \\ \xi_k^{(l)} = \frac{\mathrm{Cov}(\xi_{k+1}^{(l)} X_{k+1} + \eta_{k+1}^{(l)}, X_{k+1}|\mathcal{F}_k)}{\mathrm{Var}(X_{k+1}|\mathcal{F}_k)}, & 0 \leqslant k \leqslant M-1, \\ \eta_k^{(l)} = E((\xi_{k+1}^{(l)} - \xi_k^{(l)})X_{k+1} + \eta_{k+1}^{(l)} \mid \mathcal{F}_k), & 0 \leqslant k \leqslant M-1. \end{cases} \tag{2}$$

The choice of the quadratic criterion for risk minimization is, however, subjective. Alternatively, one can choose to minimize:

$$E\big(|C_{k+1} - C_k| \mid \mathcal{F}_k\big), \quad 0 \leqslant k \leqslant M-1. \tag{3}$$

As illustrated by Coleman et al. (2003), even if there is no analytic solution to the above piecewise linear risk minimization problem, an optimal hedging strategy can be easily computed. Criterion (3) for piecewise linear local risk minimization leads to significantly different hedging strategies and possibly better hedging results.

Another approach to hedging in an incomplete market is to consider only self-financing strategies. An optimal self-financing strategy is then chosen which best approximates H by its terminal value V_M. The quadratic criterion

for this total risk minimization is given by minimizing the L^2-norm:

$$E\big((H - V_M)^2\big) = E\left(\left(H - V_0 - \sum_{j=0}^{M-1} \xi_j \Delta X_j\right)^2\right). \tag{4}$$

By solving the total risk minimization problem (4), we obtain the initial value of the portfolio, V_0, and the number of shares, $(\xi_0, \ldots, \xi_{M-1})$. The amount invested in the bond, $(\eta_0, \ldots, \eta_M)$, is then uniquely determined since the strategy is self-financing. If the discounted stock price is given by a square integrable process with bounded mean–variance tradeoff and if the payoff is given by a square integrable random variable, then problem (4) has a unique solution. The existence and uniqueness of a total risk minimizing strategy under the quadratic criterion have been extensively studied by Schweizer (1995).

Schweizer gives an analytic formula which relates the holdings and the hedging portfolio values for the quadratic total risk minimizing strategy to the holdings and the portfolio values for the quadratic local risk minimizing strategy:

$$\begin{cases} V_0^{(t)} = \frac{E(H \prod_{j=0}^{M-1}(1-\beta_j \Delta X_j))}{E(\prod_{j=0}^{M-1}(1-\beta_j \Delta X_j))}, \\ \xi_M^{(t)} = 0, \\ \xi_k^{(t)} = \xi_k^{(l)} + \beta_k (V_k^{(l)} - V_0^{(t)} - G_k(\xi^{(t)})) + \gamma_k, \quad 0 \leqslant k \leqslant M-1, \end{cases} \tag{5}$$

where the processes $(\beta_k)_{k=0,\ldots,M-1}$ and $(\gamma_k)_{k=0,\ldots,M-1}$ are given by the formula:

$$\beta_k = \frac{E(\Delta X_k \prod_{j=k+1}^{M-1}(1 - \beta_j \Delta X_j)|\mathcal{F}_k)}{E(\Delta X_k^2 \prod_{j=k+1}^{M-1}(1 - \beta_j \Delta X_j)^2|\mathcal{F}_k)},$$

$$\gamma_k = \frac{E((V_T^{(l)} - G_T(\xi^{(l)}) - V_k^{(l)} + G_k(\xi^{(l)}))\Delta X_k \prod_{j=k+1}^{M-1}(1 - \beta_j \Delta X_j)|\mathcal{F}_k)}{E(\Delta X_k^2 \prod_{j=k+1}^{M-1}(1 - \beta_j \Delta X_j)^2|\mathcal{F}_k)}.$$

Bertsimas et al. (2001) also obtain a formula for the quadratic total risk minimizing strategy, using dynamic programming, in the case of vector-Markov price processes.

The corresponding piecewise linear total risk minimization criterion is given by the L^1-norm:

$$E\big(|H - V_M|\big) = E\left(\left|H - V_0 - \sum_{j=0}^{M-1} \xi_j \Delta X_j\right|\right). \tag{6}$$

We are interested in computing optimal hedging strategies given by the piecewise linear total risk minimization problem (6). This is a dynamic stochastic programming problem that is, in general, very difficult to solve. Since

$H - V_0 - \sum_{j=0}^{M-1} \xi_j \Delta X_j$ depends on the entire path of the stock price, a direct approach to problem (6) can be very expensive computationally. In order to see this, assume that we use Monte Carlo simulation and we generate L independent scenarios for the stock price. The total risk minimization problem (6) corresponds, in this case, to minimizing the expected total risk over all the scenarios:

$$\min_{\substack{V_0, \xi_0, \xi_j^{(k)} \\ \xi_j : \mathcal{F}_j\text{-measurable}}} \sum_{k=1}^{L} \left| H^{(k)} - V_0 - \xi_0 \Delta X_1^{(k)} - \sum_{j=1}^{M-1} \xi_j^{(k)} \Delta X_j^{(k)} \right|. \tag{7}$$

The notation $^{(k)}$ means that the option payoff, the stock price and the holdings correspond to the kth scenario. We remark that at time 0, the stock price is deterministic and, therefore, the holdings in the hedging portfolio at time 0 have to be the same for all the scenarios.

The number of unknowns in problem (7) is of order $L \cdot M$, where L is the number of scenarios and M is the number of rebalancing times. Therefore, trying to solve this problem directly is computationally very challenging when the number of scenarios is large and the rebalancing is frequent.

In order to reduce the complexity of problem (7) we try to approximate the holdings ξ_j. Spline functions have been extensively used for function approximations, since they are very attractive from a computational point of view. We choose to approximate the holdings ξ_j by unknown cubic splines.

The number of unknowns at each hedging time in the problem formulation (7) is equal to the number of scenarios; after approximating the holdings by cubic splines, the number of unknowns at each hedging time is reduced to the number of parameters in the cubic splines, which is typically very small.

An important issue to be considered when approximating the holdings in a hedging strategy by cubic splines is that the optimal hedging strategy has to be path dependent. Indeed, the total risk,

$$H - V_M = H - V_0 - \sum_{j=0}^{M-1} \xi_j \Delta X_j,$$

minimized by the optimal hedging strategy, depends on the entire path of the stock price. Although the holdings $(\xi_j)_{j=0,\ldots,M-1}$ are computed at time 0 and any measurable $(\xi_j)_{j=0,\ldots,M-1}$ is an admissible hedging strategy, intuitively, at any time t_j, $0 \leqslant j \leqslant M-1$, the optimal holdings ξ_j will have an intrinsic information about the past history of the stock price and the optimal holdings up to time t_j.

In this paper, we describe a method for solving the total risk minimization problem (6) by approximating the holdings in the optimal hedging strategy with unknown cubic splines and trying to capture the path dependency of the strategy by a simple spline formulation. The unknown cubic splines are determined

as solutions of an optimization problem that consists in minimizing the total risk over a set of scenarios for the stock price. Since the strategy computed in this way is suboptimal we have to analyze its degree of optimality. We also compare the hedging strategies based on the piecewise linear total risk minimization criterion, to the traditional strategies based on quadratic total risk minimization.

3 Total risk minimization in the Black–Scholes framework

We will first describe our method in the Black–Scholes framework. We suppose that the stock price is given by the stochastic differential equation:

$$\frac{\mathrm{d}S_t}{S_t} = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}Z_t, \tag{8}$$

where Z_t is a Brownian motion. We also assume that the writer of a European option with maturity T has only M hedging opportunities at $0 = t_0 < t_1 < \cdots < t_{M-1} < t_M := T$ to hedge his position using the underlying stock and a bond.

Using Monte Carlo simulation, we generate L independent samples for the stock price, based on Eq. (8). We want to determine the holdings in the hedging strategy such that the expected total risk over all the scenarios is minimized.

The total risk minimization problem for the piecewise linear criteria becomes:

$$\min_{\substack{V_0,\xi_0,\xi_j^{(k)} \\ \xi_j:\mathcal{F}_j\text{-measurable}}} \sum_{k=1}^{L} \left| H^{(k)} - V_0 - \xi_0 \Delta X_1^{(k)} - \sum_{j=1}^{M-1} \xi_j^{(k)} \Delta X_j^{(k)} \right|. \tag{9}$$

As before, the notation $^{(k)}$ refers to the kth scenario.

3.1 First formulation

We want to reduce the complexity of the above problem by approximating the holdings ξ_k. We first choose to ignore the fact that the hedging strategy is path dependent and assume that the amount $\xi_j X_j$ invested in the stock at any time t_j depends only on the stock price at time t_j. We will investigate the degree of optimality that can be achieved under this assumption and we will pursue subsequent refinement of this assumption. This is a natural assumption, since one should take into account the current value of the stock price, X_j, when rebalancing the portfolio at time t_j. Thus, we can assume:

$$\xi_j = D_j(X_j), \quad \forall j = 1, \ldots, M-1, \tag{10}$$

with D_j unknown functions. Let us suppose that the holdings depend continuously on the stock price, that is, D_j is a continuous function, $\forall j = 1, \ldots, M-1$. We denote by D_0 the constant function identically equal to ξ_0. The total risk minimization problem under the piecewise linear criterion becomes:

$$\min_{V_0, D_0, \ldots, D_{M-1}} \sum_{k=1}^{L} \left| H^{(k)} - V_0 - \sum_{j=0}^{M-1} D_j\big(X_j^{(k)}\big) \Delta X_j^{(k)} \right|. \tag{11}$$

In order to make the above problem computationally attractive, we assume that each function D_j is a cubic spline with fixed end conditions and spline knots placed with respect to the stock price. The function D_j is then uniquely determined by its values at the spline knots. Note that D_j is a linear function of its knot values. In this way, problem (11) becomes an L^1-optimization problem with unknowns V_0, D_0 and the values of the cubic splines D_j, $j \geqslant 1$, at their knots.

The number of knots for each spline in our implementation is typically very small (around 8) and independent of the number of scenarios. Therefore, the number of unknowns in the L^1-optimization problem (11) is of order M, where M is the number of rebalancing times. We can now solve this problem and compute the piecewise linear risk minimizing strategy that satisfies assumption (10) on the special form of the holdings ξ_j.

The question that arises is how good assumption (10) is. In order to answer this question, we will investigate the quadratic total risk minimization problem (4). We can compute the quadratic risk minimizing strategy either by solving an optimization problem similar to (11), or by using the theoretical formula (5). By comparing the hedging strategies obtained by these two methods, we will try to assert the quality of assumption (10).

We can modify the quadratic risk minimization problem (4), using an approach similar to the one described above for piecewise linear risk minimization. Under the assumption, $\xi_j = D_j(X_j)$, $\forall j = 1, \ldots, M-1$ and with the notation $D_0 \equiv \xi_0$, the problem becomes:

$$\min_{V_0, D_0, \ldots, D_{M-1}} \sum_{k=1}^{L} \left(H^{(k)} - V_0 - \sum_{j=0}^{M-1} D_j\big(X_j^{(k)}\big) \Delta X_j^{(k)} \right)^2. \tag{12}$$

We obtain, therefore, the optimal quadratic risk minimizing hedging strategy which satisfies assumption (10).

Another method for solving problem (4) is to use Schweizer's analytic solution (5) and compute the optimal quadratic risk minimizing strategy, in the general case, with no assumption on the form of the holdings. In the Black–Scholes model, the mean–variance of the stock price is not only bounded, but also deterministic. As mentioned in Schweizer's paper (Schweizer, 1995), for-

mula (5) reduces in this case to:

$$\begin{cases} V_0^{(t)} = V_0^{(l)}, \\ \xi_M^{(t)} = 0, \\ \xi_k^{(t)} = \xi_k^{(l)} + \alpha_k (V_k^{(l)} - V_0^{(l)} - G_k(\xi^{(t)})), \quad 0 \leqslant k \leqslant M-1, \end{cases} \tag{13}$$

where the process $(\alpha_k)_{k=0,\ldots,M-1}$ is given by:

$$\alpha_k = \frac{E(\Delta X_k | \mathcal{F}_k)}{E(\Delta X_k^2 | \mathcal{F}_k)}.$$

We first compute the quadratic local risk minimizing strategy, as given by formula (2). The details of this computation are given in Coleman et al. (2003). We then use formula (13) to obtain the holdings in the total risk minimizing hedging portfolio for each scenario.

The total risk minimizing hedging strategy computed from the analytical formula (13) in the above manner, is used as a benchmark for the solution of the quadratic risk minimization problem (12), in order to evaluate the validity of the assumption (10).

We also want to compare the effectiveness of the hedging strategies based on piecewise linear risk minimization and, respectively, quadratic risk minimization.

The numerical results presented below refer to hedging put options with maturity $T = 1$ and different strike prices. The initial stock price is $S_0 = 100$. The instantaneous expected return of the stock price is $\mu = 0.15$, the volatility, $\sigma = 0.2$ and the riskless rate of return, $r = 0.04$. The number of scenarios in the Monte Carlo simulation of the stock price is $L = 40{,}000$ and the number of time steps in this simulation is 600.

We have computed three risk minimizing hedging strategies:

- Strategy 1: Piecewise linear risk minimizing strategy satisfying (10).
- Strategy 2: Quadratic risk minimizing strategy satisfying (10).
- Strategy 3: Quadratic risk minimizing strategy given by the analytical formula (13).

For each of these strategies and each scenario, we compute the following:

- *Total cost*:

$$H - \sum_{k=0}^{M-1} \xi_k \Delta X_k. \tag{14}$$

This is the total amount of money necessary for the writer to implement the self-financing hedging strategy and honor the option payoff at expiry. Since the hedging strategy is self-financing, there are no intermediate costs for rebalancing the hedging portfolio.

- *Total risk*:

$$|H - V_M|. \tag{15}$$

This measures the difference between the final value of the hedging portfolio and the option payoff. The strategy being self-financing, it is the only unplanned cost or income.

Tables 1 and 2 show the average cumulative cost and average total risk over 40,000 simulated scenarios, for different number of time steps per rebalancing time. The last column in these tables correspond to the case of the static hedge, when we only have one hedging opportunity at time 0.

We remark that, in the case of Strategy 1, the average values of the cumulative cost in Table 1 and total risk in Table 2 are equal for some of the put options considered, as for example, the out-of-money put options with 1 or 2 hedging opportunities. This happens because the holdings in the optimal hedging portfolio of Strategy 1 are zero. Therefore, if the put option is not in-the-money and the number of rebalancing opportunities is sufficiently small, the optimal hedging Strategy 1 is not to hedge at all. This is intuitively quite reasonable since the likelihood of the option expiring out-of-money is large and one has no opportunity of further adjusting the hedging portfolio. The optimal hedging Strategies 2 and 3, on the other hand, still choose to hedge these particular

Table 1.
Average value of the total cost over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	2.2194	1.9764	1.0876	0.9398	0.9398
	2	2.4540	2.4033	2.3155	2.0400	1.7421
	3	2.4838	2.4387	2.3474	2.0429	1.7454
95	1	3.7878	3.6356	3.2435	1.6648	1.6648
	2	3.9512	3.8830	3.7647	3.4006	2.9735
	3	3.9770	3.9188	3.8022	3.4018	2.9745
100	1	5.8421	5.7082	5.5074	4.0392	2.7269
	2	5.9183	5.8396	5.6983	5.2566	4.6948
	3	5.9413	5.8773	5.7399	5.2565	4.6928
105	1	8.3549	8.2549	8.1113	7.2494	5.5301
	2	8.3613	8.2809	8.1280	7.6307	6.9449
	3	8.3866	8.3221	8.1724	7.6303	6.9392
110	1	11.2609	11.1988	11.0950	10.6364	9.2160
	2	11.2566	11.1789	11.0264	10.4994	9.7148
	3	11.2858	11.2221	11.0713	10.5007	9.7072

Notes. Average total cost for put options with $T = 1$, different strike prices and number of timesteps per rebalancing time, for strategies: 1 – piecewise linear with (10), 2 – quadratic with (10) and 3 – quadratic given by analytical formula; $S_0 = 100$, $\mu = 0.15$, $\sigma = 0.2$, $r = 0.04$.

Table 2.
Average value of the total risk over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	0.6031	0.7822	0.9276	0.9398	0.9398
	2	0.6312	0.8410	1.1212	1.5727	1.7707
	3	0.5336	0.7450	1.0377	1.5799	1.7759
95	1	0.7761	1.0419	1.3829	1.6648	1.6648
	2	0.7918	1.0771	1.4687	2.1945	2.6222
	3	0.6885	0.9641	1.3592	2.1993	2.6251
100	1	0.9790	1.2921	1.7293	2.5544	2.7269
	2	0.9877	1.3144	1.7784	2.7944	3.5117
	3	0.8295	1.1636	1.6479	2.7914	3.5119
105	1	1.1000	1.4535	1.9668	3.1622	3.9566
	2	1.1068	1.4677	2.0051	3.2892	4.3184
	3	0.9465	1.3180	1.8694	3.2774	4.3170
110	1	1.1240	1.5192	2.0798	3.4688	4.7912
	2	1.1308	1.5344	2.1240	3.6189	4.9366
	3	1.0147	1.4171	2.0036	3.6027	4.9355

Notes. Average total risk for the hedging of put options with different strike prices and different number of time steps per rebalancing time, for the three strategies and in the setup described in Table 1.

put options. We remark that out-of-money put options with more hedging opportunities are hedged by Strategy 1. Experiments show that out-of-money put options which are closer to expiry will be hedged by Strategy 1.

When the rebalancing is infrequent, the average values of the total risk for the quadratic risk minimizing Strategies 2 and 3 are very close. The same can be observed for the cumulative cost. However, as the rebalancing becomes frequent enough, the total risk for Strategy 2 becomes larger than the total risk for Strategy 3. The results maintain the same trend even if we increase the number of spline knots or change their position. This suggests that the constraint (10), on the form of the holdings leads to supplementary risk and a better assumption has to be found.

The numerical results in Tables 1 and 2 illustrate that the hedging strategies based on the piecewise linear and, respectively, quadratic risk minimization perform differently in terms of average cumulative cost and risk. In the case of the in-the-money put options, the values of the average cumulative cost are very close for all the three strategies. However, as the option becomes out-of-money and the rebalancing is less frequent, the average cumulative cost for Strategy 1 is almost half the average cumulative cost of the quadratic strategies. The average total risk has the same trend. Nevertheless, since it may be possible to eliminate part of the total risk for Strategies 1 and 2, by using a less restrictive constraint than (10), the above results do not show very clearly the difference between the piecewise linear and the quadratic risk minimiz-

ing strategies. The numerical results obtained with a better assumption on the form of the holdings will allow further discussion on this subject.

3.2 Second formulation

As illustrated above, the constraint that the holdings at any time t_j depend only on the current stock price, $\xi_j = C_j(X_j)$, may be too restrictive. In order to obtain a better formulation, let us analyze in more detail the holdings satisfying assumption (10). Consider the particular case of the at-the-money put options with 6 hedging opportunities. Figure 2 shows the number of shares in the optimal hedging portfolio after the third rebalancing opportunity, for the quadratic risk minimizing Strategies 2 and 3.

We can see that in the case of Strategy 3, for the same value of the current stock price, we may have different number of shares in the hedging portfolio for the different scenarios. This is because this hedging strategy depends not only on the current value of the stock price, but also on the path of the stock price up to the current time. However, since the holdings for Strategy 2 satisfy assumption (10), they only depend on the current stock price and this assumption can become too restrictive. Note, however, that the holdings obtained under (10) capture quite well the trend of the optional holdings. To further reduce the risk, we have to incorporate the dependence on the path of the stock price in the assumption on the form of the holdings.

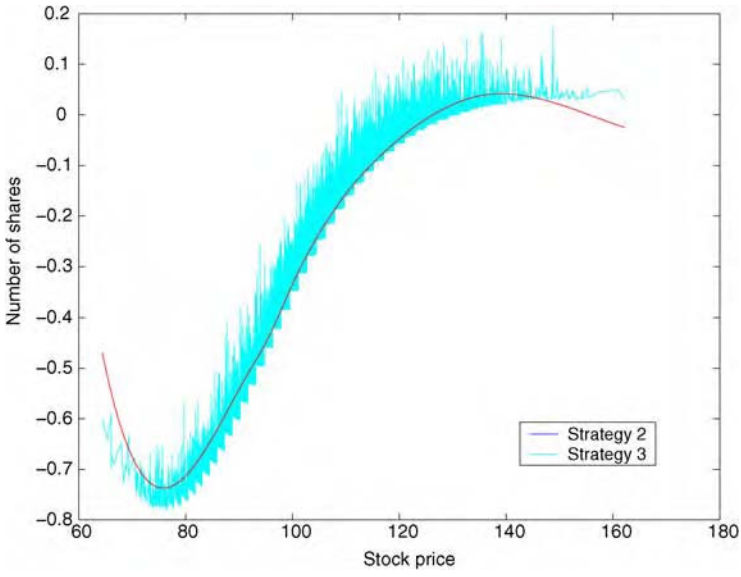


Fig. 2. Number of shares in the hedging portfolio after the third rebalancing time for the at-the-money put option with 6 rebalancing opportunities.

Strategy 2 considers only the hedging strategies for which the amount invested in the stock at any time t_j depends only on the stock price X_j at time t_j. It may be more natural to assume, however, that the investment in the stock at time t_j also depends on the cumulative gain up to time t_j. We assume that the holdings depend linearly on the past gain, specifically:

$$\xi_j = D_j(X_j) + \frac{1}{X_j}\sum_{i=0}^{j-1} \xi_i \Delta X_i, \quad \forall j = 1, \ldots, M-1$$

with D_j unknown cubic splines. As before, we make the convention $D_0 \equiv \xi_0$. After some algebraic manipulation and ignoring the higher order terms containing products $\Delta X_{i_1} \Delta X_{i_2}$, we obtain:

$$\xi_j = D_j(X_j) + \frac{1}{X_j}\sum_{i=0}^{j-1} D_i(X_i) \Delta X_i, \quad \forall j = 0, \ldots, M-1.$$

We introduce more degrees of freedom in the above formulation by allowing the effect of the current stock price, X_j, on the holdings at time t_j to be different from the effect of the past stock prices, $X_0, \ldots, X_{j-1}$.

The assumption on the form of the holdings ξ_j becomes:

$$\xi_j = D_j(X_j) + \frac{1}{X_j}\sum_{i=0}^{j-1} \tilde{D}_i(X_i) \Delta X_i, \quad \forall j = 0, \ldots, M-1, \tag{16}$$

where for $j \geqslant 1$, D_j and $\tilde{D}_j$ are unknown cubic splines with fixed end conditions and spline knots, while D_0, $\tilde{D}_0$ are constant functions. With this formulation, the piecewise linear optimization problem (6) becomes:

$$\min_{V_0, D_j, \tilde{D}_j} \sum_{k=1}^{L} \left| H^{(k)} - V_0 \right.$$
$$\left. - \sum_{j=0}^{M-1} \left(D_j\big(X_j^{(k)}\big) + \sum_{i=0}^{j-1} \tilde{D}_j\big(X_j^{(k)}\big) \frac{\Delta X_i^{(k)}}{X_j^{(k)}} \right) \Delta X_j^{(k)} \right|. \tag{17}$$

Problem (17) can be interpreted, similarly to problem (11), as a L^1-optimization problem with unknowns V_0, D_0, $\tilde{D}_0$ and the values of the cubic splines D_j, $\tilde{D}_j$, $j \geqslant 1$ at their knots.

The corresponding formulation for the quadratic risk minimization criterion is:

$$\min_{V_0, D_j, \tilde{D}_j} \sum_{k=1}^{L} \Bigg(H^{(k)} - V_0 - \sum_{j=0}^{M-1} \Bigg(D_j\big(X_j^{(k)}\big) + \sum_{i=0}^{j-1} \tilde{D}_j\big(X_j^{(k)}\big) \frac{\Delta X_i^{(k)}}{X_j^{(k)}} \Bigg) \Delta X_j^{(k)} \Bigg)^2. \tag{18}$$

We note that the number of knots for each spline is usually small (around 8). The number of unknowns in the above problems is approximately double to the number of unknowns in the previous formulation.

The optimization problems (17) and (18) allow us to compute the optimal piecewise linear and, respectively, quadratic risk minimizing strategies satisfying assumption (16) on the form of the holdings in the hedging portfolio. We can now investigate the quality of this assumption using the three strategies:

- Strategy 1: Piecewise linear risk minimizing strategy satisfying (16).
- Strategy 2: Quadratic risk minimizing strategy satisfying (16).
- Strategy 3: Quadratic risk minimizing strategy given by the analytical formula (13).

We first re-examine the case considered in Fig. 3 of the at-the-money put option with 6 hedging opportunities. The number of shares in the optimal hedging portfolio for Strategies 2 and 3, after the third rebalancing time is shown in Fig. 3. We remark that the values of the holdings for the optimal quadratic Strategy 2 satisfying constraint (16) follow closely the values of the holdings for the theoretical quadratic Strategy 3.

Tables 3 and 4 show the average values over 40,000 scenarios of the cumulative cost and total risk, as defined before, for the above hedging strategies and different numbers of hedging opportunities.

We remark that in the case of one hedging opportunity, the assumptions (10) and (16) on the form of the holdings coincide and, therefore, the last column in Tables 3 and 4 has the same results as the last column in Tables 1 and 2, respectively.

As noticed before, the optimal hedging Strategy 1 for some of the put options which are not in-the-money and have very few rebalancing opportunities, is not to hedge at all. This is shown by the fact that the holdings in the hedging portfolios for these options are zero, which implies that the average cumulative cost and the average total risk are equal.

In contrast with the numerical results presented earlier, the quadratic Strategies 2 and 3 now yield very close values for the average cumulative cost in Table 3 and, respectively, the average total risk in Table 4. We conclude that imposing the constraint (16) on the form of the holdings in the hedging portfolio does not affect significantly the optimal value of the average total hedging risk over the 40,000 simulated scenarios.

The numerical results suggest that assumption (16) leads to smaller average total hedging risk than assumption (10). In the case of the quadratic risk minimization, the average total hedging risk is very close to optimal. Therefore, we

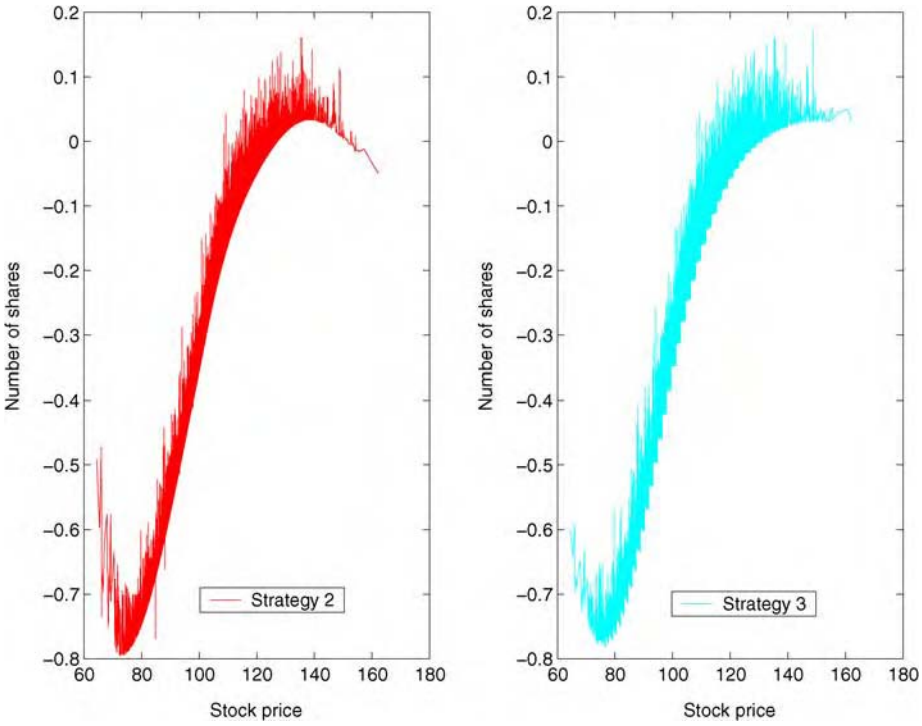


Fig. 3. Number of shares in the hedging portfolio after the third rebalancing time for the at-the-money put option with 6 rebalancing opportunities.

use the optimization problems (17) and (18) to compute the optimal hedging strategies under the piecewise linear and the quadratic risk minimizing criteria.

Tables 3 and 4 allow a clearer comparison of the hedging strategies based on the two criteria for risk minimization. We remark that the performance of these strategies depends on the moneyness of the options and on the number of rebalancing opportunities. The piecewise linear risk minimizing strategy yields a smaller average cumulative cost and risk for almost all the options considered. However, for in-the-money put options the values for the average cumulative cost and, respectively, total risk are close for all three strategies. The differences tend to increase as the put options are out-of-money and the rebalancing is less frequent. For the out-of-money put options with only 1 or 2 hedging opportunities the average cumulative cost for Strategy 1 is almost half the average cumulative cost for Strategies 2 and 3. The same happens for the average total risk.

Even if the market is incomplete due to the discrete hedging, many practitioners are still using delta hedging in order to hedge an option in the current framework. They choose a self-financing strategy such that the initial value of the hedging portfolio, V_0, is given by the value of the option at t_0, as computed

Table 3.
Average value of the total cost over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	2.2728	2.1093	1.5031	0.9398	0.9398
	2	2.4504	2.4086	2.3224	2.0388	1.7421
	3	2.4838	2.4387	2.3474	2.0429	1.7454
95	1	3.7964	3.6640	3.4080	1.6648	1.6648
	2	3.9443	3.8885	3.7741	3.3983	2.9735
	3	3.9770	3.9188	3.8022	3.4018	2.9745
100	1	5.8223	5.6896	5.5067	4.0644	2.7269
	2	5.9118	5.8455	5.7119	5.2530	4.6948
	3	5.9413	5.8773	5.7399	5.2565	4.6928
105	1	8.2982	8.1835	8.0393	7.2893	5.5301
	2	8.3584	8.2882	8.1412	7.6261	6.9449
	3	8.3866	8.3221	8.1724	7.6303	6.9392
110	1	11.2072	11.1146	10.9945	10.6934	9.2160
	2	11.2569	11.1881	11.0413	10.4945	9.7148
	3	11.2858	11.2221	11.0713	10.5007	9.7072

Notes. Average total cost for put options with different strike prices and number of time steps per rebalancing time, for the three strategies: 1 – piecewise linear with (16), 2 – quadratic with (16) and 3 – quadratic given by analytical formula; the same setup as in Table 1.

Table 4.
Average value of the total risk over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	0.5033	0.6819	0.8874	0.9398	0.9398
	2	0.5450	0.7497	1.0325	1.5722	1.7707
	3	0.5336	0.7450	1.0377	1.5799	1.7759
95	1	0.6575	0.9062	1.2512	1.6648	1.6648
	2	0.6952	0.9662	1.3551	2.1908	2.6222
	3	0.6885	0.9641	1.3592	2.1993	2.6251
100	1	0.8246	1.1269	1.5635	2.5524	2.7269
	2	0.8563	1.1789	1.6518	2.7843	3.5117
	3	0.8295	1.1636	1.6479	2.7914	3.5119
105	1	0.9380	1.2800	1.7897	3.1551	3.9566
	2	0.9722	1.3319	1.8802	3.2738	4.3184
	3	0.9465	1.3180	1.8694	3.2774	4.3170
110	1	1.0140	1.3806	1.9099	3.4619	4.7912
	2	1.0460	1.4279	2.0079	3.6025	4.9366
	3	1.0147	1.4171	2.0036	3.6027	4.9355

Notes. Average total risk for put options with different strike prices and number of time steps per rebalancing time, for the three strategies and in the setup described in Table 3.

by the Black–Scholes formula and the number of shares, ξ_k, at any hedging time t_k is equal to the delta of the option at t_k,

$$\xi_k = \left(\frac{\partial V}{\partial S}\right)_{t_k},$$

where V denotes the value of the option as given by the Black–Scholes formula. However, delta hedging insures a risk-free replication of the option only if the hedging is continuous. In the case of discrete rebalancing, delta hedging is no longer optimal since the corresponding portfolio is only instantaneously risk-free and the risk-free position does not last till the next rebalancing time. Tables 5 and 6 show the average values of the cumulative cost and risk over the 40,000 generated scenarios for the delta hedging strategy in comparison to the piecewise linear and quadratic risk minimizing strategies satisfying assumption (16) – Strategies 1 and 2, respectively.

We remark that when the rebalancing is frequent, the values of the total hedging cost and risk for the delta hedging strategy are very close, though slightly larger than the corresponding values for the piecewise linear and quadratic total risk minimizing strategies. However, as the number of rebalancing opportunities decreases, delta hedging an option leads to much larger hedging cost and risk than hedging the option by any of the two optimal hedging strategies for total risk minimization.

Table 5.
Average value of the total cost over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	2.2728	2.1093	1.5031	0.9398	0.9398
	2	2.4504	2.4086	2.3224	2.0388	1.7421
	Delta	2.5583	2.5859	2.6454	2.8838	3.2819
95	1	3.7964	3.6640	3.4080	1.6648	1.6648
	2	3.9443	3.8885	3.7741	3.3983	2.9735
	Delta	4.0702	4.1028	4.1763	4.4830	4.9793
100	1	5.8223	5.6896	5.5067	4.0644	2.7269
	2	5.9118	5.8455	5.7119	5.2530	4.6948
	Delta	6.0483	6.0897	6.1734	6.5382	7.1098
105	1	8.2982	8.1835	8.0393	7.2893	5.5301
	2	8.3584	8.2882	8.1412	7.6261	6.9449
	Delta	8.5011	8.5505	8.6407	9.0457	9.6607
110	1	11.2072	11.1146	10.9945	10.6934	9.2160
	2	11.2569	11.1881	11.0413	10.4945	9.7148
	Delta	11.4019	11.4537	11.5484	11.9712	12.5952

Notes. Average total cost for put options with different strike prices and number of time steps per rebalancing time, for the three strategies: 1 – piecewise linear with (16), 2 – quadratic with (16) and 3 – delta hedging; the same setup as in Table 3.

Table 6.
Average value of the total risk over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	0.5033	0.6819	0.8874	0.9398	0.9398
	2	0.5450	0.7497	1.0325	1.5722	1.7707
	Delta	0.6366	0.8935	1.2681	2.2099	3.2836
95	1	0.6575	0.9062	1.2512	1.6648	1.6648
	2	0.6952	0.9662	1.3551	2.1908	2.6222
	Delta	0.8042	1.1325	1.6160	2.8786	4.2846
100	1	0.8246	1.1269	1.5635	2.5524	2.7269
	2	0.8563	1.1789	1.6518	2.7843	3.5117
	Delta	0.9481	1.3385	1.9128	3.4582	5.1359
105	1	0.9380	1.2800	1.7897	3.1551	3.9566
	2	0.9722	1.3319	1.8802	3.2738	4.3184
	Delta	1.0576	1.4881	2.1282	3.8736	5.7216
110	1	1.0140	1.3806	1.9099	3.4619	4.7912
	2	1.0460	1.4279	2.0079	3.6025	4.9366
	Delta	1.1144	1.5725	2.2450	4.0892	5.9833

Notes. Average total risk for put options with different strike prices and number of time steps per rebalancing time, for the three strategies and in the setup described in Table 5.

Next we analyze the distributions of the cumulative cost and total risk for the at-the-money put option with 6 hedging opportunities. The average cumulative cost from Table 3 is 5.5067 for Strategy 1, 5.7119 for Strategy 2 and 5.7399 for Strategy 3. The histograms for each strategy of the cumulative cost over the 40,000 simulated scenarios are presented in Fig. 4. We mention that all three strategies have very few values of the cumulative cost larger than the range of values illustrated in Fig. 4, however, we chose this range in order to make the figure clearer.

The distribution of the cumulative cost for Strategy 1 is more asymmetric about its mean compared to the distributions for Strategies 2 and 3. About 60% of the cumulative costs for Strategy 1 are less than the mean, while in the case of the quadratic Strategies 2 and 3 the median is almost equal to the mean. The skewness of the distributions, which is another indication of the asymmetry of the data, is equal to 1.9012 for Strategy 1, 0.8017 for Strategy 2 and 0.8394 for Strategy 3.

We remark, however, that while Strategy 1 has a larger probability of smaller hedging cost, it also has a small probability of larger hedging costs than Strategies 2 and 3.

The next figure, Fig. 5, shows the histograms of the total risk over the simulated scenarios, for each hedging strategy. As in the case of Fig. 4, the range of values in Fig. 5 was chosen for clarity, though all the strategies lead to very few values of the total risk larger than the values in the chosen interval.

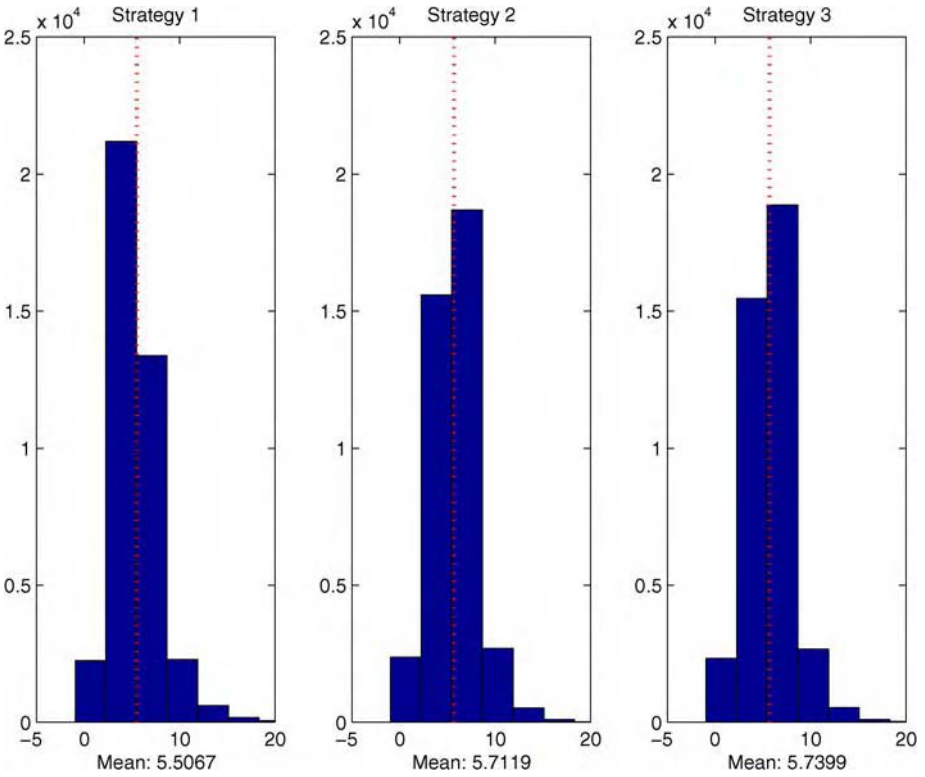


Fig. 4. Histograms of the total hedging cost over 40,000 scenarios.

The distributions of the total hedging risk for the three strategies have similar shapes. However, the mean for Strategy 1 is smaller than the mean for the quadratic strategies. The mean values of the total risk, as given in Table 4, are 1.5635, 1.6518 and, respectively, 1.6479. 65% of the total risk for Strategy 1 is less than the mean, while this happens 62% of the time for Strategies 2 and 3. The skewness in the case of Strategy 1 is 3.4414, larger than the skewness for Strategy 2, 2.0153, and Strategy 2, 2.1058. We note, however, that, as in the case of the total hedging cost, Strategy 1 has also a small probability of larger risk than Strategies 2 and 3. We remark that the distributions of Strategies 2 and 3, for both cumulative cost and risk, are very similar, another indication that (16) is sufficiently flexible to capture the optimal risk performance.

A similar behavior of the strategies based on the piecewise linear and quadratic criteria has been observed in the case of the local risk minimization, as shown by Coleman et al. (2003). Table 7 presents, for comparison, the average cumulative cost over the same 40,000 scenarios for the optimal piecewise linear (Strategy 1) and quadratic (Strategy 2) local risk minimizing hedging strategies. We do not include the results for the average risk, since the risk

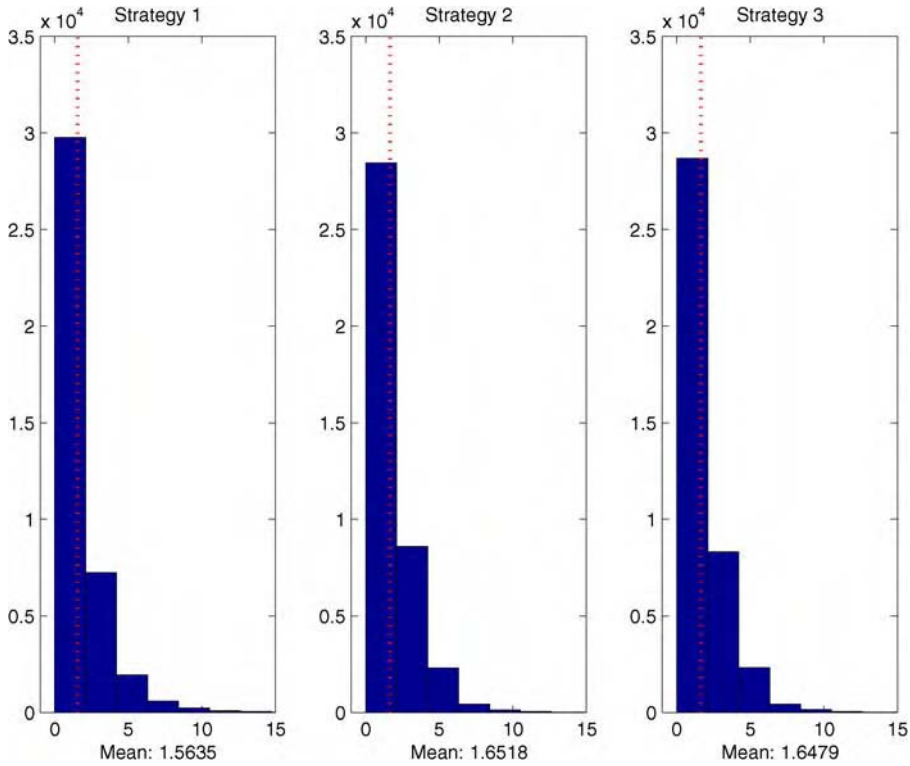


Fig. 5. Histograms of the total hedging risk over 40,000 scenarios.

measure has different meanings in the case of the local risk minimization and the total risk minimization.

As mentioned by Schäl (1994), when the stock price has a deterministic mean–variance tradeoff, the expected total hedging cost for the optimal quadratic local risk minimizing strategy is equal to the expected total hedging cost for the optimal quadratic total risk minimizing strategy. We remark that the average cumulative cost for the quadratic local risk minimizing Strategy 2 in Table 7 is very close to the average cumulative cost for the quadratic total risk minimizing Strategies 2 and 3 in Table 3 for all the put options considered. Schäl (1994) suggests the interpretation of the total hedging cost as a fair hedging price for the option. However, an example given by Mercurio and Vorst (1996), shows that this is not always appropriate.

We note that, in the case of static hedging, that is only one hedging opportunity, the local risk minimization and the total risk minimization criteria coincide. This is why the numerical results for the piecewise linear and the theoretical quadratic risk minimizing strategies in the last column of Table 3 are the same as the corresponding results in Table 7.

Table 7.
Average value of the total cost over 40,000 scenarios for local risk minimization.

Strike	Strategy	# of time steps per rebalancing time				
		25	50	100	300	600
90	1	2.1933	2.1043	1.8592	1.1690	0.9398
	2	2.4846	2.4377	2.3487	2.0424	1.7454
95	1	3.7284	3.6485	3.3907	2.1243	1.6648
	2	3.9785	3.9178	3.8036	3.4008	2.9745
100	1	5.7803	5.7698	5.5225	4.2964	2.7269
	2	5.9433	5.8765	5.7414	5.2550	4.6928
105	1	8.3483	8.4152	8.1908	7.4178	5.5301
	2	8.3889	8.3220	8.1738	7.6285	6.9392
110	1	11.3760	11.5276	11.3383	11.0652	9.2160
	2	11.2883	11.2226	11.0723	10.4989	9.7072

Notes. Average total for the hedging of put options with different strike prices and number of rebalancing opportunities, for the two strategies: 1 – piecewise linear local risk minimization, 2 – quadratic local risk minimization; the same setup as in Table 3.

In the case of the local risk minimization the hedging performance of the strategies also depends on the moneyness of the options and on the number of rebalancing opportunities, with the average cumulative cost for the piecewise linear local risk minimizing strategy being the smaller for the out-of-money and at-the-money put options. However, for in-the-money put options, the quadratic local risk minimizing strategy is slightly better, even though the values are close. The total risk minimization shows an improvement in terms of total hedging cost for the piecewise linear criterion, especially in the case of in-the-money put option. As a result, the average cumulative cost for the piecewise linear total risk minimizing strategy is the smallest for almost all the put options considered.

As shown by Coleman et al. (2003), the values of the optimal hedging portfolios for local risk minimization satisfy discrete hedging put-call parity. This is also true in the case of the total risk minimization, the proof being very similar.

Suppose that we have computed the optimal holdings ξ^p, η^p in the portfolio for hedging a put option with maturity T, discounted strike price K and M hedging opportunities at $0 = t_0 < t_1 < \cdots < t_{M-1} < t_M := T$. We can derive a relation between these holdings and the corresponding optimal holdings ξ^c, η^c for the call option on the same underlying asset and with the same maturity, strike price and hedging opportunities. We have the following property:

$$\begin{cases} \xi_k^c = \xi_k^p + 1, \\ \eta_k^c = \eta_k^p - K \end{cases}$$

for all $0 \leqslant k \leqslant M - 1$.

Moreover, the discounted values of the portfolios for hedging the put and the call options, V_k^p and V_k^c, satisfy the following put-call parity relation for all $0 \leqslant k \leqslant M$:

$$V_k^c - V_k^p = X_k - K.$$

Similarly, the relation between the cumulative costs for the call and put options is given by:

$$C_k^c = C_k^p + X_0 - K,$$

for all $0 \leqslant k \leqslant M$.

Therefore, if we know the optimal strategy for hedging the put option, we can compute the optimal strategy for the call, directly, without solving any optimization problems.

4 Total risk minimization in a stochastic volatility framework

In this section we assume that the stock price follows a Heston type stochastic volatility model (Heston, 1993). The discounted stock price X and its volatility Y satisfy a stochastic differential equation of the form:

$$\begin{aligned} \frac{\mathrm{d}X_t}{X_t} &= \alpha\,\mathrm{d}t + Y_t\,\mathrm{d}Z_t, \\ \mathrm{d}Y_t &= \left(\frac{4\beta\theta - \delta^2}{8Y_t} - \frac{\beta}{2}Y_t\right)\mathrm{d}t + \frac{\delta}{2}\,\mathrm{d}Z_t' \end{aligned} \tag{19}$$

where Z_t and Z_t' are Brownian motions with instantaneous correlation ρ.

In the Heston type model, the square of the volatility, $F := Y^2$ is a Cox–Ingersoll–Ross type process satisfying the stochastic differential equation:

$$\mathrm{d}F_t = \beta(\theta - F_t)\,\mathrm{d}t + \delta\sqrt{F_t}\,\mathrm{d}Z_t'. \tag{20}$$

As in the previous section, we assume the writer of a European option wants to hedge his position using only the underlying stock and a bond, but he only has a finite number of hedging opportunities.

Formula (5) given by Schweizer (1995), or the formula presented by Bertsimas et al. (2001), can be used to compute the optimal quadratic total risk minimizing strategy. We compute both the piecewise linear and quadratic risk minimizing strategies as given by the optimization problems (17) and (18) using Monte Carlo implementation.

Since the formulation of problems (17) and (18) depends on the entire stock price path, we are interested in generating strongly convergent discrete path approximations to the stochastic differential equations (19) and (20). We use Euler's method for Eqs. (19) and (20) to generate scenarios for the stock price and volatility.

The parameters for our numerical experiments are chosen as in Heath et al. (2001a, 2001b), in which the authors investigate continuous hedging under the total and local quadratic risk minimizing criteria and provide comparative numerical results for a class of stochastic volatility models. The values of the parameters are $\alpha = 0.5$, $\beta = 5$, $\theta = 0.04$, $\delta = 0.6$ and $\rho = 0$. As emphasized by Heath et al. (2001a, 2001b), these parameters satisfy Feller's test for explosions: $\beta\theta \geqslant \frac{1}{2}\delta$, which insures a positive solution for F_t in the stochastic differential equation (20). We generate 10,000 scenarios using 1024 time steps in Euler's method. We have also performed numerical experiments for 20,000 simulated scenarios, the results being very close in value to the results presented below. The initial stock price and volatility are $X_0 = 100$ and $Y_0 = 0.2$. The riskless rate of return is $r = 0.04$. As before, we want to hedge put options with maturity $T = 1$ and different strike prices.

We first assume that the holdings in the hedging portfolio depend on the current stock price and the past gains, their form being given by the constraint (16):

$$\xi_j = D_j(X_j) + \frac{1}{X_j}\sum_{i=0}^{j-1} \tilde{D}_i(X_i)\Delta X_i, \quad \forall j = 0, \ldots, M-1.$$

We remark that this constraint assumes the holdings are independent of the current volatility. This is attractive, since the volatility is not observable in the market. On the other hand, since the volatility is no longer constant in the current framework, it may be reasonable to assume that it also affects the form of the holdings. We will investigate later a different constraint on the form of the holdings which takes into account the volatility. However, the new formulation, while being computationally more expensive to implement, does not improve significantly the average total hedging cost and risk.

We compute the total risk minimizing strategies satisfying assumption (16):

- Strategy 1: Piecewise linear risk minimizing strategy.
- Strategy 2: Quadratic risk minimizing strategy.

Tables 8 and 9 present the average values of the total hedging cost and risk over the 10,000 simulated scenarios. We remark that the last column in these tables corresponds to the static hedging, when we only have one rebalancing opportunity, at time 0.

The above numerical results follow the trend observed in the Black–Scholes framework. For out-of-money and at-the-money put options the average cumulative cost and risk for the piecewise linear risk minimizing strategy are much smaller than the corresponding values for the quadratic risk minimizing strategy. The differences increase as the rebalancing is less frequent. For the deep out-of-money put options with very few hedging opportunities the values for the piecewise linear risk minimizing strategy are almost half the values for the quadratic risk minimizing strategy.

Table 8.
Average value of the total cost over 10,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		16	64	128	512	1024
90	1	1.9433	1.5446	1.1637	1.0199	1.0199
	2	2.3366	2.2365	2.2137	1.9469	1.7340
95	1	3.4682	3.2307	3.0079	1.7710	1.7738
	2	3.7234	3.6141	3.5726	3.2049	2.9003
100	1	5.4967	5.2277	5.1111	3.8699	2.8902
	2	5.5977	5.4786	5.4225	4.9567	4.5512
105	1	7.9197	7.7034	7.6502	7.0733	5.8629
	2	8.0112	7.8777	7.8106	7.2709	6.7681
110	1	10.8262	10.7099	10.6651	10.5153	9.5471
	2	10.9231	10.7769	10.7051	10.1219	9.5382

Notes. Average total cost for the hedging of put options with $T = 1$, different strike prices and number of rebalancing opportunities, for the two strategies satisfying (16): 1 – piecewise linear, 2 – quadratic; $X_0 = 100$, $Y_0 = 0.2$, $r = 0.04$, $\alpha = 0.5$, $\beta = 5$, $\theta = 0.04$, $\delta = 0.6$ and $\rho = 0$.

Table 9.
Average value of the total risk over 10,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time				
		16	64	128	512	1024
90	1	0.8395	0.9099	0.9901	1.0199	1.0199
	2	0.9727	1.0942	1.2546	1.7399	1.8985
95	1	1.1469	1.2728	1.4854	1.7737	1.7738
	2	1.2599	1.4251	1.6598	2.4190	2.7518
100	1	1.4342	1.5745	1.8701	2.7670	2.8902
	2	1.5274	1.7164	2.0283	3.1000	3.6495
105	1	1.6076	1.7925	2.1315	3.4513	4.1089
	2	1.7032	1.9303	2.3000	3.6521	4.4442
110	1	1.7004	1.9156	2.2754	3.7799	4.8597
	2	1.7915	2.0499	2.4533	4.0204	5.0373

Notes. Average total risk for the hedging of put options with different strike prices and number of rebalancing opportunities, for the two strategies and in the setup described in Table 8.

In the case of the in-the-money put options, the two strategies yield close values for the average cumulative cost and risk, with the piecewise linear risk minimizing strategy being better in most of the cases.

We can also analyze the distributions of the total hedging cost and risk for the two hedging strategies. Figure 6 shows the histograms of the total hedging

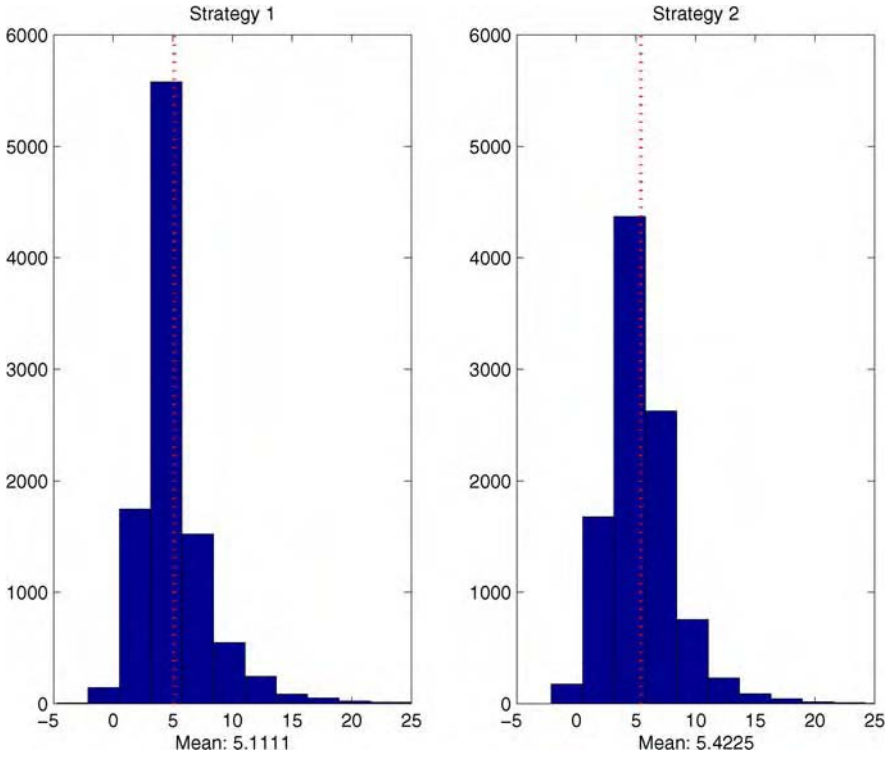


Fig. 6. Histograms of the total hedging cost over 10,000 scenarios.

cost over the 10,000 simulated scenarios for each strategy, in the case of the at-the-money put option with 8 hedging opportunities.

The average cumulative costs, given in Table 8, are 5.1111 for Strategy 1, and 5.4225 for Strategy 2. As in the Black–Scholes framework, the distribution of the cumulative cost for the piecewise linear risk minimizing strategy is more asymmetric about its mean than the distribution of the quadratic risk minimizing strategy. In the case of Strategy 1, 65% of the cumulative costs for Strategy 1 are less than the mean, while this happens only 55% of the time for Strategy 2. The skewness is 2.7526 for Strategy 1 and 1.3711 for Strategy 2. However, we remark again that piecewise linear risk minimization may lead, with a very small probability, to larger total hedging cost than the quadratic risk minimization.

Figure 7 presents the histograms of the total hedging risk for the same at-the-money put option with 8 hedging opportunities. As shown in Table 9, the average total hedging risk is 1.8701 in the case of Strategy 1 and 2.0283 in the case of Strategy 2.

The distributions of the total risk for both strategies are asymmetric about their mean. However, the mean for Strategy 1 is smaller than the mean for

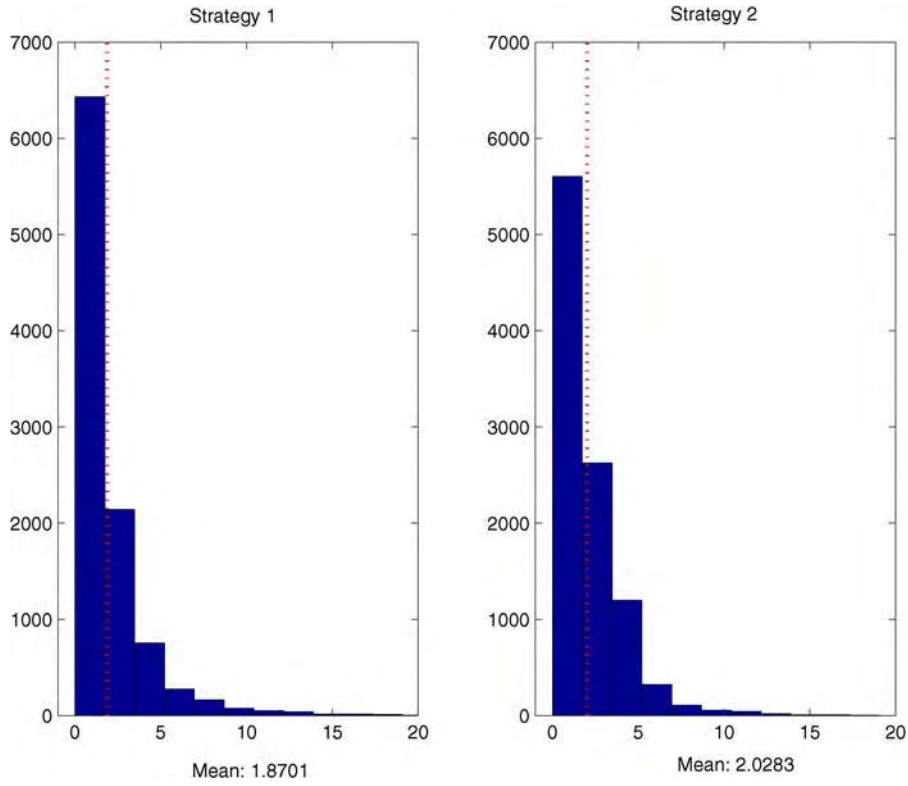


Fig. 7. Histograms of the total hedging risk over 10,000 scenarios.

Strategy 2. Strategy 1 yields smaller than the mean total risk 67% of the time, while this happens for 62% of the total risk for Strategy 2. The skewness is 4.0549 for Strategy 1 and 2.5346 for Strategy 2. The total risk for the piecewise linear risk minimizing strategy has a very small probability of larger values than the cumulative cost for the quadratic risk minimizing strategy.

We mention that the range of values illustrated in Figs. 6 and 7 was chosen for clarity, but both strategies can lead to values of the cumulative cost and risk larger than the values in the selected interval.

It is interesting to analyze the quadratic risk minimizing Strategy 2 as the number of hedging opportunities increases and compare it to the quadratic risk minimizing strategy for continuous trading. Such an analysis requires, however, a very thorough investigation and the simulation of a larger number of scenarios. For a very brief comparison, we illustrate the case of the in-the-money put option with maturity $T = 1$ and strike price $100 * \exp(r \cdot T)$, where r is the riskless rate of return. Heath et al. (2001a, 2001b) compute the expected cumulative hedging cost and the expected squared net loss $E((H - V_M)^2)$ for the continuous hedging of this option under the quadratic risk measure. They

Table 10.
Average total hedging cost and squared net loss for Strategy 2 over 10,000 scenarios.

Time steps	Cost	Net loss
1024	6.3182	32.9453
512	6.8041	22.5713
128	7.2715	9.5157
64	7.3155	6.4883
16	7.3260	3.7875

Notes. Average total cost and squared net loss for the hedging of the put option with $T = 1$ and strike price $100 * \exp(r \cdot T)$, for the quadratic risk minimizing strategy; the same setup as described in Table 8.

obtain an expected hedging cost of 7.691 and an expected squared net loss of 3.685. Table 10 shows the average over 10,000 of the cumulative hedging cost and squared net loss for the quadratic risk minimizing Strategy 2 as the number of time steps per rebalancing time decreases. We remark that, as the number of hedging opportunities increases, the average values of the cumulative hedging cost and squared net loss in Table 10, approach the values given by Heath et al. (2001a, 2001b).

We have remarked earlier in this section that the constraint (16) on the form of the holdings does not take into account the volatility Y_t. It may be reasonable to include the effect of the volatility on the holdings in the hedging portfolio and use the following constraint:

$$\xi_j = D_j(X_j, Y_j) + \frac{1}{X_j} \sum_{i=0}^{j-1} \tilde{D}_i(X_i, Y_i) \Delta X_i, \quad \forall j = 0, \ldots, M-1. \tag{21}$$

The unknown functions D_j, $\tilde{D}_j$, $j = 1, \ldots, M-1$, are now bicubic splines with fixed end conditions and knots placed with respect to the stock price and volatility. For each $j = 1, \ldots, M-1$, D_j and $\tilde{D}_j$ depend on the stock price and the volatility at time t_j. We assume, as before, that D_0 and $\tilde{D}_0$ are constant functions.

Solving an L^1-optimization problem similar to (17) and, respectively, an L^2-optimization problem similar to (18), we compute the total risk minimizing strategies satisfying assumption (21):

- Strategy 1: Piecewise linear risk minimizing strategy.
- Strategy 2: Quadratic risk minimizing strategy.

Since the assumption (21) involves bicubic splines, computing the above optimal strategies is much more expensive than computing the optimal strategies satisfying (16).

The average values of the cumulative hedging cost and risk for these two strategies over the 10,000 simulated scenarios are presented below, in Tables 11

Table 11.
Average value of the total cost over 10,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time					
		With assumption (18)			With assumption (13)		
		128	512	1024	128	512	1024
90	1	1.2055	1.0186	1.0199	1.1637	1.0199	1.0199
	2	2.1913	1.9507	1.7340	2.2137	1.9469	1.7340
95	1	2.9708	1.7715	1.7738	3.0079	1.7710	1.7738
	2	3.5455	3.2149	2.9003	3.5726	3.2049	2.9003
100	1	5.0478	3.9188	2.8902	5.1111	3.8699	2.8902
	2	5.3959	4.9729	4.5512	5.4225	4.9567	4.5512
105	1	7.6027	7.0814	5.8629	7.6502	7.0733	5.8629
	2	7.7734	7.2961	6.7681	7.8106	7.2709	6.7681
110	1	10.6063	10.5019	9.5471	10.6651	10.5153	9.5471
	2	10.6836	10.1544	9.5382	10.7051	10.1219	9.5382

Notes. Average total cost for the hedging of put options with different strike prices and number of rebalancing opportunities, for the two strategies satisfying (21): 1 – piecewise linear, 2 – quadratic; the same setup as described in Table 8.

Table 12.
Average value of the total risk over 10,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time					
		With assumption (21)			With assumption (16)		
		128	512	1024	128	512	1024
90	1	0.9464	1.0193	1.0199	0.9901	1.0199	1.0199
	2	1.2028	1.7407	1.8985	1.2546	1.7399	1.8985
95	1	1.4197	1.7726	1.7738	1.4854	1.7737	1.7738
	2	1.6107	2.4152	2.7518	1.6598	2.4190	2.7518
100	1	1.8175	2.7492	2.8902	1.8701	2.7670	2.8902
	2	1.9414	3.0880	3.6495	2.0283	3.1000	3.6495
105	1	2.1104	3.4145	4.1089	2.1315	3.4513	4.1089
	2	2.2950	3.6276	4.4442	2.3000	3.6521	4.4442
110	1	2.2754	3.7335	4.8597	2.2754	3.7799	4.8597
	2	2.4102	3.9797	5.0373	2.4533	4.0204	5.0373

Notes. Average total risk for the hedging of put options with different strike prices and number of rebalancing opportunities, for the two strategies and in the setup described in Table 11.

and 12, respectively. In order to make the comparison easier, we also reproduce the corresponding results from Tables 8 and 9. We remark that in the case of the static hedging, assumptions (16) and (21) coincide. This is why, in

Tables 11 and 12, the columns for 1024 time steps per rebalancing time, which correspond to static hedging in our implementation, coincide.

Computing the optimal strategies satisfying the constraint (21) on the form of the holdings is expensive, however, these strategies do not lead to significantly better cumulative hedging cost or risk, as can be seen by comparing the values of the cumulative cost and risk for these strategies to the corresponding values for the optimal hedging strategies satisfying the constraint (16). Moreover, assumption (21) relies on the values of the volatility, which are not directly observable in the market. In conclusion, it seems reasonable to compute the optimal hedging strategies in this framework by solving the optimization problems (17) and (18), even if their formulation takes into account only the dependence of the holdings in the hedging portfolio on the stock price path.

The numerical results presented in this section refer to hedging put options. However, as mentioned at the end of Section 3, hedging call options is closely related to hedging put options on the same underlying asset and with the same maturity and strike price. The optimal hedging portfolio values satisfy discrete hedging put-call parity. Moreover, if the holdings in the optimal portfolio for hedging the put options are known, the optimal holdings for the call options can be computed directly, without solving any optimization problems.

5 Shortfall risk minimization

An important criticism of the quadratic risk minimizing criterion, which is also valid for the piecewise linear risk measure, is the fact that it penalizes symmetrically losses, as well as gains.

It has been argued (see Bertsimas et al., 2001) that, in the case of pricing an option, a symmetric risk measure is the natural choice, since we do not know a priori if the option is being sold or purchased. However, when hedging an option, one tries to replicate the option payoff by constructing a hedging portfolio and he may be interested in penalizing only the costs and not also the profits from his position.

We will investigate here only the perspective of the writer of an option. When using a self-financing strategy to hedge an option with payoff H and maturity T, the total risk for the writer of the option is given by the difference between the payoff H and the final value of the hedging strategy, V_M. Even if V_M does not match exactly H, if $V_M \geqslant H$ the writer is still on the safe side, that is, he can cover the option payoff with no suplementary inflow of capital. Therefore, the writer of the option may prefer to choose a hedging strategy that minimizes only the shortfall risk, $E((H - V_M)^+)$:

$$\min E\big((H - V_M)^+\big), \tag{22}$$

and not the total risk, $E(|H - V_M|)$ or $E((H - V_M)^2)$.

A self-financing hedging strategy such that $V_M \geqslant H$, a.s., is called a super-replicating strategy. Unfortunately, the minimum initial cost of a super-replicating strategy is often too high. Moreover, in practice, one may be inclined not to use a super-replicating hedging strategy if he can make higher profits by accepting the risk of a loss.

In order to see that it can be quite expensive to super-replicate an option, we compare the minimum initial cost of a super-replicating strategy – obtained by minimizing $E((H-V_M)^+)$ – with the initial cost of the total risk minimizing strategies described in Section 3.1 – computed by minimizing $E(|H-V_M|)$ and $E((H-V_M)^2)$, respectively. The numerical results refer to the hedging put options with maturity $T=1$ and different strike prices when we only have a finite number of hedging opportunities at $0=t_0<t_1<\cdots<t_M:=T$. The stock price follows a Black–Scholes model with instantaneous expected return $\mu=0.15$ and volatility $\sigma=0.2$. The initial stock price is $S_0=100$. We generate 40,000 scenarios for the stock price using Monte Carlo simulation. The riskless rate of return is $r=0.04$.

An optimal super-replicating strategy for (22) can be obtained in a similar way to the computation of a total risk minimizing strategy described in Section 3.1, by assuming that the optimal holdings have the special form given by (16). Moreover, since,

$$(H-V_M)^+=\frac{1}{2}\big(H-V_M+|H-V_M|\big) \tag{23}$$

problem (22) can be implemented as a linear programming problem.

Table 13 shows the minimum initial cost for a super-replicating strategy satisfying assumption (16), in comparison with the initial cost of the piecewise linear total risk minimizing strategy – Strategy 1 – and the quadratic total risk minimizing strategy – Strategy 2 – satisfying the same assumption.

We can see from Table 13 that buying the initial portfolio for super-hedging is much more expensive than buying the initial portfolio for total risk minimization. Therefore, computing a hedging strategy by simply minimizing the shortfall risk $E((H-V_M)^+)$ is not very attractive from a practical point of view, even if a super-replicating strategy prevents the risk of any loss at the maturity of the option. In these conditions, an investor who still wants to penalize only the shortfall risk, but has a given initial capital and is willing to accept some risk of loss, may choose an optimal self-financing hedging strategy in the following way:

$$\begin{aligned}&\min\ E\big((H-V_M)^+\big)\\ &\text{s.t.}\quad V_0 \text{ given.}\end{aligned} \tag{24}$$

The above criteria for minimizing the shortfall risk has been studied by Föllmer and Leukert (2000), and Runggaldier (2001).

Alternative to penalizing the positive values of $H-V_M$, by minimizing $E((H-V_M)^+)$, one may try to penalize those values which are above the mean.

Table 13.
Initial portfolio cost.

Strike	Strategy	# of time steps per rebalancing time			
		50	100	300	600
90	Super-replicate	7.4806	10.3742	19.5669	28.1378
	1	1.9022	1.0070	0.0000	0.0000
	2	2.4086	2.3224	2.0388	1.7421
95	Super-replicate	9.7100	12.7437	22.2787	32.2861
	1	3.5152	3.0875	0.0000	0.0000
	2	3.8885	3.7741	3.3983	2.9735
100	Super-replicate	12.3146	15.3754	24.9454	35.7017
	1	5.5279	5.2248	2.7110	0.0000
	2	5.8455	5.7119	5.2530	4.6948
105	Super-replicate	15.3226	18.1656	27.7273	39.3592
	1	8.0693	7.8209	6.5076	3.2595
	2	8.2882	8.1412	7.6261	6.9449
110	Super-replicate	18.9710	21.4535	30.8217	43.1454
	1	11.0098	10.9945	10.1126	7.6382
	2	11.1881	11.0413	10.4945	9.7148

Notes. Initial portfolio cost for put options with different strike prices and number of time steps per rebalancing time, for the three strategies: super-replicating, 1 – piecewise linear and 2 – quadratic; the same setup as in Table 1.

This corresponds to minimizing:

$$E\big(\big(H - V_M - E(H - V_M)\big)^+\big). \tag{25}$$

However, note that, since for a self-financing strategy, $V_M = V_0 + \sum_{k=0}^{M-1} \xi_k \Delta X_k$, we have:

$$\begin{aligned} &H - V_M - E(H - V_M) \\ &\quad = H - E(H) - \sum_{k=0}^{M-1} \xi_k \Delta X_k + E\left(\sum_{k=0}^{M-1} \xi_k \Delta X_k\right). \end{aligned}$$

Therefore, the initial value of the hedging portfolio, V_0 cannot be determined by minimizing (25). In these conditions, a natural idea is to impose the constraint:

$$E(H - V_M) = 0 \quad \Leftrightarrow \quad V_0 = E\left(H - \sum_{k=0}^{M-1} \xi_k \Delta X_k\right), \tag{26}$$

that is, the initial value of the hedging portfolio is equal to the expected value of the difference between the option payoff and the cumulative gain of the

portfolio. With this constraint, criterion (25) becomes:

$$\begin{aligned} &\min\ E\big((H - V_M)^+\big) \\ &\text{s.t.}\quad E(H - V_M) = 0. \end{aligned} \tag{27}$$

By (23), this criterion is equivalent to:

$$\begin{aligned} &\min\ E\big(|H - V_M|\big) \\ &\text{s.t.}\quad E(H - V_M) = 0. \end{aligned} \tag{28}$$

Assuming that the holdings have the special form given by (16):

$$\xi_j = D_j(X_j) + \frac{1}{X_j}\sum_{i=0}^{j-1} \tilde{D}_i(X_i)\Delta X_i, \quad \forall j = 0, \ldots, M-1,$$

an optimal strategy for the above problem can be computed is a similar way to the piecewise linear total risk minimization problem (6).

We remark that the shortfall risk minimization problem (27) is not equivalent to problem (24), since (27) imposes a relation between the optimal holdings ξ_k and the initial value of the hedging portfolio, V_0.

In order to investigate the two shortfall risk minimization criteria (24) and (27), we first compute the optimal hedging strategy for the second criterion, (27), then using the initial value of this hedging strategy as given value for V_0, we calculate the optimal holdings for the strategy based on the first criterion, (24).

We denote by Strategy 3, the optimal strategy solving the first shortfall risk minimization problem, (24), and by Strategy 4, the optimal strategy for the second problem, (27). We remark the initial portfolio values, V_0, are the same for both strategies, however, the holdings, ξ_k, are different. For comparison with the minimum initial cost super-replicating strategy, Table 14 illustrates the values of V_0 for Strategy 4, for the same put options as in Table 13.

Table 14.
Initial portfolio cost.

Strike	# of time steps per rebalancing time			
	50	100	300	600
90	2.2065	2.0562	1.4911	1.2486
95	3.7236	3.5833	2.8130	2.3168
100	5.7089	5.5760	4.8485	3.9832
105	8.1940	8.0620	7.5616	6.3626
110	11.1355	11.0106	10.7330	9.5202

Notes. Initial portfolio cost for hedging put options with different strike prices and number of time steps per rebalancing time, for the optimal shortfall risk minimizing strategy solving (27); the same setup as in Table 1.

We remark that the initial portfolio values for Strategies 3 and 4 are much smaller than the initial values for the minimal cost super-replicating strategy and they are comparable to the initial portfolio values for the total risk minimizing strategies.

Strategies 3 and 4 have a reasonable initial cost compared to the super-replicating strategy. However, this reduction in the initial cost has been achieved by allowing a nonzero probability of a loss. While a super-replicating strategy prevents any loss, Strategies 3 and 4 have a nonzero shortfall risk. Table 15 illustrates the average values of the shortfall risk, $(H - V_M)^+$, over 40,000 scenarios for the hedging Strategies 3 and 4. We note that the shortfall risk increases as the options become more in-the-money and we rebalance less frequently. Moreover, since Strategy 3 minimizes the shortfall risk for a given initial portfolio, this strategy yields smaller values of the shortfall risk than Strategy 4, which has the same initial investment.

We can also compute the average values of the shortfall risk, $(H - V_M)^+$, over the same 40,000 paths, for the piecewise linear and quadratic total risk minimizing Strategies 1 and 2, respectively. These values will certainly be larger than the corresponding values for Strategies 3 and 4, which are shortfall risk minimizing strategies. However, the results provide interesting information about the behavior of the total risk minimizing strategies. The average shortfall risk for Strategies 1 and 2 is illustrated in Table 16.

We remark from Table 16 that the quadratic total risk minimizing Strategy 2 always yields smaller average shortfall risk than the piecewise linear risk mini-

Table 15.
Average value of the shortfall risk over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time			
		50	100	300	600
90	3	0.2826	0.4280	0.6768	0.7578
	4	0.3437	0.4735	0.6904	0.7632
95	3	0.3638	0.5571	0.9822	1.1761
	4	0.4570	0.6391	1.0279	1.1957
100	3	0.4349	0.6782	1.2353	1.6182
	4	0.5597	0.7847	1.3547	1.6781
105	3	0.4944	0.7534	1.3846	1.9844
	4	0.6498	0.9008	1.6037	2.1252
110	3	0.5288	0.8007	1.4518	2.1905
	4	0.7035	0.9693	1.7426	2.4653

Notes. Average value of the shortfall risk for hedging put options with different strike prices and number of time steps per rebalancing time, for the optimal shortfall risk minimizing strategies solving (24) and (27); the same setup as in Table 1.

Table 16.
Average value of the shortfall risk over 40,000 scenarios for the piecewise linear and quadratic total risk minimizing strategies.

Strike	Strategy	# of time steps per rebalancing time			
		50	100	300	600
90	1	0.4446	0.7533	0.9398	0.9398
	2	0.3920	0.5299	0.7868	0.8854
95	1	0.5306	0.8057	1.6648	1.6648
	2	0.4905	0.6850	1.0966	1.3111
100	1	0.6222	0.9088	1.9555	2.7269
	2	0.5868	0.8259	1.3938	1.7558
105	1	0.7043	0.9979	1.9898	3.1136
	2	0.6690	0.9420	1.6385	2.1592
110	1	0.7650	1.0773	2.0424	3.1845
	2	0.7214	1.0160	1.8026	2.4683

Notes. Average value of the shortfall risk for hedging put options with different strike prices and number of time steps per rebalancing time, for the optimal total risk minimizing Strategies 1 and 2; the same setup as in Table 1.

mizing Strategy 1. Using the relation:

$$|H - V_M| = (H - V_M)^+ + (V_M - H)^+, \tag{29}$$

we can analyze the average values of the shortfall risk, $(H - V_M)^+$, from Table 16, in comparison with the average values of the total hedging risk, $|H - V_M|$, from Table 4. While in the case of Strategy 2, the average shortfall risk is approximately half the average total risk, in the case of Strategy 1, these values are much closer, especially for out-of-money put options. By (29), it follows that Strategy 1 typically under-hedges the options, while Strategy 2 shows no trend for either under-hedging, or over-hedging.

We will now investigate the cumulative hedging cost. Table 17 illustrates the average values of the cumulative hedging cost over 40,000 scenarios for the shortfall risk minimizing Strategies 3 and 4. For comparison we include the corresponding values from Table 3 for the piecewise linear and quadratic total risk minimizing strategies satisfying (16).

As illustrated in Table 17, even if the two shortfall risk minimizing strategies start with the same investment in the hedging portfolio, Strategy 4, which has to satisfy the constraint $E(H - V_M) = 0$, yields larger values of the average cumulative hedging cost than Strategy 3. We also note that using a quadratic criterion for minimizing the risk leads to the largest hedging cost, as can be seen by comparing the cost for Strategy 2 to the cost of the other three strategies. The performance of the hedging strategies also depends on the moneyness of the options and the number of rebalancing opportunities: the piecewise linear total risk minimization has the smallest average cost when the put options are

Table 17.
Average value of the total cost over 40,000 scenarios.

Strike	Strategy	# of time steps per rebalancing time			
		50	100	300	600
95	1	3.6640	3.4080	1.6648	1.6648
	2	3.8885	3.7741	3.3983	2.9735
	3	3.2156	3.1629	2.5151	2.1555
	4	3.7236	3.5833	2.8130	2.3168
100	1	5.6896	5.5067	4.0644	2.7269
	2	5.8455	5.7119	5.2530	4.6948
	3	5.0506	5.0474	4.2289	3.6124
	4	5.7089	5.5760	4.8485	3.9832
105	1	8.1835	8.0393	7.2893	5.5301
	2	8.2882	8.1412	7.6261	6.9449
	3	7.3748	7.3405	6.5371	5.6365
	4	8.1940	8.0620	7.5616	6.3626

Notes. Average total hedging cost for put options with different strike prices and number of time steps per rebalancing time, for the total risk minimizing strategies: 1 – piecewise linear and 2 – quadratic and the shortfall risk minimizing strategies: 3 – strategy solving (24), 4 – strategy solving (27); the same setup as in Table 1.

out-of-the-money and the rebalancing in infrequent, however, as the options become in-the-money or the number of hedging opportunities increases, the shortfall risk minimization criterion (24) is the least expensive on average.

As in Section 3.1, we investigate the distributions of the shortfall risk and cumulative cost for the shortfall risk minimizing Strategies 3 and 4, in the particular case of the at-the-money put options with 6 hedging opportunities. The histograms of the shortfall risk, $(H - V_M)^+$, over the 40,000 simulated scenarios are presented in Fig. 8. We mention that the strategies have a few values of the shortfall risk outside the represented interval, however, we chose this range to make the figure clearer.

From Table 15 the average values of the shortfall risk are 0.6782 for Strategy 3 and 0.7847 for Strategy 4. The distributions of the shortfall risk for the two strategies are very similar in the chosen interval. However, Strategy 3 has a longer right tail, outside the interval. This can be seen from the values of the skewness: 6.6456 for Strategy 3, compared to 4.2451 for Strategy 4.

The histograms of the cumulative cost for the shortfall risk minimizing Strategies 3 and 4 are illustrated in Fig. 9. As before, the range of the figure has been chosen for clarity.

The average values of the cumulative cost for Strategies 3 and 4 are 5.0474 and 5.5760, respectively, as can be seen from Table 17. We remark that Strategy 3 is more asymmetric than Strategy 4. The values of the skewness, 3.2878 for Strategy 3 and 1.7171 for Strategy 4, show that Strategy 3 has also a longer right tail.

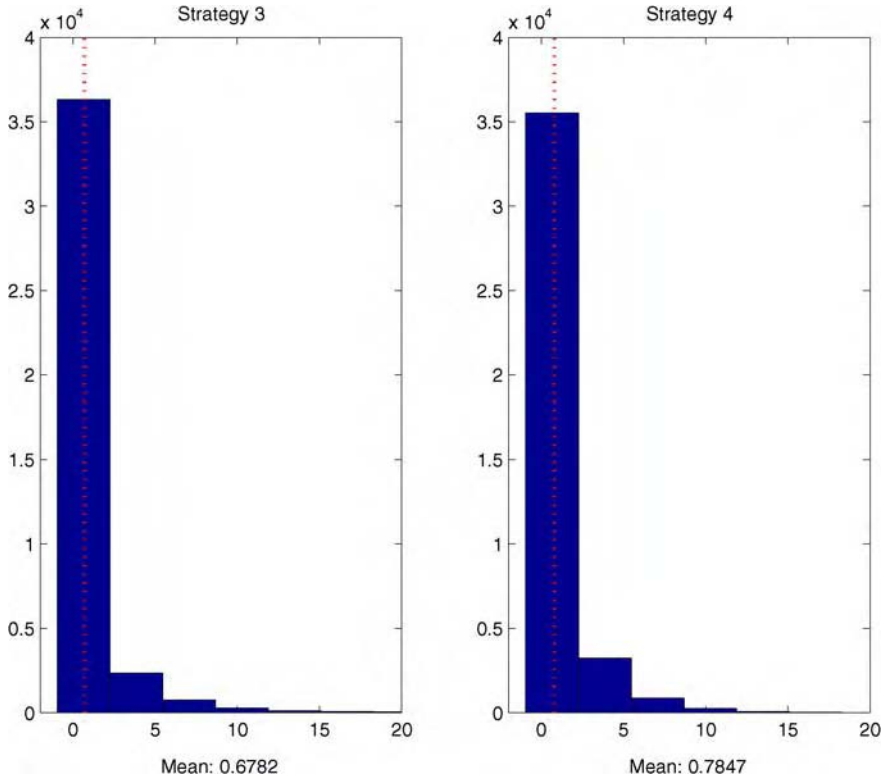


Fig. 8. Histograms of the shortfall risk over 40,000 scenarios.

As illustrated in this section, the shortfall risk minimizing strategies have attractive features, they have smaller average loss and, possibly, smaller cumulative hedging cost than the total risk minimizing strategies. However, when choosing between shortfall and total risk minimization, one has to take into account the fact that shortfall risk minimization can only be used for hedging purposes, while total risk minimization can be used for both hedging and pricing, since the initial value of a total risk minimizing strategy may be considered as a "fair value" of the option. Moreover, when hedging an option, the shortfall risk measure is appropriate if one is inclined to penalize only the costs, and not the profits of his position; if one prefers to penalize both losses and gains, he has to choose a symmetric risk measure, such as the total risk measure.

6 Conclusions

In a complete market, there exists a unique self-financing strategy that exactly replicates the option payoff. Market completeness is not, however, a realistic assumption. For example, introducing stochastic volatility or volatil-

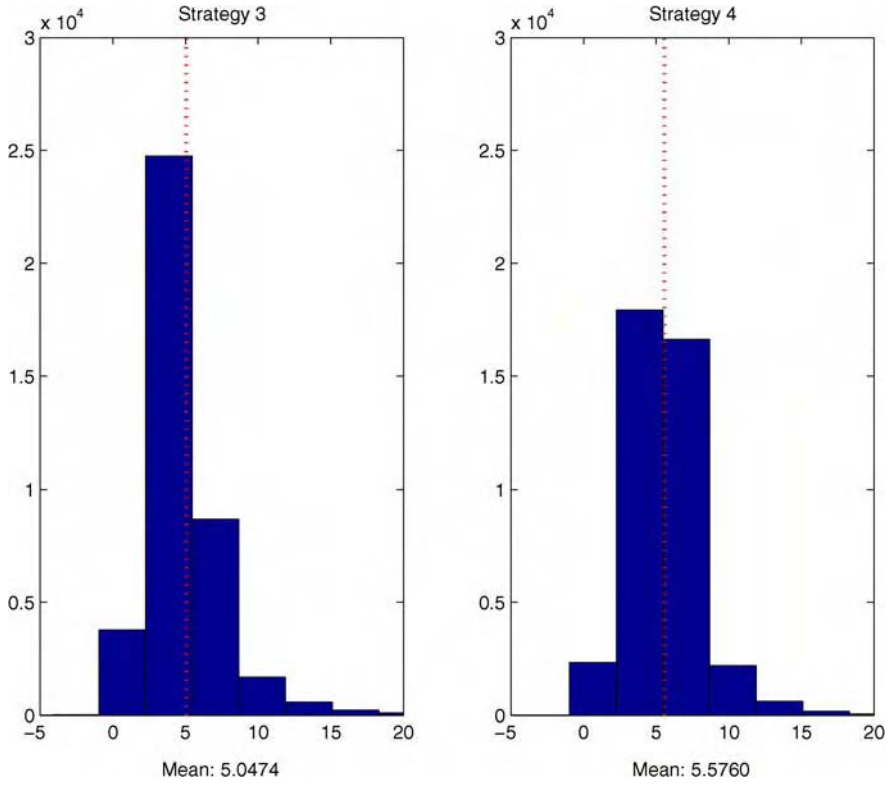


Fig. 9. Histograms of the cumulative cost over 40,000 scenarios.

ity with jumps in the Black–Scholes model in order to explain the market data, or allowing for discrete hedging, leads to an incomplete market. If the market is incomplete, the optimal hedging strategy for an option depends on the criterion for measuring the risk. The traditional strategies found in the literature are based on quadratic risk measures.

We investigate alternative piecewise linear risk minimizing criteria for total-risk minimization. Unfortunately, there are no analytic solutions to the piecewise linear risk minimization problem. Since a direct approach to this dynamic stochastic programming problem may be computationally very expensive, we obtain the optimal piecewise linear risk minimizing strategies using Monte Carlo simulations and approximating the holdings in the hedging portfolio by cubic splines. We analyze this approach in the Black–Scholes and stochastic volatility frameworks.

The numerical results illustrate that, as in the case of the local risk minimization, the piecewise linear total risk minimization criterion typically leads to smaller average hedging cost and risk. We remark that the hedging performance of the optimal strategies depends on the moneyness of the options and

on the number of rebalancing opportunities. The hedging strategies based on piecewise linear risk minimization have quite different, and often preferable, properties compared to the traditional, quadratic risk minimizing strategies. The distributions of the cumulative cost and risk show that these new strategies have a larger probability of small cost and risk, though they also have a very small probability of larger cost and risk. We also remark that in the stochastic framework analyzed in this paper, the volatility does not significantly affect the average total cost and risk of the hedging strategies.

Comparing the hedging performance of the optimal strategies for piecewise linear and quadratic total risk minimization to the performance of the shortfall risk minimizing strategies, we note that the quadratic criterion yields the largest values of the average cumulative hedging cost. Shortfall risk minimization may lead to smaller average cumulative hedging cost than piecewise linear risk minimization, depending on the moneyness of the options and the number of hedging opportunities.

By analyzing the values of the shortfall risk for the piecewise linear and quadratic total risk minimizing hedging strategies, we infer that the piecewise linear criterion typically leads to options being under-hedged, while quadratic total risk minimization shows no trend for either over-hedging, or under-hedging the options.

A shortfall risk measure may be more attractive than a total risk measure when one tries to hedge an option and he is inclined to penalize only the costs of his position. However, shortfall risk minimization cannot be used for pricing the option, while total risk minimization can be used for both hedging and pricing. Moreover, when one prefers to penalize both losses and gains, a shortfall risk measure is no longer appropriate.

References

Bertsimas, D., Kogan, L., Lo, A. (2001). Hedging derivative securities and incomplete markets: An ϵ-arbitrage approach. *Operations Research* 49, 372–397.

Coleman, T.F., Li, Y., Patron, M. (2003). Discrete hedging under piecewise-linear risk-minimization. *Journal of Risk* 5 (3).

Föllmer, H., Leukert, P. (2000). Efficient hedging: Cost versus shortfall risk. *Finance and Stochastics* 4, 117–146.

Föllmer, H., Schweizer, M. (1989). Hedging by sequential regression: An introduction to the mathematics of option trading. *ASTIN Bulletin* 1, 147–160.

Frey, R. (1997). Derivative asset analysis in models with level-dependent and stochastic volatility. *CWI Quarterly* 10 (1), 1–34.

Heath, D., Platen, E., Schweizer, M. (2001a). A comparison of two quadratic approaches to hedging in incomplete markets. *Mathematical Finance* 11, 385–413.

Heath, D., Platen, E., Schweizer, M. (2001b). Numerical comparison of local risk-minimisation and mean–variance hedging. In: Jouini, E., Cvitanic, J., Musiela, M. (Eds.), *Option Pricing, Interest Rates and Risk Management*. Cambridge Univ. Press, pp. 509–537.

Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6, 327–343.

Hull, J.C., White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42, 281–300.

El Karoui, N., 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.

Longstaff, F., Schwartz, E.S. (2001). Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies* 14 (1), 113–147.

Mercurio, F., Vorst, T.C.F. (1996). Option pricing with hedging at fixed trading dates. *Applied Mathematical Science* 3, 135–158.

Runggaldier, W.J. (2001). Adaptive and robust control procedures for risk minimization under uncertainty. In: Menaldi, J.L., Rofman, E., Sulem, A. (Eds.), *Optimal Control and Partial Differential Equations, Volume in Honour of Prof. Alain Bensoussan's 60th Birthday*. IOS Press, pp. 549–557.

Schäl, M. (1994). On quadratic cost criteria for option hedging. *Mathematics of Operation Research* 19 (1), 121–131.

Schweizer, M. (1995). Variance-optimal hedging in discrete time. *Mathematics of Operation Research* 20, 1–32.

Schweizer, M. (2001). A guided tour through quadratic hedging approaches. In: Jouini, E., Cvitanic, J., Musiela, M. (Eds.), *Option pricing, interest rates and risk management*. Cambridge Univ. Press, pp. 538–574.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15015-5

Chapter 15

Queuing Theoretic Approaches to Financial Price Fluctuations*

Erhan Bayraktar

Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA
E-mail: erhan@umich.edu

Ulrich Horst

Department of Mathematics, The University of British Columbia, 1984 Mathematics Road, Vancouver, BC V6T 1Z2, Canada
E-mail: horst@math.ubc.ca

Ronnie Sircar

Operations Research & Financial Engineering Department, Princeton University, E-Quad, Princeton, NJ 08544, USA
E-mail: sircar@princeton.edu

Abstract

One approach to the analysis of stochastic fluctuations in market prices is to model characteristics of investor behavior and the complex interactions between market participants, with the aim of extracting consequences in the aggregate. This agent-based viewpoint in finance goes back at least to the work of Garman [Garman, M. (1976). Market microstructure. Journal of Financial Economics 3, 257–275] and shares the philosophy of statistical mechanics in the physical sciences. We discuss recent developments in market microstructure models. They are capable, often through numerical simulations, to explain many stylized facts like the emergence of herding behavior, volatility clustering and fat tailed returns distributions. They are typically queuing-type models, that is, models of order flows, in contrast to classical economic equilibrium theories of utility-maximizing, rational, "representative" investors. Mathematically, they are analyzed using tools of functional central limit theorems, strong approximations and weak convergence. Our main examples focus on investor inertia, a trait that is well-documented, among other behavioral qualities, and modeled using semi-Markov switching processes. In particular, we show how inertia may lead to the phenomenon of long-range dependence in stock prices.

* E. Bayraktar's work supported in part by NSF Research Grant, DMS-0604491. R. Sircar's work supported in part by NSF Research Grant, DMS-0306357.

1 Introduction

Modeling market microstructure in order to understand the effects of many individual investors on aggregate demand and price formation is both a classical area of study in economics, and a rapidly growing activity among researchers from a variety of disciplines, partly due to modern-day computational power for large-scale simulations, and the increased availability of price and order-book data. Among the benefits of this type of analysis, whether mathematical or simulation-based, is the design of better models of macroscopic financial variables such as prices, informed by microscopic (investor-level) features, that can then be utilized for improved forecasts, investment and policy decisions.

The approach we discuss here is to identify characteristics common to large groups of investors, for example prolonged inactivity or inertia, and study the resulting price dynamics created by order flows. Typically, we are interested in understanding the microstructure effects on the aggregate quantity through approximations from stochastic process limit theorems when there is a large number of investors.

In this point of view, we model right away the behavior of individual traders rather than characterizing agents' investment decisions as solutions to individual utility maximization problems. Such an approach has also been taken in Garman (1976), Föllmer and Schweizer (1993), Lux (1998) and Föllmer et al. (2005), for example. As pointed out by O'Hara in her influential book Market Microstructure Theory (O'Hara, 1995), it was Garman's 1976 paper (Garman, 1976) that "inaugurated the explicit study of market microstructure." There, he explains the philosophy of this approach as follows: "The usual development here would be to start with a theory of individual choice. Such a theory would probably include the assumption of a stochastic income stream [and] probabilistic budget constraints ... But here we are concerned rather with aggregate market behavior and shall adopt the attitude of the physicist who cares not whether his individual particles possess rationality, free will, blind ignorance or whatever, as long as his statistical mechanics will accurately describe the behavior of large ensembles of those particles." This approach is also common in much of the econophysics literature (see the discussion in Farmer et al., 2005, for example), and is of course prevalent in queuing models of telephone calls or Internet traffic (Chen and Yao, 2001), where interest is not so much on causes of phone calls or bandwidth demand, but on phenomenological models and their overall implications. As one econophysicist explained it in reaction to the usual battle-cry of the classical economist about rational behavior, when AT&T uses queuing models, it doesn't ask *why* you call your grandmother.

In this article, we provide an outline to recent surveys on agent-based computational models and analytical models based on dynamical systems, while our focus is on developing limit theorems for queuing models of investor behavior, which apply modern methods from stochastic analysis to models based on economic intuition and empirical evidence. The goal is in obtaining insights

into market dynamics by understanding price formation from typical behavioral qualities of individual investors.

The remainder of this paper is summarized as follows: in Section 2, we briefly survey some recent research on agent based models. These models relate the behavioral qualities of investors and quantitative features of the stock price process. We give a relevant literature review of Queuing Theory approaches to the modeling of stock price dynamics in Section 2.3. In Section 2.4, we discuss evidence of investor inertia in financial markets, and we study its effect on stock price dynamics in Section 3. Key tools are a functional central limit theorem for semi-Markov processes and approximation results for integrals with respect to fractional Brownian motion, that establish a link between investor inertia and long range dependence in stock price returns. These are extended in Section 3.2 to allow for the *feedback* of the stock price into agents' investment decisions, using methods and techniques from *state dependent queuing networks*. We establish approximation results for the stock price in a non-Walrasian framework in which the order rates of the agents depend on the stock price and exogenously specified investor sentiment. Section 4 concludes and discusses future directions.

2 Agent-based models of financial markets

In mathematical finance, the dynamics of asset prices are usually modeled by trajectories of some exogenously specified stochastic process defined on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Geometric Brownian motion has long become the canonical reference model of financial price dynamics. Since prices are generated by the demand of market participants, it is of interest to support such an approach by a microeconomic model of interacting agents. In recent years there has been increasing interest in agent-based models of financial markets where the demand for a risky asset comes from many agents with interacting preferences and expectations. These models are capable of reproducing, often through simulations, many "stylized facts" like the emergence of herding behavior (Föllmer et al., 2005; Lux, 1998); volatility clustering (Cont, 2004; Lux and Marchesi, 2000), or fat-tailed distributions of stock returns (Cont and Bouchaud, 2000), that are observed in financial data.

In contrast to the traditional framework of an economy with a utility-maximizing representative agent, agent-based models typically comprise many *heterogeneous* traders who are so-called *boundedly rational*. Behavioral finance models assume that market participants do not necessarily share identical expectations about the future evolution of asset prices or assessments about a stock's fundamental value. Instead, agents are allowed to use rule of thumb strategies when making their investment decisions and to switch randomly between them as time passes. Following the seminal article of Frankel and Froot

(1986), one typically distinguishes *fundamentalists*, *noise traders* and *chartists*.[1] A fundamentalist bases his forecasts of future asset prices upon market fundamentals and economic factors such as dividends, quarterly earnings or GDP growth rates. He invests in assets he considers undervalued, that is, he invests in assets whose price is beneath his subjective assessment of the fundamental value. Chartists, on the other hand, base their trading strategy upon observed historical price patterns such as trends. Technical traders try to extrapolate future asset price movements from past observations. Fundamentalists and chartists typically coexist with fractions varying over time as agents are allowed to change their strategies in reaction to either the strategies' performances or the choices of other market participants. Some of these changes can be self reinforcing when agents tend to follow the more successful strategies or agents. This may lead to temporary deviations of prices from the benchmark fundamental or rational expectations prices generating bubbles or crashes in periods when technical trading predominates. Fundamentalists typically have a stabilizing impact on stock prices.

In this section, we review some agent-based models of financial markets. Our focus will be on a class probabilistic models in which asset price dynamics are modeled as stochastic processes in a random environment of investor sentiment. These models are perhaps most amenable to rigorous mathematical results. Behavioral finance models based on deterministic dynamical systems are covered only briefly as they are discussed extensively in a recent survey by Hommes (2006). For results on evolutionary dynamics of financial markets we refer to Hens and Schenk-Hoppé (2005), Estigneev et al. (2006), or Sandroni (2000) and references therein.

2.1 Stock prices as temporary equilibria in random media

Föllmer and Schweizer (1993) argue that asset prices should be viewed as a sequence of temporary equilibrium prices in a random environment of investor sentiment; see also (Föllmer, 1994). In reaction to a proposed price p in period t, agent $a \in \mathbb{A}$ forms a random excess demand $e_t^a(p, \omega)$, and the actual asset price $P_t(\omega)$ is determined by the market clearing condition of zero total excess demand. In Föllmer and Schweizer (1993), individual excess demand involves some *exogenous* liquidity demand and an *endogenous* amount obtained by comparing the proposed price p with some reference level $\hat{S}_t^a$. This dependence is linear on a logarithmic scale and individual excess demand takes the form

$$e_t^a(p, \omega) := c_t^a(\omega)\big(\log \hat{S}_t^a(\omega) - \log p\big) + \eta_t^a(\omega) \tag{2.1}$$

with nonnegative random coefficients $c_t^a(\omega)$. Here $\eta_t^a(w)$ is the individual's liquidity demand. The logarithmic equilibrium price $S_t(\omega) := \log P_t(\omega)$ is then

[1] Survey data showing the importance of chartist trading rules among financial practitioners can be found in, e.g., Taylor and Allen (1992) and Frankel and Froot (1987).

determined via the market clearing condition $\sum_{a\in\mathbb{A}} e^a(P_t(\omega),\omega) = 0$. It is thus formed from an aggregate of individual price assessments and liquidity demands. If the agents have no sense of the direction of the market and simply take the last logarithmic price S_{t-1} as their reference level, i.e., if $\log \hat{S}_t^a = S_{t-1}$, then the log-price dynamics reduces to an equation of the form

$$S_t = S_{t-1} + \eta_t$$

were η_t denotes the aggregate liquidity demand. In this case the dynamics of logarithmic prices reduces to a simple random walk model if the aggregate liquidity demand is independent and identically distributed over time. This is just the discretized version of the Black–Scholes–Samuelson geometric Brownian motion model.

A *fundamentalists* bases his investment decision on the idea that asset prices will move closer to his subjective benchmark fundamental value F^a. In a simple log-linear case

$$\log \hat{S}_t^a := S_{t-1} + \alpha_t^a\big(F^a - S_{t-1}\big) \tag{2.2}$$

for some random coefficient $0 < \alpha_t^a < 1$. If only such *information traders* are active on the market, the resulting logarithmic stock price process takes the form of a mean-reverting random walk in the random environment $\{\alpha_t\}_{t\in\mathbb{N}}$ ($\alpha_t = \{\alpha_t^a\}_{a\in\mathbb{A}}$). A combination of information trading and a simple form of *noise trading* where some agents take the proposed price seriously as a signal about the underlying fundamental value replacing F^a in (2.2) by p leads to a class of discrete time Ornstein–Uhlenbeck processes. Assuming for simplicity that subjective fundamentals equal zero the logarithmic price process takes the form

$$S_t - S_{t-1} = \tilde{\gamma}_t S_{t-1} + \gamma_t \tag{2.3}$$

with random coefficients $\tilde{\gamma}_t$ and γ_t. These coefficients describe the fluctuations in the proportion between fundamentalist and noise traders. When noise trading predominates, $\tilde{\gamma}_t$ becomes negative and the price process transient. Asset prices behave in a stable manner when the majority of the agents adopts a fundamentalist benchmark.

2.1.1 Random environment driven by interactive Markov processes

Let us now discuss a possible source of randomness driving the environment for the evolution of stock prices. Extending an earlier approach in Föllmer (1994), Horst (2005) analyzes a situation with countably many agents located on some integer lattice $\mathbb{A}$ where the environment is generated by an underlying Markov chain with an interactive dynamics. There is a set C of possible characteristics or trading strategies. An agent's state $x_t^a \in C$ specifies her reference level for the following period. The environment is then driven by a Markov chain

$$\Pi(x_t;\cdot) = \prod_{a\in\mathbb{A}} \pi_a(x_t;\cdot) \tag{2.4}$$

where $x_t = (x_t^a)_{a\in\mathbb{A}}$ denotes the current configuration of reference levels. The distribution $\pi_a(x_t;\cdot)$ of an agent's state in the following period may depend both on the current states of some "neighbors" and signals about the aggregate behavior. Information about aggregate behavior is carried in the empirical distribution $\varrho(x_t)$ or, more generally, the empirical field $R(x_t)$ associated to the configuration x_t. The empirical field is defined as the weak limit

$$R(x_t) := \lim_{n\to\infty} \frac{1}{|\mathbb{A}_n|} \sum_{a\in\mathbb{A}_n} \delta_{\theta^a x_t}(\cdot)$$

along an increasing sequence of finite sub-populations $\mathbb{A}_n \uparrow \mathbb{A}$ and $(\theta^a)_{a\in\mathbb{A}}$ denotes the canonical shift group on the space of all configurations. Due to the dependence of the transition probabilities $\pi_a(x;\cdot)$ on aggregate behavior, the kernel Π does not have the Feller property, and so standard convergence results for Feller processes on compact state spaces do not apply. As shown by Föllmer and Horst (2001) and Horst (2002) the evolution of aggregate behavior on the level of empirical fields can be described by a Markov chain. In Horst (2005), it is the $\{R(x_t)\}_{t\in\mathbb{N}}$ process that generates the environment:

$$(\tilde{\gamma}_t, \gamma_t) \sim Z\big(R(x_t);\cdot\big) \quad \text{for a suitable stochastic kernel } Z.$$

Under a weak interaction condition the process $\{R(x_t)\}_{t\in\mathbb{N}}$ settles down to a unique limiting distribution. Hence asset prices asymptotically evolve in a stationary and ergodic random environment. This allows us to approximate the discrete time process $\{S_t\}_{t\in\mathbb{N}}$ by the unique strong solution to the stochastic differential equation

$$\mathrm{d}Z_t = Z_t\,\mathrm{d}X_t + \mathrm{d}\tilde{X}_t$$

where X and $\tilde{X}$ are Brownian motions with drift and volatility; see Föllmer and Schweizer (1993) or Horst (2005) for details.

2.1.2 *Feedback effects*

The random environment in (Horst, 2005) is generated by a Markov process describing the evolution of individual behavior. While this approach is capable of capturing some interaction and imitation effects such as word-of-mouth advertising unrelated to market events, the dynamics of $\{x_t\}_{t\in\mathbb{N}}$ lacks a dependence on asset price dynamics. The model by Föllmer et al. (2005) captures feedback effects from stock prices into the environment. At the same time it allows for trend chasing. A trend chaser or chartist bases his expectation of future asset prices and hence his trading strategy upon observed historical price patterns such as trends. In Föllmer et al. (2005), for instance, the chartist's benchmark level takes the form

$$\log \hat{S}_t^a := S_{t-1} + \beta_t^a (S_{t-1} - S_{t-2}). \tag{2.5}$$

A combination of the trading strategies (2.2) and (2.5) yields a class of asset price processes that can be described by a higher order stochastic difference

equation. In Föllmer et al. (2005) the agents use one of a number of predictors which they obtain from financial "gurus" to forecast future price movements. The agents evaluate the gurus' performance over time. Performances are measured by weighted sums of past profits the strategies generate. The probability of choosing a given guru is related to the guru's success. As a result, the configuration x_t of individual choices at time t is a random variable whose distribution depends on the current vector of performance levels U_{t-1}. This dependence of the agents' choices on performances introduces a feedback from past prices into the random environment. Loosely speaking one obtains a difference equation of the form (2.3) where

$$(\tilde{\gamma}_t, \gamma_t) \sim Z(U_t; \cdot) \quad \text{for a suitable stochastic kernel } Z.$$

While prices can temporarily deviate from fundamental values, the main result in Föllmer et al. (2005) shows that the price process has a unique stationary distribution, and time averages converge to their expected value under the stationary measure if the impact of trend chasing is weak enough.

2.1.3 *Multiplicity of equilibria*

As argued by Kirman (1992), in a random economy with many heterogeneous agents, a natural idea of an equilibrium is not a particular state, but rather a distribution of states reflecting the proportion of time the economy spends in each of the states. In the context of microstructure models where liquidity trading or interaction effects prevent asset prices from converging pathwise to some steady state, stationary distributions for asset prices are thus a natural notion of equilibrium. In this sense, the main result in (Föllmer et al., 2005) may be viewed as an existence and uniqueness result for equilibria in financial markets with heterogeneous agents. Horst and Wenzelburger (2005) study a related model with many small investors where performances are evaluated according historic returns or Sharpe ratios. In the limit of an infinite set of agents the dynamics of asset prices can be described by a path dependent linear stochastic difference equation of the form

$$Y_t = A(\varrho_{t-1})Y_{t-1} + B(\varrho_{t-1}, \epsilon_t).$$

Here $\{\epsilon_t\}_{t\in\mathbb{N}}$ is an exogenous i.i.d sequence of noise trader demand and ϱ_{t-1} denotes the empirical distribution of the random vector $Y_0, Y_1, \ldots, Y_{t-1}$. While the models shares many of the qualitative features of Horst (2005) and Föllmer et al. (2005), it allows for multiple limiting distributions of asset prices. If the interaction between different agents is strong enough, asset prices converge in distribution to a *random* limiting measure. Randomness in the limiting distribution may be viewed as a form of market incompleteness generated by contagious interaction effects.

2.1.4 *Interacting agent models in an overlapping generations framework*

The work in Horst and Wenzelburger (2005) is based on earlier work by Böhm et al. (2000), Böhm and Wenzelburger (2005), and Wenzelburger (2004).

These authors developed a dynamic analysis of endogenous asset price formations in the context of overlapping generations economies where agents live for two periods and the demand for the risky asset comes from young households. They investigate the impact of different forecasting rules on both asset price and wealth dynamics under the assumption that agents are myopic and therefore boundedly rational, mean–variance maximizers. Böhm et al. (2000) study asset prices and equity premia for a parameterized class of examples and investigate the role of risk aversion and of subjective as well as rational beliefs. It is argued that realistic parameter values explain Mehra and Prescott's equity premium puzzle (Mehra and Prescott, 1985). The model is generalized in Wenzelburger (2004) to a model with an arbitrary number of risky assets and heterogeneous beliefs, thus generalizing the classical CAPM. A major result is conditions under which a learning scheme converges to rational expectations for one investor while other investors have non-rational beliefs. A second major result is the notion of a *modified market portfolio* along with a generalization of the security market line result stating that in a world of heterogeneous myopic investors, modified market portfolios are *mean–variance efficient in the classical sense of CAPM*, regardless of the diversity of beliefs of other agents. See Böhm and Chiarella (2005) for a related approach.

2.1.5 Feedback effects from program trading, large agents and illiquidity

A different type of feedback effect, from the actions of a large group of *program traders* or large influential agents has been modeled in the financial mathematics literature. In the 1990s, following the Brady report that attributed part of the cause of the 1987 stock market crash to program trading by institutions following portfolio insurance strategies, researchers analyzed the feedback effect from option Delta-hedging by a significant fraction of market participants on the price dynamics of the underlying security. See, for example, Frey and Stremme (1997), Sircar and Papanicolaou (1998), Schönbucher and Wilmott (2000) and Platen and Schweizer (1998).

Related analyses can be found in models where there is a large investor whose actions move the price, for example Jonsson and Keppo (2002), and where there is a market depth function describing the impact of order size on price, for example Cetin et al. (2004). A cautionary note on all such models is that, under sensible conditions, they do not explain the implied volatility smile/skew that is observed in modern options markets (in fact they predict a reverse smile). This would suggest that program trading, large agent or illiquidity effects are second order phenomena as far as derivatives markets are concerned, compared with the impacts of jumps or stochastic volatility.

There has also been some recent empirical work on estimating the market depth function, in particular the tail of the distribution governing how order size impacts trading price: see Farmer and Lillo (2004) and Gabaix et al. (2003).

2.2 Stock prices and random dynamical systems

An important branch of the literature on agent-based financial market models analyzes financial markets in which the dynamics of asset prices can be described by a deterministic dynamical system. The idea is to view agent-based models as highly nonlinear deterministic dynamical systems and markets as *complex adaptive systems*, with the evolution of expectations and trading strategies coupled to market dynamics. Many such models, when simulated, generate time paths of prices which switch from one expectations regime to another generating *rational routes to randomness*, i.e., chaotic price fluctuations. As these models are considerably more complex than the ones reviewed in the previous section, analytical characterizations of asset price processes are typically not available. However, when simulated, these model generate much more realistic time paths of prices explaining many of the stylized facts observed in real financial markets.

Particularly relevant contributions include the early work of Day and Huang (1990), Frankel and Froot (1986) and the work of Brock and Hommes (1997). The latter studies a model in which boundedly rational agents can use one of two forecasting rules or investment strategies. One of them is costly but when all agents use it, the emerging price process is stable. The other is cheaper but when used by many individuals induces unstable behavior of the price process. Their model has periods of stability interspersed with bubble like behavior. In Brock and Hommes (1998) the same authors introduced the notion of *Adaptive Belief Systems* (ABS), a "financial market application of the evolutionary selection of expectation rules" analyzed in Brock and Hommes (1997). An ABS may be viewed as asset pricing models derived form mean-variance optimization with heterogeneous beliefs. As pointed out in Hommes (2006), "a convenient feature of an ABS is that it can be formulated in terms of (price) deviations from a benchmark fundamental and (...) can therefore be used in experimental and empirical testing of deviations from the (rational expectations) benchmark." Recently, several modifications of ABSs have been studied. While in Brock and Hommes (1998) the demand for a risky asset comes from agents with constant absolute risk aversion utility functions and the number of trader types is small, Chiarella and He (2001) and Brock et al. (2005) developed models of interaction of portfolio decisions and wealth dynamics with heterogeneous agents whose preferences are described by logarithmic CRRA utility functions and many types of traders, respectively. Gaunersdorfer (2000) extends the work in Brock and Hommes (1997) to the case of time-varying expectations about variances of conditional stock returns.[2]

[2] There are many other papers utilizing dynamical system theory to analyze asset price dynamics in behavioral finance models. For a detailed survey, we refer the interested reader to Hommes (2006).

2.3 Queuing models and order book dynamics

The aforementioned models differ considerably in their degree of complexity and analytical tractability, but they are all based on the idea that asset price fluctuations can be described by a sequence of temporary price equilibria. All agents submit their demand schedule to a market maker who matches individual demands in such a way that markets clear. While such an approach is consistent with dynamic microeconomic theory, it should only be viewed as a first steps towards a more realistic modeling of asset price formation in large financial markets. In real markets, buying and selling orders arrive at different points in time, and so the economic paradigm that a Walrasian auctioneer can set prices such that the markets clear at the end of each trading period typically does not apply. In fact, almost all automated financial trading systems function as continuous double auctions. They are based on electronic *order books* in which all unexecuted limit orders are stored and displayed while awaiting execution. While analytically tractable models of order book dynamics would be of considerable value, their development has been hindered by the inherent complexity of limit order markets. So far, rigorous mathematical results have only been established under rather restrictive assumptions on aggregate order flows by, e.g., Mendelson (1982), Luckock (2003) and Kruk (2003). Statistical properties of continuous double auctions are often analyzed in the econophysics literature, e.g., Smith et al. (2003) and references therein.

Microstructure models with asynchronous order arrivals where orders are executed immediately rather than awaiting the arrival of a matching order and where asset prices move into the order to market imbalance are studied by, e.g. Garman (1976); Lux (1995, 1998, 1997) or Bayraktar et al. (2006). These models may be viewed as an intermediate step towards a more realistic modeling of electronic trading systems.

A convenient mathematical framework for such models, which we will develop in detail in Section 3.2, is based on the theory of state-dependent queuing networks (see Mandelbaum et al., 1998 or Mandelbaum and Pats, 1998 for detailed discussions of Markovian queuing networks). Underlying this approach is the idea that the dynamics of order arrivals follows a Poisson-type process with price dependent rates and that a buying (selling) order increases (decreases) the stock price by a fixed amount (on a possibly logarithmic scale to avoid negative prices).

More precisely, the arrival times of aggregate buying and selling orders are specified by independent Poisson processes Π_+ and Π_- with price and time dependent rates λ_+ and λ_-, respectively, that may also depend on investor characteristics or random economic fundamentals. In the simplest case the logarithmic price process $\{S_t\}_{t\geqslant 0}$ takes the form

$$S_t = S_0 + \Pi_+\left(\int_0^t \lambda_+(S_u, u)\,\mathrm{d}u\right) - \Pi_-\left(\int_0^t \lambda_-(S_u, u)\,\mathrm{d}u\right).$$

The excess order rate $\lambda_+(S_u, u) - \lambda_-(S_u, u)$ may be viewed as a measure of aggregate excess demand while $\Pi_+(\int_0^t \lambda_+(S_u, u)\,\mathrm{d}u) - \Pi_-(\int_0^t \lambda_-(S_u, u)\,\mathrm{d}u)$ denotes the accumulated net order flow up to time t. In a model with many agents and after suitable rescaling the asset price process may be approximated by a deterministic process while the fluctuations around this first-order approximation can typically be described by an Ornstein–Uhlenbeck diffusion.

Recently, such queuing models have also been applied to modeling the credit risk of large portfolios by Davis and Esparragoza (2004). They approximate evolution of the loss distribution of a large portfolio of credit instruments over time. We further elaborate on queuing theoretic approaches to stock price dynamics in Section 3. Before that, we introduce a common investor trait, investor inertia, and show the effects of this common trait on stock prices.

2.4 Inertia in financial markets

The models mentioned previously assume that agents trade the asset in each period. At the end of each trading interval, agents update their expectations for the future evolution of the stock price and formulate their excess demand for the following period. However, small investors are not so efficient in their investment decisions: they are typically inactive and actually trade only occasionally. This may be because they are waiting to accumulate sufficient capital to make further stock purchases; or they tend to monitor their portfolios infrequently; or they are simply scared of choosing the wrong investments; or they feel that as long-term investors, they can defer action; or they put off the time-consuming research necessary to make informed portfolio choices. Long uninterrupted periods of inactivity may be viewed as a form of investor inertia.

2.4.1 Evidence of inertia

Investor inertia is a common experience and is well documented. The New York Stock Exchange (NYSE)'s survey of individual shareownership in the United States, "Shareownership2000" (Grasso, 2000), demonstrates that many investors have very low levels of trading activity. For example they find that "23 percent of stockholders with brokerage accounts report no trading at all, while 35 percent report trading only once or twice in the last year." The NYSE survey also reports (Table 28) that the average holding period for stocks is long, for example 2.9 years in the early 1990s. Empirical evidence of inertia also appears in the economic literature. For example, Madrian and Shea (2001) looked at the reallocation of assets in employees' individual 401(k) (retirement) plans and found "a status quo bias resulting from employee procrastination in making or implementing an optimal savings decision." A related study by Hewitt Associates (a management consulting firm) found that in 2001, four out of five plan participants did not do any trading in their 401(k)s. Madrian and Shea explain that "if the cost of gathering and evaluating the information needed to make a 401(k) savings decision exceeds the short-run benefit from doing so, individuals will procrastinate." The prediction of Prospect Theory (see Kahneman

and Tversky, 1979) that investors tend to hold onto losing stocks too long has also been observed in Shefrin and Statman (1985). Another typical cause is that small investors seem to find it difficult to reverse investment decisions, as is discussed even in the popular press. A recent newspaper column (by Russ Wiles in the Arizona Republic, November 30, 2003) states: "Perhaps more than anything, investor inertia is a key force (in financial markets). When the news turns sour, people tend to hold off on buying rather than bail out. In 2002, the toughest market climate in a generation and a year with ample Wall Street scandals, equity funds suffered cash outflows of just one percent."

2.4.2 Inertia and long range dependencies in financial time series

One of the outcomes of a limit analysis of an agent-based model of investor inertia is a stock price process based on fractional Brownian motion, which exhibits long-range dependence (that is correlation or memory in returns). This is discussed in Section 3.1. In particular, the limit fluctuation process is a *fractional Brownian motion*.

We recall that fractional Brownian motion B^H with *Hurst* parameter $H \in (0, 1]$ is an almost surely continuous and centered Gaussian process with auto-correlation

$$\mathbb{E}\{B_t^H B_s^H\} = \frac{1}{2}\big(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\big). \tag{2.6}$$

Remark 2.1. Note that the case $H = \frac{1}{2}$ gives standard Brownian motion. Also note that the auto-correlation function is positive definite if and only if $H \in (0, 1]$.

Bayraktar et al. (2004) studied an asymptotically efficient wavelet-based estimator for the Hurst parameter, and analyzed high frequency S&P 500 index data over the span of 11.5 years (1989–2000). It was observed that, although the Hurst parameter was significantly higher than the efficient markets value of $H = \frac{1}{2}$ up through the mid-1990s, it started to fall to that level over the period 1997–2000 (see Fig. 1). This might be explained by the increase in Internet trading in that period, which is documented, for example, in NYSE's "Shareownership2000" (Grasso, 2000; Barber and Odean, 2001; and Choi et al., 2002), in which it is demonstrated that "after 18 months of access, the Web effect is very large: trading frequency doubles." Indeed, as reported in Barber and Odean (2002), "after going online, investors trade more actively, more speculatively and less profitably than before." Similar empirical findings to that of Bayraktar et al. (2004) were recently reached, using a completely different statistical technique by Bianchi (2005).

Thus, the dramatic fall in the estimated Hurst parameter in the late 1990s can be thought of as *a posteriori* validation of the link the limit theorem in Bayraktar et al. (2006) provides between investor inertia and long-range dependence in stock prices. We review this model in Section 3.1. An extension

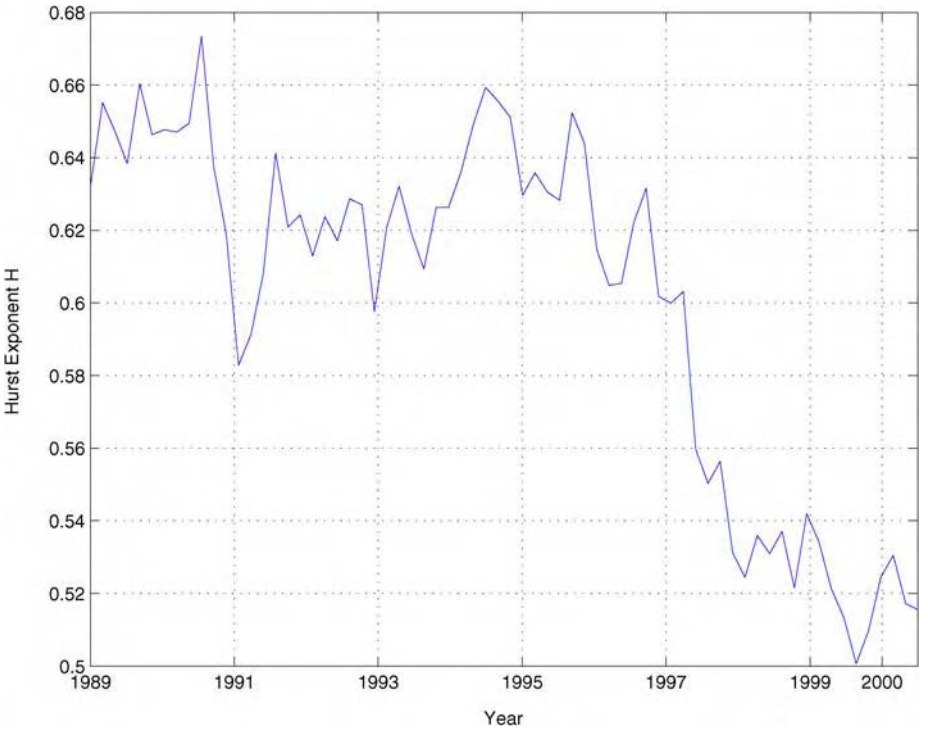


Fig. 1. Estimates of the Hurst exponent of the S&P 500 index over 1990s, taken from Bayraktar et al. (2004).

based on state dependent queuing networks with semi-Markov switching is discussed in Section 3.2.

3 Microstructure models with inert investors

We illustrate the use of microstrucure, or agent-based models, combined with limit theorems by focusing on investor inertia as a very common characteristic among small and casual market participants. In Section 3.1 we summarize earlier work (Bayraktar et al., 2006) that established a mathematical link between inertia, long-range dependence in stock returns and potential short-lived arbitrage opportunities for other 'sophisticated' parties. Section 3.2 contains an extension allowing for feedback effects from current prices into the agents' order rates.

3.1 A microstructure model without feedback

We now introduce the basic concepts and notation of the market microstructure model analyzed in Bayraktar et al. (2006) that will serve as basis for the

more sophisticated model in Section 3.2. We start with a financial market with a set $\mathbb{A} := \{a_1, a_2, \ldots, a_N\}$ of *agents* trading a single risky asset. Each agent $a \in \mathbb{A}$ is associated with a continuous-time stochastic process $x^a = \{x_t^a\}_{t \geqslant 0}$ on a finite state space E describing his *trading activity*.

We take a pragmatic approach to specify the demand. Instead of formulating an individual optimization problem under budget constraints for the agents, we start right away with the agent's order rates. The agent $a \in \mathbb{A}$ accumulates the asset at a rate $\Psi_t x_t^a$ at time $t \geqslant 0$. Here x_t^a may be negative indicating that the agent is selling. The random process $\Psi = \{\Psi_t\}_{t \geqslant 0}$ describes the evolution of the size of a typical trade. It can also be interpreted as a stochastic elasticity coefficient (the reaction of the price to the market imbalance). We assume that Ψ is a continuous non-negative semi-martingale which is independent of the processes x^a and that $0 \in E$. The agents do not trade at times when $x_t^a = 0$. The holdings of the agent $a \in \mathbb{A}$ and the "market imbalance" at time $t \geqslant 0$ are thus given by, respectively,

$$\int_0^t \Psi_s x_s^a \, \mathrm{d}s \quad \text{and} \quad \sum_{a \in \mathbb{A}} \int_0^t \Psi_s x_s^a \, \mathrm{d}s. \tag{3.1}$$

Remark 3.1. In our continuous time model, buyers and sellers arrive at different points in time. Hence the economic paradigm that a Walrasian auctioneer can set prices such that the markets clear at the end of each trading period does not apply. Rather, temporary imbalances between demand and supply will occur. Prices will reflect the extent of market imbalance.

All the orders are received by a single market maker. The market maker clears all trades and prices in reaction to the evolution of market imbalances, the only component driving asset prices. Reflecting the idea that an individual agent has diminishing impact on market dynamics if the number of traders is large, we assume that the impact of an individual order is inversely proportional to the number of possible traders: a buying (selling) order increases (decreases) the logarithmic stock price by $1/N$. The pricing rule for the evolution of the logarithmic stock price process $S^N = \{S_t^N\}_{t \geqslant 0}$ is linear and taken to be:

$$\mathrm{d}S_t^N = \frac{1}{N} \sum_{a \in \mathbb{A}} \Psi_t x_t^a \, \mathrm{d}t. \tag{3.2}$$

In order to incorporate the idea of market inertia, the agents' trading activity is modeled by independent and identically distributed *semi-Markov* processes x^a. Semi-Markov processes are tailor-made to model individual traders' inertia as they generalize Markov processes by removing the requirement of exponentially distributed, and therefore thin-tailed, holding (or sojourn) times. Since the processes x^a are independent and identically distrib-

uted, it is enough to specify the dynamics of some "representative" process $x = \{x_t\}_{t \geqslant 0}$.

3.1.1 Semi-Markov processes

A semi-Markov process x defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is specified in terms of random variables $\xi_n : \Omega \to E$ and $T_n : \Omega \to \mathbb{R}_+$, satisfying $0 = T_1 \leqslant T_1 \leqslant \cdots$ almost surely and

$$\begin{aligned}&\mathbb{P}\{\xi_{n+1} = j,\, T_{n+1} - T_n \leqslant t \mid \xi_1, \ldots, \xi_n; T_1, \ldots, T_n\}\\ &\quad = \mathbb{P}\{\xi_{n+1} = j,\, T_{n+1} - T_n \leqslant t \mid \xi_n\}\end{aligned}$$

for each $n \in \mathbb{N}$, $j \in E$ and all $t \in \mathbb{R}_+$, through the relation

$$x_t = \sum_{n \geqslant 0} \xi_n \mathbf{1}_{[T_n, T_{n+1})}(t). \tag{3.3}$$

In economic terms, the representative agent's mood in the random time interval $[T_n, T_{n+1})$ is given by ξ_n. The distribution of the length of the interval $T_{n+1} - T_n$ may depend on the sequence $\{\xi_n\}_{n \in \mathbb{N}}$ through the states ξ_n and ξ_{n+1}. This allows us to assume different distributions for the lengths of the agents' active and inactive periods, and in particular to model inertia as a heavy-tailed sojourn time in the zero state.

Remark 3.2. In the present analysis of investor inertia, we do not allow for feedback effects of prices into agents' investment decisions. While such an assumption might be justified for small, non-professional investors, it is clearly desirable to allow active traders' investment decisions to be influenced by asset prices. We discuss such an extension in the next section.

We assume that x is temporally homogeneous under the measure $\mathbb{P}$, that is,

$$Q(i, j, t) \triangleq \mathbb{P}\{\xi_{n+1} = j,\ T_{n+1} - T_n \leqslant t \mid \xi_n = i\} \tag{3.4}$$

is independent of $n \in \mathbb{N}$. By Çinlar (1975, Proposition 1.6), this implies that $\{\xi_n\}_{n \in \mathbb{N}}$ is a homogeneous Markov chain on E whose transition probability matrix (p_{ij}) is given by

$$p_{ij} = \lim_{t \to \infty} Q(i, j, t).$$

Clearly, x is an ordinary temporally homogeneous Markov process if Q takes the form

$$Q(i, j, t) = p_{ij}\big(1 - e^{-\lambda_i t}\big). \tag{3.5}$$

We also assume that the *embedded Markov chain* $\{\xi_n\}_{n \in \mathbb{N}}$ satisfies $p_{ij} > 0$ so that $\{\xi_n\}_{n \in \mathbb{N}}$ has a unique stationary distribution. The conditional distribution function of the length of the n-th sojourn time, $T_{n+1} - T_n$, given ξ_{n+1} and ξ_n is specified in terms of the *semi-Markov kernel* $\{Q(i, j, t);\ i, j \in E,\ t \geqslant 0\}$ and

the transition matrix P by

$$G(i,j,t) := \frac{Q(i,j,t)}{p_{ij}} = \mathbb{P}\{T_{n+1} - T_n \leqslant t \mid \xi_n = i,\ \xi_{n+1} = j\}. \tag{3.6}$$

The semi-Markov processes are assumed to satisfy the following conditions.

Assumption 3.3.

(i) The average sojourn time at state $i \in E$ is finite:

$$m_i := \mathbb{E}[T_{n+1} - T_n \mid \xi_n = i] < \infty. \tag{3.7}$$

Here $\mathbb{E}$ denotes the expectation operator with respect to $\mathbb{P}$.

(ii) There exists a constant $1 < \alpha < 2$ and a locally bounded function $L : \mathbb{R}_+ \to \mathbb{R}_+$ which is slowly varying at infinity (e.g. log), i.e.,

$$\lim_{t\to\infty} \frac{L(xt)}{L(t)} = 1 \quad \text{for all } x > 0,$$

such that

$$\mathbb{P}\{T_{n+1} - T_n \geqslant t \mid \xi_n = 0\} \sim t^{-\alpha} L(t). \tag{3.8}$$

Here we use the notation $f(t) \sim g(t)$ for two functions $f, g : \mathbb{R}_+ \to \mathbb{R}_+$ to mean that $\lim_{t\to\infty} f(t)/g(t) = 1$.

(iii) The distributions of the sojourn times at state $i \neq 0$ satisfy

$$\lim_{t\to 0} \frac{\mathbb{P}\{T_{n+1} - T_n \geqslant t \mid \xi_n = i\}}{t^{-(\alpha+1)} L(t)} = 0. \tag{3.9}$$

(iv) The distribution of the sojourn times in the various states have continuous and bounded densities with respect to Lebesgue measure on $\mathbb{R}_+$.

The key parameter is the tail index α of the sojourn time distribution of the inactive state zero. Condition (3.8) is satisfied if, for instance, the length of the sojourn time at state $0 \in E$ is distributed according to a Pareto distribution. The idea of inertia is then reflected by (3.9): the probability of long uninterrupted trading periods is small compared to the probability of an individual agent being inactive for a long time. In fact, it is natural to think of the sojourn times in the various active states as being thin tailed as in the exponential distribution since small investors typically do not trade persistently.

3.1.2 A limit theorem for financial markets with inert investors

We assume that the semi-Markov processes x^a are stationary. Stationarity can be achieved by a suitable specification of the common distribution of the initial states and initial sojourn times. We denote the resulting measure on the canonical path space by $\mathbb{P}^*$. Independence and stationarity of the semi-Markov processes guarantees that the logarithmic price process can be approximated

pathwise by the process $\{s_t\}_{t\geqslant 0}$ defined by

$$s_t = \mu \int_0^t \Psi_s \, \mathrm{d}s \quad \text{where } \mu := \mathbb{E}^* x_0^a$$

when the number of agents grows to infinity. Our functional central limit theorem for stationary semi-Markov processes shows that after suitable scaling, the fluctuations around $(s_t)_{t\geqslant 0}$ can be approximated in law by a process with long range dependence. The convergence concept we shall use is weak convergence with respect to the measure $\mathbb{P}^*$ of the Skorohod space $\mathbb{D}$ of all right continuous processes. We write $\mathcal{L}\text{-}\lim_{n\to\infty} Y^n = Y$ if $\{Y^n\}_{n\in\mathbb{N}}$ is a sequence of $\mathbb{D}$-valued stochastic processes that converges weakly to the process Y.

The convergence result is formulated in terms of a scaling limit for the processes $\{x^a_{Tt}\}_{t\geqslant 0}$ ($T \in \mathbb{N}$). For T large, x^a_{Tt} is a "speeded-up" semi-Markov process. In other words, the investors' individual trading dispensations are evolving on a faster scale than Ψ. Observe, however, that we are not altering the main qualitative feature of the model: agents still remain in the inactive state for relatively much longer times than in an active state. In the rescaled model the logarithmic asset price process $S^{N,T}$ is given by

$$S_t^{N,T} = \frac{1}{N} \int_0^t \sum_{a\in\mathbb{A}} \Psi_u x^a_{Tu} \, \mathrm{d}u. \tag{3.10}$$

The central limit theorem allows us to approximate the fluctuations around the first-order approximation as $N \to \infty$. In terms of the Gaussian processes X^T and Y^T defined by

$$X_t^T \triangleq \mathcal{L}\text{-}\lim_{N\to\infty} T^{1-H} \frac{1}{\sqrt{N}} \sum_{a=1}^{N} \bigl(x^a_{Tt} - \mu t\bigr) \quad \text{and} \quad Y_t^T \triangleq \int_0^t X_s^T \, \mathrm{d}s, \tag{3.11}$$

with $H = (3-\alpha)/2$, the fluctuations around the first-order approximation can be approximated by an integral of the elasticity coefficient with respect to Y^T:

$$\mathcal{L}\text{-}\lim_{N\to\infty} \sqrt{N}\bigl\{S_t^{N,T} - \mu t\bigr\}_{0\leqslant t\leqslant 1} = \left\{ \int_0^t \Psi_s \, \mathrm{d}Y_s^T \right\}_{0\leqslant t\leqslant 1}.$$

In order to see more clearly the effects of investor inertia, we rescale the price process in space and time and let T tend to infinity. In a benchmark model with many agents where $\Psi \equiv 1$ these fluctuations, when suitably normalized, can be described by a fractional Brownian motion B^H if $T \to \infty$. The Hurst coefficient is related to the degree of investor inertia.

Theorem 3.4. *(See Bayraktar et al., 2006.) Let $H = \frac{3-\alpha}{2}$. Assume that $\Psi \equiv 1$, that Assumption 3.3 holds and that $\mu \neq 0$. Then there exists $\sigma > 0$ such that*

$$\mathcal{L}\text{-}\lim_{T\to\infty} \mathcal{L}\text{-}\lim_{N\to\infty} T^{1-H} \frac{\sqrt{N}}{\sqrt{L(T)}} \{S_t^{N,T} - \mu t\}_{0\leqslant t\leqslant 1} = \{\sigma B_t^H\}_{0\leqslant t\leqslant 1}. \tag{3.12}$$

To generalize this result to a market in which the agents' order rates are coupled by a stochastic elasticity coefficient as in (3.2), we need the following approximation result for stochastic integrals of continuous semi-martingales with respect to fractional Brownian motion.

Theorem 3.5. *(See Bayraktar et al., 2006.) Let $\{\Psi^n\}_{n\in\mathbb{N}}$ be a sequence of good semimartingales and $\{Z^n\}_{n\in\mathbb{N}}$ be a sequence of $\mathbb{D}$-valued stochastic processes that satisfy*

(i) *The sample paths of the processes Z^n are almost surely of zero quadratic variation on compact sets, and $\mathbb{P}\{Z_0^n = 0\} = 1$.*
(ii) *The stochastic integrals $\int \Psi^n \,\mathrm{d}Z^n$ and $\int Z^n \,\mathrm{d}Z^n$ exist as limits in probability of Stieltjes-sums, and the sample paths $t \mapsto \int_0^t Z_s^n \,\mathrm{d}Z_s^n$ and $t \mapsto \int_0^t \Psi_s^n \,\mathrm{d}Z_s^n$ are càdlàg.*

If Ψ is a continuous semimartingale and if B^H is a fractional Brownian motion process with Hurst parameter $H > \frac{1}{2}$, then the convergence $\mathcal{L}\text{-}\lim_{n\to\infty}(\Psi^n, Z^n) = (\Psi, B^H)$ implies the convergence

$$\mathcal{L}\text{-}\lim_{n\to\infty} \left(\Psi^n, Z^n, \int \Psi^n \,\mathrm{d}Z^n\right) = \left(\Psi, B^H, \int \Psi \,\mathrm{d}B^H\right).$$

As an immediate corollary to Theorem 3.5 we see that the fluctuations of the price process (3.10) around its first-order approximation converge in distribution to a stochastic integral with respect to fractional Brownian motion.

Corollary 3.6. *Let Ψ be a continuous semi martingale with Doob–Meyer decomposition $\Psi = M + A$. If $\mathbb{E}\{[M, M]_T\} < \infty$, $\mathbb{E}\{|A|_T\} < \infty$ and $\mu \neq 0$, then there exists $\sigma > 0$ such that*

$$\begin{aligned} &\mathcal{L}\text{-}\lim_{T\to\infty} \mathcal{L}\text{-}\lim_{N\to\infty} T^{1-H} \frac{\sqrt{N}}{\sqrt{L(T)}} \left\{S_t^{N,T} - \mu \int_0^t \Psi_s \,\mathrm{d}s\right\}_{0\leqslant t\leqslant 1} \\ &\quad = \left\{\sigma \int_0^t \Psi_s \,\mathrm{d}B_s^H\right\}_{0\leqslant t\leqslant 1}. \end{aligned} \tag{3.13}$$

The increments of a fractional Brownian motion with Hurst coefficient $H \in (\frac{1}{2}, 1]$ are positively correlated. The correlation increases in H. Thus, the limit theorem reveals that, in isolation, investor inertia may lead to long

range dependence in asset returns. Indeed, a greater degree of inactivity, represented by a smaller tail index α, leads to a larger H, and so greater positive correlation between returns. Since fractional Brownian motion is not a semimartingale, it may also lead to arbitrage opportunities for other traders whose impact has not been considered in the model so far. Explicit arbitrage strategies for various models were constructed in, e.g. Bayraktar and Poor (2005).

Remark 3.7. In a model without inertia where all the sojourn time distributions are thin-tailed, the logarithmic stock price fluctuations can be approximated in law by a process of the form

$$\left\{\int_0^t \Psi_s \, \mathrm{d}W_s\right\}_{0\leqslant t\leqslant 1} \tag{3.14}$$

where W is a standard Brownian motion. Thus, when all traders' mood processes are standard Markov processes and Ψ is constant, we recover in the limit the standard Black–Scholes–Samuelson geometric Brownian motion model.

The approach of studying queuing systems through their limiting behavior has a long history in many applications, see Whitt (2002), for example. This analysis of investor inertia is built upon the works of Taqqu et al. (1997) on Internet traffic. However, even the simple model we have discussed so far shows how economic applications lead to new mathematical challenges: in the teletraffic application, it is sufficient to consider a binary (on/off) state space, but when agents buy, sell or do nothing, there must be at least three states. This requires different techniques from the binary case. Our functional central limit theorems for stationary semi-Markov processes may also serve as a mathematical basis for proving heavy-traffic limits in the multilevel network models studied in, e.g. Duffield and Whitt (1998a, 1998b).

3.2 A limit theorem with feedback effects

The model in the previous section assumes that investors' actions affect the price, but prices did not affect the agents' demands. This assumption might be justified for Internet or new economy stocks where no accurate information about the actual underlying fundamental value is available. In such a situation, price is not always a good indicator of value and is often ignored by uninformed small investors. In general, however, it is certainly desirable to allow for feedback effects from current prices into the agents' order rates. In this section we extend our previous model to allow for feedback effects from prices into the agents' order rates. At the same time we provide a unified mathematical framework for analyzing microstructure models with asynchronous order arrivals. Our approach is based on methods and techniques from state dependent Markovian service networks. Mathematically, it extends earlier results in

Anisimov (2002) beyond semi-Markov models with thin-tailed sojourn time distributions.

3.2.1 The dynamics of logarithmic asset prices

Let us now be more precise about the probabilistic structure our model. We assume that the agents' orders arrive with an order rate that depends on the price and the investor sentiment. Each order is good for one unit of the stock. Specifically, we associate to each agent $a \in \mathbb{A}$ two independent standard Poisson processes $\{\Pi^a_+(t)\}_{t\geqslant 0}$ and $\{\Pi^a_-(t)\}_{t\geqslant 0}$, a stationary semi-Markov process x^a on E satisfying Assumption 3.3, and bounded Lipschitz continuous *rate functions* $\lambda_\pm : E \times \mathbb{R} \to \mathbb{R}^+$. The rate functions along with the Poisson processes $\Pi^a_\pm$ specify the arrivals times of buying and selling orders. The agent's holdings at time $t \geqslant 0$ are given by

$$\Pi^a_+\left(\int_0^t \lambda_+\big(x^a_u, S^N_u\big)\,\mathrm{d}u\right) - \Pi^a_-\left(\int_0^t \lambda_-\big(x^a_u, S^N_u\big)\,\mathrm{d}u\right) \tag{3.15}$$

where $\{S^N_t\}_{t\geqslant 0}$ denotes the logarithmic asset price process. As before, a buying (selling) order increases (decreases) the logarithmic price by $1/N$. Assuming for simplicity that $S^N_0 = 0$, we thus obtain

$$\begin{aligned} S^N_t = {} & \frac{1}{N}\sum_{a\in\mathbb{A}} \Pi^a_+\left(\int_0^t \lambda_+\big(x^a_u, S^N_u\big)\,du\right) \\ & - \frac{1}{N}\sum_{a\in\mathbb{A}} \Pi^a_-\left(\int_0^t \lambda_-\big(x^a_u, S^N_u\big)\,du\right). \end{aligned} \tag{3.16}$$

Remark 3.8.

(i) In the model studied in the previous section, the agents continuously accumulated the stock at rates specified by semi-Markov processes. Our current models assume that stocks are purchased at random points in times. The arrival times of buying and selling times follow exponential distributions conditional on random arrival rates that depend on current prices and exogenous semi-Markov processes.

(i) As before, we think of x^a as being the investor's "mood" (for trading) process. Loosely speaking, $\lambda_+(x^a_t, s) - \lambda_-(x^a_t, s)$ may be viewed as the agent's excess demand at time t at a logarithmic price level s, given his trading mood x^a_t.

(ii) To develop a model of interaction, in which the participants are inert, out of (3.15), it is natural to assume that $\lambda_\pm(0, s) \equiv 0$ and that the buying and selling rates $\lambda_+(x, \cdot)$ and $\lambda_-(x, \cdot)$ are increasing, resp. decreasing, in the second variable meaning that meaning high (low) prices temper buying (selling) rates.

The sum of independent Poisson processes is a Poisson process with intensity given by the sum of the intensities. As a result, the logarithmic price process satisfies the equality

$$S_t^N = \frac{1}{N}\Pi_+\left(\sum_{a=1}^{N}\int_0^t \lambda_+\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - \frac{1}{N}\Pi_-\left(\sum_{a=1}^{N}\int_0^t \lambda_-\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) \tag{3.17}$$

in *distribution* where Π_+ and Π_- are independent standard Poisson processes. Since our focus will be on a limit result for the *distribution* of the price process as the number of agents grows to infinity, we may with no loss of generality assume that the logarithmic price process is *defined* by (3.17) rather than (3.16).

Assumption 3.9.

(i) The rate functions $\lambda_\pm$ are uniformly bounded.
(ii) For each $x \in E$, the rate functions $\lambda_\pm(x,\cdot)$ are continuously differentiable with first derivative bounded in absolute value by some constant L.

Our convergence results will be based on the following strong approximation result which allows for a pathwise approximation of a Poisson process by a standard Brownian motion living on the same probability space.

Lemma 3.10. *(See* Kurtz, 1978*.) A standard Poisson process* $\{\Pi(t)\}_{t\geqslant 0}$ *can be realized on the same probability space as a standard Brownian motion* $\{B(t)\}_{t\geqslant 0}$ *in such a way that the almost surely finite random variable*

$$\sup_{t\geqslant 0}\frac{|\Pi(t) - t - B(t)|}{\log(2\vee t)}$$

has a finite moment generating function in the neighborhood of the origin and in particular finite mean.

In view of Assumption 3.9(i), the strong approximation result yields the following alternative representation of the logarithmic asset price process:

$$S_t^N = \frac{1}{N}\left\{\sum_{a=1}^{N}\int_0^t \lambda\big(x_u^a, S_u^N\big)\,\mathrm{d}u + B_+\left(\sum_{a=1}^{N}\int_0^t \lambda_+\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - B_-\left(\sum_{a=1}^{N}\int_0^t \lambda_-\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right)\right\} + \mathcal{O}\left(\frac{\log N}{N}\right), \tag{3.18}$$

where $\lambda(x_u^a, \cdot)$ denotes the excess order rate of the agent $a \in \mathbb{A}$, given his mood for trading x_u^a and $\mathcal{O}(\log N/N)$ holds uniformly over compact time intervals. Using this representation of the logarithmic price process our goal is to prove approximation results for the process $\{S_t^N\}_{t\geqslant 0}$. In a first step we show that it can almost surely be approximated by the trajectory of an ordinary differential equation ("fluid limit"). In the subsequent step, we apply a result from (Bayraktar et al., 2006) to show that, after suitable scaling, the fluctuations around this first-order approximation can be described in terms of a fractional process $\{Z_t\}_{t\geqslant 0}$ of the form

$$\mathrm{d}Z_t = \mu_t Z_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}B_t^H.$$

In a benchmark model without feedback, where the order rates do not depend on current prices, the process $\{Z_t\}_{t\geqslant 0}$ reduces to a fractional Brownian motion. That is, we recover the type of results of Section 3.1.2 with the alternative model presented in this section.

3.2.2 First-order approximation

In order to prove our first convergence result, it is convenient to denote by

$$\lambda(x, s) \triangleq \lambda_+(x, s) - \lambda_-(x, s) \tag{3.19}$$

the accumulated net order rate at a given logarithmic price level $s \in \mathbb{R}$ and trading mood $x \in E$ and by

$$\bar{\lambda}(s) \triangleq \bar{\lambda}_+(s) - \bar{\lambda}_-(s)$$

the expected excess order flow where

$$\bar{\lambda}_\pm(s) \triangleq \int_E \lambda_\pm(x, s)\nu(\mathrm{d}x),$$

and ν is the stationary distribution of the semi-Markov process x_t. We are first going to show that in a financial market with many agents the dynamics of the logarithmic price process can be approximated by the solution $\{s_t\}_{t\geqslant 0}$ to the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} s_t = \bar{\lambda}(s_t), \tag{3.20}$$

with initial condition $s_0 = 0$. To this end, we need to prove that the average excess order rate converges almost surely to the expected excess order flow uniformly on compact time intervals.

Lemma 3.11. *Uniformly on compact time intervals*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{a=1}^{N} \int_0^t \lambda_\pm\big(x_u^a, s_u\big)\,\mathrm{d}u = \int_0^t \bar{\lambda}_\pm(s_u)\,\mathrm{d}u \quad \mathbb{P}^*\text{-a.s.} \tag{3.21}$$

Proof. The stationary semi Markov processes x^a are independent, and so the random variables $\int_0^t \lambda(x_u^a, s_u)\,\mathrm{d}u$ $(a = 1, 2, \ldots)$ are also independent. Thus, the law of large numbers for independent random variables along with Fubini's theorem (to exchange the sum and the integral) and bounded convergence theorem (to exchange the limit and the integral) yields convergence for each t. In order to prove that the convergence holds uniformly over compact time intervals we will use uniform law of large numbers of Potscher and Prucha (1989). Denoting $\mathrm{D}_E[0, t]$ the class of all cádlág functions $y : [0, t] \to E$ we need to show that the maps $q_\pm : \mathrm{D}_E[0, t] \times [0, t] \to \mathbb{R}$ defined by

$$q_\pm(y, t) \triangleq \int_0^t \lambda_\pm\big(y(u), s_u\big)\,\mathrm{d}u$$

are continuous. Since the rate functions are bounded, it is enough to show that the map $y \mapsto \int_0^{\cdot} \lambda_\pm(y(u), s_u)\,\mathrm{d}u$ is continuous uniformly over compact time intervals.

To this end, we denote by d the metric defined in Ethier and Kurtz (1986, (3.5.2)) which induces the Skorohod topology in $D_E[0, t]$ and recall that $\lim_{n\to\infty} d(y_n, y) = 0$ if and only if

$$\lim_{n\to\infty} \sup_{0\leqslant s\leqslant t} \big|y_n \circ \tau_n(s) - y(s)\big| = 0 \tag{3.22}$$

for a suitable sequence of strictly increasing time-shifts τ_n; see Ethier and Kurtz (1986, p. 117) for details. Let $\{y_n\}$ denote a sequence in $D_E[0, t]$ that converges to y and put

$$\lambda_\pm^n(u) \triangleq \lambda_\pm\big(y_n(u), s_u\big).$$

In view of the transformation formula for Lebesgue integrals and because $\tau(0) = 0$ and $\tau_n^{-1}(t) \leqslant t$ we obtain

$$\begin{aligned}
\int_0^t \big[\lambda_\pm^n(u) - \lambda_\pm(u)\big]\,\mathrm{d}u &= \int_0^{\tau_n^{-1}(t)} \big[\lambda_\pm^n \circ \tau_n(u)\tau_n'(u) - \lambda_\pm(u)\big]\,\mathrm{d}u \\
&\quad - \int_{\tau_n^{-1}(t)}^t \lambda_\pm(u)\,\mathrm{d}u \\
&= \int_0^{\tau_n^{-1}(t)} \big[\lambda_\pm^n \circ \tau_n(u) - \lambda_\pm(u)\big]\,\mathrm{d}u \\
&\quad + \int_0^{\tau_n^{-1}(t)} \lambda_\pm^n \circ \tau_n(u)\big[\tau_n'(u) - 1\big]\,\mathrm{d}u
\end{aligned}$$

$$- \int_{\tau_n^{-1}(t)}^{t} \lambda_{\pm}(u)\,\mathrm{d}u.$$

By Ethier and Kurtz (1986, (3.5.5)–(3.5.7)),

$$\lim_{n\to\infty} \sup_{0\leqslant u\leqslant t} \left|\tau_n'(u) - 1\right| = 0 \quad \text{and} \quad \lim_{n\to\infty} \sup_{0\leqslant u\leqslant t} \left|\tau_n^{-1}(u) - u\right| = 0$$

so that the last two terms on the right-hand side of the inequality above vanish uniformly on compact time intervals. As far as the first term is concerned, observe that boundedness of the rate function's derivative with respect to the second argument yields

$$\begin{aligned}&\left|\lambda_{\pm}\big(y_n\circ\tau_n(u), s_{\tau_n(u)}\big) - \lambda_{\pm}\big(y(u), s_u\big)\right| \\ &\quad\leqslant L\left|y_n\circ\tau_n(u) - y(u)\right| + L\left|s\circ\tau_n(u) - s(u)\right|.\end{aligned}$$

As a continuous function s is uniformly continuous over compact time intervals. This, along with (3.22) yields

$$\lim_{n\to\infty} \sup_{0\leqslant u\leqslant t} \left|\lambda_{\pm}\big(y_n\circ\tau_n(u), s_{\tau_n(u)}\big) - \lambda_{\pm}\big(y(u), s_u\big)\right| = 0$$

so that the maps $q_{\pm}$ are indeed continuous. Thus, the uniform law of large numbers yields

$$\lim_{N\to\infty} \frac{1}{N}\sum_{a=1}^{N} q_{\pm}\big(x_u^a, u\big) = \lim_{N\to\infty} \frac{1}{N}\sum_{a=1}^{N} \int_0^u \lambda_{\pm}\big(x_v^a, s_v\big)\,\mathrm{d}v = \lambda_{\pm}(\mu, s_u)$$

almost surely uniformly on compact time intervals. □

We are now ready to state and prove our functional law of large numbers.

Theorem 3.12. *As $N \to \infty$, the sequence of stochastic processes $\{S_t^N\}_{t\geqslant 0}$ $(N \in \mathbb{N})$ converges almost surely to the deterministic process $\{s_t\}_{t\geqslant 0}$:*

$$\lim_{N\to\infty} S_t^N = s_t \quad \mathbb{P}^*\text{-a.s.}$$

where the convergence is uniform over compact time intervals.

Proof. In view of the strong approximation result formulated in Lemma 3.10 and because the rate functions are uniformly bounded,

$$\begin{aligned}&\left| \mathit{\Pi}_{\pm}\left(\sum_{a=1}^{N}\int_0^t \lambda_{\pm}\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - \sum_{a=1}^{N}\int_0^t \lambda_{\pm}\big(x_u^a, S_u^N\big)\,\mathrm{d}u \right. \\ &\quad \left. - B_{\pm}\left(\sum_{a=1}^{N}\int_0^t \lambda_{\pm}\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right)\right|\end{aligned}$$

is of the order $O(\log N)$ almost surely where $B_\pm$ are the Brownian motions used in (3.18). Since the rate functions are uniformly bounded, the law of iterated logarithm for Brownian motion yields

$$\lim_{N\to\infty} \sup_{u\leqslant t} \frac{1}{N} B_\pm\left(\sum_{a=1}^{N} \int_0^u \lambda_\pm\big(x_v^a, S_v^N\big)\,\mathrm{d}v\right) = 0 \quad \mathbb{P}^*\text{-a.s.}$$

It follows from this and Lemma 3.11 above, that the quantities

$$B_t^N \triangleq \frac{1}{N}\left|B_+\left(\sum_{a=1}^{N} \int_0^t \lambda_+\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - B_-\left(\sum_{a=1}^{N} \int_0^t \lambda_-\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right)\right|$$

and

$$\Lambda_t^N \triangleq \left|\frac{1}{N}\sum_{a=1}^{N} \int_0^t \big\{\lambda\big(x_u^a, s_u\big) - \bar{\lambda}(s_u)\big\}\,\mathrm{d}u\right|$$

converge to zero uniformly over compact time intervals as $N \to \infty$.

Let us now fix $\epsilon > 0$. Due to Lemma 3.10 there exists $N^* \in \mathbb{N}$ such that for all $N \geqslant N^*$ and uniformly on compact time sets, for $l \leqslant t$ we can write

$$\begin{aligned}
\big|S_l^N - s_l\big| &\leqslant \left|\frac{1}{N}\sum_{a=1}^{N} \int_0^l \lambda\big(x_u^a, S_u^N\big)\,\mathrm{d}u - \int_0^l \bar{\lambda}(s_u)\,\mathrm{d}u\right| + B_l^N + \epsilon \\
&\leqslant \left|\frac{1}{N}\sum_{a=1}^{N} \int_0^l \big\{\lambda\big(x_u^a, S_u^N\big) - \lambda\big(x_u^a, s_u\big)\big\}\,\mathrm{d}u\right| \\
&\quad + \Lambda_l^N + B_l^N + \epsilon \quad \mathbb{P}^*\text{-a.s.}
\end{aligned}$$

Lipschitz continuity of the rate functions yields

$$\begin{aligned}
\big|S_l^N - s_l\big| &\leqslant L\int_0^l \sup_{0\leqslant r\leqslant u} \big|S_r^N - s_r\big|\,\mathrm{d}u + \Lambda_l^N + B_l^N + \epsilon \\
&\leqslant L\int_0^t \sup_{0\leqslant r\leqslant u} \big|S_r^N - s_r\big|\,\mathrm{d}u + \sup_{0\leqslant r\leqslant t} \Lambda_r^N + \sup_{0\leqslant r\leqslant t} B_r^N + \epsilon
\end{aligned}$$

$\mathbb{P}^*$-a.s.

for some $L > 0$ and so

$$\sup_{0\leqslant r\leqslant t}\left|S_r^N - s_r\right| \leqslant L\int_0^t \sup_{0\leqslant r\leqslant u}\left|S_r^N - s_r\right| \mathrm{d}u + \sup_{0\leqslant r\leqslant t} \Lambda_r^N + \sup_{0\leqslant r\leqslant t} B_r^N + \epsilon \quad \mathbb{P}^*\text{-a.s.} \tag{3.23}$$

Now, an application of Gronwall's lemma yields

$$\sup_{0\leqslant r\leqslant t}\left|S_r^N - s_r\right| \leqslant \Big(\sup_{0\leqslant r\leqslant t} \Lambda_r^N + \sup_{0\leqslant r\leqslant t} B_r^N + \epsilon\Big)e^{Lt} \quad \mathbb{P}^*\text{-a.s.}$$

for all $N \geqslant N^*$. This proves our assertion. □

3.2.3 Second-order approximation

In this section we analyze the fluctuations of the logarithmic price process around its first-order approximation. We are interested in the distribution of asset prices around their first-order approximation as $N \to \infty$. In view of the representation (3.18) and by self-similarity of Brownian motion we may thus assume that $\{S_t^N\}_{t\geqslant 0}$ is defined by the integral equation:

$$S_t^N = \frac{1}{N}\sum_{a=1}^N \int_0^t \lambda\big(x_u^a, S_u^N\big)\,\mathrm{d}u + \frac{1}{\sqrt{N}} B_+\left(\frac{1}{N}\sum_{a=1}^N \int_0^t \lambda_+\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - \frac{1}{\sqrt{N}} B_-\left(\frac{1}{N}\sum_{a=1}^N \int_0^t \lambda_-\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) + O\left(\frac{\log N}{N}\right). \tag{3.24}$$

As we shall see, the fluctuations around the first-order approximation are driven by two Gaussian processes. The first,

$$X_t \triangleq B_+\left(\int_0^t \bar{\lambda}_+(s_u)\,\mathrm{d}u\right) - B_-\left(\int_0^t \bar{\lambda}_-(s_u)\,\mathrm{d}u\right), \tag{3.25}$$

captures the randomness in the agents' trading times. The second, $\{Y_t\}_{t\geqslant 0}$, is defined in terms of the integral of a non-stationary Gaussian process whose covariance function depends on the first-order approximation. It captures the second source randomness generated by the agents' trading activity. Specifically,

$$Y_t \triangleq \int_0^t y_s\,\mathrm{d}s, \tag{3.26}$$

where $\{y_t\}_{t\geqslant 0}$ denotes the centered Gaussian process whose covariance function γ is given by the covariance function of the stochastic process

$\{\lambda(x_t, s_t)\}_{t\geqslant 0}$, i.e.,

$$\gamma(t,u) \triangleq \mathbb{E}\big[\lambda(x_t,s_t)\lambda(x_u,s_u)\big] - \bar{\lambda}(s_t)\bar{\lambda}(s_u). \tag{3.27}$$

It turns out that the fluctuations can be approximated in distribution by the process $\{Z_t\}_{t\geqslant 0}$ which satisfies the integral equation

$$Z_t = \int_0^t \bar{\lambda}'(s_u) Z_u \,\mathrm{d}u + Y_t + X_t. \tag{3.28}$$

Our goal is to establish the following second-order approximation for the asset price process in an economy with many market participants.

Theorem 3.13. *The fluctuations of the market imbalance $\{S_t^N\}_{0\leqslant t\leqslant 1}$ around its first-order approximation can be described by the process $\{Z_t\}_{0\leqslant t\leqslant 1}$ defined in (3.28). More precisely,*

$$\mathcal{L}\text{-}\lim_{N\to\infty} \sqrt{N}\big\{S_t^N - s_t\big\}_{0\leqslant t\leqslant 1} = \{Z_t\}_{0\leqslant t\leqslant 1}.$$

The proof of Theorem 3.13 requires some preparation. For notational convenience we introduce stochastic processes $Q^N = \{Q_t^N\}_{0\leqslant t\leqslant 1}$, $Y^N = \{Y_t^N\}_{0\leqslant t\leqslant 1}$ and $X^N = \{X_t^N\}_{0\leqslant t\leqslant 1}$ by, respectively,

$$Q_t^N \triangleq \sqrt{N}\big(S_t^N - s_t\big) \quad \text{and} \quad Y_t^N \triangleq \sum_{a=1}^N \int_0^t \frac{\lambda(x_u^a, s_u) - \bar{\lambda}(s_u)}{\sqrt{N}}\,\mathrm{d}u, \tag{3.29}$$

and

$$X_t^N \triangleq B_+\left(\frac{1}{N}\sum_{a=1}^N \int_0^t \lambda_+\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right) - B_-\left(\frac{1}{N}\sum_{a=1}^N \int_0^t \lambda_-\big(x_u^a, S_u^N\big)\,\mathrm{d}u\right). \tag{3.30}$$

We first prove convergence in distribution of the sequence $\{(X^N, Y^N)\}_{N\in\mathbb{N}}$ to (X, Y).

Proposition 3.14. *The sequence $\{(X^N, Y^N)\}_{N\in\mathbb{N}}$ converges in distribution to the process (X, Y) defined by (3.25) and (3.26).*

Proof. For any $\alpha \in (0, \frac{1}{2})$ and $T > 0$, there exist integrable and hence almost surely finite random variables $M_\pm$ such that for all $t_1, t_2 \leqslant T$ we have

$$\big|B_\pm(t_1) - B_\pm(t_2)\big| \leqslant M_\pm |t_1 - t_2|^\alpha \quad \mathbb{P}^*\text{-a.s.},$$

see, for instance, Karatzas and Shreve (1991, Remark 2.12). Thus, the first-order approximation shows that the sequence of processes $\{X^N\}_{N\in\mathbb{N}}$ converges almost surely to X on any compact time interval. Since the processes

$$\int_0^t \frac{\lambda(x_u^a, s_u) - \bar{\lambda}(s_u)}{\sqrt{N}}\,\mathrm{d}u$$

have Lipschitz continuous sample paths and the semi-Markov processes are independent, the central limit theorem for Lipschitz processes (Whitt (2002, Corollary 7.2.1)) shows that $\{Y^N\}_{N\in\mathbb{N}}$ converges in distribution to the Gaussian process Y. As a result, both sequences $\{X^N\}_{N\in\mathbb{N}}$ and $\{Y^N\}_{N\in\mathbb{N}}$ are tight. Since $\{X^N\}_{n\in\mathbb{N}}$ is also C-tight, the sequence $\{(X^N, Y^N)\}_{N\in\mathbb{N}}$ is tight. It is therefore enough to prove weak convergence of the finite dimensional distributions of the process (X^N, Y^N) to the finite dimensional distributions of (X, Y).

In order to establish weak convergence of the one-dimensional distributions we fix a Lipschitz continuous functions with compact support $F : \mathbb{R}^2 \to \mathbb{R}$. We may with no loss of generality assume that both the Lipschitz constant and the diameter of the support of F equal one. In this case

$$\begin{aligned}&\Big|\int F\big(X_t^N, Y_t^N\big)\,\mathrm{d}\mathbb{P}^* - \int F\big(X_t, Y_t^N\big)\,\mathrm{d}\mathbb{P}^*\Big| \\ &\quad\leqslant \int \min\big\{\big|X_t^N - X_t\big|, 1\big\}\,\mathrm{d}\mathbb{P}^*.\end{aligned}$$

In view of the convergence properties of the sequence $\{X^N\}_{N\in\mathbb{N}}$, there exists, for any $\epsilon > 0$, a constant $N^* \in \mathbb{N}$ such that

$$\sup_{0\leqslant t\leqslant 1}\int \min\big\{\big|X_t^N - X_t\big|, 1\big\}\,\mathrm{d}\mathbb{P}^* \leqslant \epsilon \quad \text{for all } N \geqslant N^*.$$

This yields

$$\lim_{N\to\infty}\Big|\int F\big(X_t^N, Y_t^N\big)\,\mathrm{d}\mathbb{P}^* - \int F\big(X_t, Y_t^N\big)\,\mathrm{d}\mathbb{P}^*\Big| = 0.$$

Since the random variables X_t and Y_t^N are independent, we also have that

$$\lim_{N\to\infty}\int F\big(X_t, Y_t^N\big)\,\mathrm{d}\mathbb{P}^* = \int F(X_t, Y_t)\,\mathrm{d}\mathbb{P}^*.$$

This proves vague convergence[3] of the one-dimensional marginal distributions of (X^N, Y^N) to the one-dimensional distributions of (X, Y) and hence weak

[3] A sequence of probability measure $\{\mu_n\}$ converges to a measure μ in the vague topology if $\lim_{n\to\infty}\int f\,\mathrm{d}\mu_n = \int f\,\mathrm{d}\mu$ for all continuous functions f with bounded support. The vague limit μ is not necessarily a probability measure. However, if there is an *a priori* reason that μ *is* a probability

convergence. Weak Convergence of the finite dimensional distributions follows from similar considerations. □

The following "compact containment condition" is key to the second-order approximation.

Lemma 3.15.

(i) *The sequence of stochastic processes* $\{Q^N\}_{N\in\mathbb{N}}$ *is bounded in probability. That is, for any* $\epsilon > 0$, *there exists* $N^* \in \mathbb{N}$ *and* $K < \infty$ *such that*

$$\mathbb{P}^*\Big[\sup_{0\leqslant t\leqslant 1}\big|Q_t^N\big| > K\Big] < \epsilon \quad \textit{for all } N \geqslant N^*. \tag{3.31}$$

(ii) *If* $f^N = \{f_t^N\}_{t\geqslant 0}$ *be a sequence of non-negative random processes such that*

$$\lim_{N\to\infty}\int_0^1 f_u^N\,\mathrm{d}u = 0 \quad \textit{in probability}, \tag{3.32}$$

then, for all $\delta > 0$,

$$\lim_{N\to\infty}\mathbb{P}^*\left[\sup_{0\leqslant t\leqslant 1}\left|\int_0^t Q_u^N f_u^N\,\mathrm{d}u\right| > \delta\right] = 0.$$

Proof. (i) The strong approximation for Brownian motion yields the representation

$$Q_t^N = \frac{\int_0^t \sum_{a=1}^N \{\lambda(x_u^a, S_u^N) - \lambda(x_u^a, s_u)\}\,\mathrm{d}u}{\sqrt{N}} + Y_t^N + X_t^N + O\left(\frac{\log N}{\sqrt{N}}\right). \tag{3.33}$$

By Proposition 3.14 the sequence $\{(X^N, Y^N)\}_{n\in\mathbb{N}}$ is tight, and hence it is bounded in probability (see e.g. Duffield and Whitt, 1998a). As a result, Lipschitz continuity of the rate functions yields

$$\sup_{0\leqslant t\leqslant 1}\big|Q_t^N\big| \leqslant L\int_0^T \sup_{0\leqslant t\leqslant u}\big|Q_u^N\big|\,\mathrm{d}u + \sup_{0\leqslant t\leqslant 1}\big|Y_t^N\big| + \sup_{0\leqslant t\leqslant 1}\big|X_t^N\big| + \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right),$$

measure, then weak convergence of $\{\mu_n\}$ to μ can be established by analyzing integrals of continuous and hence Lipschitz continuous functions with bounded support. See e.g. Bauer (1992) and Billingsley (1995).

for some $L > 0$. Hence, by Gronwall's inequality,

$$\sup_{0\leqslant t\leqslant 1}|Q_t^N| \leqslant e^{LT}\left[\sup_{0\leqslant t\leqslant 1}|Y_t^N| + \sup_{0\leqslant t\leqslant 1}|X_t^N| + \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right)\right] \quad \mathbb{P}^*\text{-a.s.}$$

This proves (i).

(ii) Let us fix $\epsilon > 0$. There exists a constant N^* such that when $N \geqslant N^*$ there exist sets Ω_N and A_N such that

$$\int_0^1 f_u^N \,\mathrm{d}u < \frac{\epsilon}{2} \quad \text{on } \Omega_N \text{ and such that } \quad \mathbb{P}^*[\Omega_N] \geqslant 1 - \frac{\epsilon}{2}$$

and

$$\sup_{0\leqslant t\leqslant 1}|Q_t^N| < K \quad \text{on } A_N \text{ and such that } \quad \mathbb{P}^*[A_N] \geqslant 1 - \frac{\epsilon}{2}.$$

Hence

$$\sup_{0\leqslant t\leqslant 1}\left|\int_0^t Q_u^N f_u^N \,\mathrm{d}u\right| \leqslant \sup_{0\leqslant t\leqslant 1}|Q_t^N| \int_0^1 f_u^N \,\mathrm{d}u < K\epsilon \quad \text{on } A_N \cap \Omega_N.$$

□

Proof of Theorem 3.13. Let us first define a sequence of stochastic processes $\widetilde{Q}^N = \{\widetilde{Q}_t^N\}_{0\leqslant t\leqslant 1}$ by

$$\widetilde{Q}_t^N \triangleq \int_0^t \bar{\lambda}'(s_u)\widetilde{Q}_u^N \,\mathrm{d}u + Y_t^N + X_t^N.$$

By the continuous mapping theorem and Lemma 3.14 the sequence $\{\widetilde{Q}^N\}_{N\in\mathbb{N}}$ converges in distribution to the process Z defined in (3.28). It is now enough to show that

$$\lim_{N\to\infty}\sup_{0\leqslant t\leqslant 1}\left|Q_t^N - \widetilde{Q}_t^N\right| = 0 \quad \text{in probability.} \tag{3.34}$$

To this end, let $E_t^N \triangleq Q_t^N - \tilde{Q}_t^N$. From the definition of $\tilde{Q}_t^N$ and the representation (3.33) of Q_t^N we obtain

$$E_t^N = \int_0^t \bar{\lambda}'(s_u)E_u^N \,\mathrm{d}u + \frac{1}{\sqrt{N}}\int_0^t \sum_{a=1}^N \left\{\lambda\left(x_u^a, S_u^N\right) - \lambda\left(x_u^a, s_u\right)\right\}\mathrm{d}u$$
$$- \int_0^t \bar{\lambda}'(s_u)Q_u^N \,\mathrm{d}u$$

$$= \int_0^t \bar{\lambda}'(s_u) E_u^N \,\mathrm{d}u + \int_0^t \left(\frac{1}{N} \sum_{a=1}^N \lambda'\big(x_u^a, s_u\big) - \bar{\lambda}'(s_u) \right) Q_u^N \,\mathrm{d}u$$

$$+ \int_0^t \left(\frac{1}{N} \sum_{a=1}^N \lambda'\big(x_u^a, \xi_u^N\big) - \lambda'\big(x_u^a, s_u\big) \right) Q_u^N \,\mathrm{d}u.$$

The second equality follows from the mean value theorem for $\lambda(x_u^a, \cdot)$,

$$\lambda\big(x_u^a, S_u^N\big) - \lambda\big(x_u^a, s_u\big) = \frac{1}{\sqrt{N}} \lambda'\big(x_u^a, \xi_u^N\big) Q_u^N,$$

where ξ_u^N lies between $\frac{1}{N} S_u^N$ and s_u. We put

$$f_u^{N,1} \triangleq \frac{1}{N} \sum_{a=1}^N \lambda'\big(x_u^a, s_u\big) - \bar{\lambda}'(s_u) \quad \text{and}$$

$$f_u^{N,2} \triangleq \frac{1}{N} \sum_{a=1}^N \lambda'\big(x_u^a, \xi_u^N\big) - \lambda'\big(x_u^a, s_u\big)$$

in order to obtain

$$\sup_{0 \leqslant s \leqslant t} \big|E_s^N\big| \leqslant L \int_0^t \sup_{0 \leqslant s \leqslant u} \big|E_s^N\big| \,\mathrm{d}u + \left| \sup_{0 \leqslant s \leqslant t} \int_0^s \big|f_u^{N,1}\big| Q_u^N \,\mathrm{d}u \right|$$

$$+ \left| \sup_{0 \leqslant s \leqslant t} \int_0^s \big|f_u^{N,2}\big| Q_u^N \,\mathrm{d}u \right|.$$

The processes $|f^{N,1}|$ and $|f^{N,2}|$ satisfy the condition (3.32) of Lemma 3.15 by the law of large numbers. Thus, an application of Gronwall's lemma yields (3.34). □

3.2.4 Approximation by a fractional Ornstein–Uhlenbeck process

So far we have shown that the fluctuations of the logarithmic price process around its first-order approximation can be described in terms of an Ornstein–Uhlenbeck process Z driven by two Gaussian processes X and Y. In order to see more clearly the effects of investor inertia on asset processes we need to better understand the dynamics of Y. As before, this will be achieved by a proper scaling of the semi-Markov processes x^a in time and the price process in space. Specifically, we introduce a family of processes $S^{N,T}$ $(T \in \mathbb{N})$ with

initial value 0 by

$$S_t^{N,T} = \frac{1}{NT}\left\{ \Pi_+\left(T \sum_{a=1}^{N} \int_0^t \lambda_+\big(x_{Tu}^a, S_u^{N,T}\big)\, \mathrm{d}u \right) - \Pi_-\left(T \sum_{a=1}^{N} \int_0^t \lambda_-\big(x_{Tu}^a, S_u^{N,T}\big)\, \mathrm{d}u \right) \right\}.$$

The strong approximation result for Poisson processes with respect to Brownian motion allow us to represent the process $\{S_t^{N,T}\}_{t\geqslant 0}$ as in (3.18) with the semi-Markov processes $\{x_t^a\}_{t\geqslant 0}$ replaced by the "speeded-up" processes $\{x_{Tt}^a\}_{t\geqslant 0}$. Moreover, by Lemma 3.11, the sequence of processes

$$\Lambda_t^{N,T} \triangleq \left| \frac{1}{N} \sum_{a=1}^{N} \int_0^t \big\{ \lambda\big(x_{Tu}^a, s_u\big) - \bar{\lambda}(s_u) \big\}\, \mathrm{d}u \right|$$

converges to zero uniformly over compact time intervals as $N \to \infty$. Following the same line of arguments as in the proof of Proposition 3.12 it can then be shown that for any $T > 0$

$$\lim_{N\to\infty} S_t^{N,T} = s_t \quad \mathbb{P}^*\text{-a.s.} \tag{3.35}$$

Here $\{s_t\}_{t\geqslant 0}$ denotes the deterministic process defined by the ordinary differential equation (3.20) with initial condition $s_0 = 0$ and the convergence holds uniformly over compact time intervals. Thus, the first-order approximation is independent of T. By analogy to (3.24)–(3.28) introduce a Gaussian process Y^T by

$$Y_t^T \triangleq \int_0^t y_s^T\, \mathrm{d}s, \tag{3.36}$$

where $\{y_t^T\}_{t\geqslant 0}$ denotes the centered Gaussian process with covariance function

$$\gamma^T(t,u) \triangleq \mathbb{E}\big[\lambda(x_{Tt}, s_t)\lambda(x_{Tu}, s_u)\big] - \bar{\lambda}(s_t)\bar{\lambda}(s_u).$$

Following the same arguments in the proof of Theorem 3.13, we see that as the number of agents tends to infinity the price fluctuations round the fluid limit can be approximated in distribution by a process $\{Z_t^T\}_{t\geqslant 0}$ of the form

$$Z_t^T = \int_0^t \bar{\lambda}'(s_u) Z_u^T\, \mathrm{d}u + Y_t^T + \frac{1}{\sqrt{T}} X_t.$$

Proposition 3.16. *For any T, the fluctuations of the logarithmic price process $\{S_t^{N,T}\}_{0\leqslant t\leqslant 1}$ around its first-order approximation can be described by the process $\{Z_t^T\}_{0\leqslant t\leqslant 1}$. More precisely,*

$$\mathcal{L}\text{-}\lim_{N\to\infty} \sqrt{N}\big\{S_t^{N,T} - s_t\big\}_{0\leqslant t\leqslant 1} = \big\{Z_t^T\big\}_{0\leqslant t\leqslant 1}.$$

To take the T-limit, we need the following assumption on the structure of the rate functions.

Assumption 3.17. The rate function λ defined in (3.19) can be written as

$$\lambda(x, s) = f(x)g(s) + h(s). \tag{3.37}$$

Moreover, the function f in (3.37) is one-to-one and $\hat{\mu} \triangleq f(0) \neq \mathbb{E}^* f(x_0)$.

Example 3.18. The previous assumption is always satisfied if $(x_t^a)_{t\geqslant 0}$ is a stationary on/off process, i.e., if $E = \{0, 1\}$. In this case

$$x_t^a = \frac{\lambda(x_t^a, s_t) - \lambda(0, s_t)}{\lambda(1, s_t) - \lambda(0, s_t)},$$

and the representation (3.37) holds with

$$f(x) = x, \qquad g(s) = \lambda(1, s) - \lambda(0, s) \quad \text{and} \quad h(s) = \lambda(0, s).$$

We are now ready to show that the fluctuations of the logarithmic stock price around its first-order approximation behaves like a fractional Ornstein–Uhlenbeck process.

Theorem 3.19. *Under the Assumptions 3.9 and 3.17 we have that*

$$\mathcal{L}\text{-}\lim_{T\to\infty} \mathcal{L}\text{-}\lim_{N\to\infty} T^{1-H} \frac{\sqrt{N}}{\sqrt{L(T)}} \big\{S_t^{N,T} - s_t\big\}_{0\leqslant t\leqslant 1} = \big\{\hat{Z}_t\big\}_{0\leqslant t\leqslant 1}.$$

Here $\hat{Z}$ denotes unique solution starting at zero to the stochastic differential equation

$$\mathrm{d}\hat{Z}_t = \bar{\lambda}'(s_t)\hat{Z}_t\,\mathrm{d}t + \sigma g(s_t)\,\mathrm{d}B_t^H$$

where B^H is a fractional Brownian motion with Hurst coefficient $H = \frac{3-\alpha}{2}$. The integral with respect to B^H is understood as a limit in probability of Stieltjes sums.

Proof. The proof uses modifications of arguments given in the proof of Theorem 3.13 and the approximation result for integrals with respect to fractional Brownian motion in Bayraktar et al. (2006).

(i) In a first step we study the dynamics of the process $\{Y_t^{N,T}\}_{t\geqslant 0}$ defined by

$$Y_t^{N,T} = \sum_{a=1}^{N} \int_0^t \frac{\lambda(x_{Tu}^a, s_u) - \bar{\lambda}(s_u)}{\sqrt{N}}\, \mathrm{d}u.$$

Under Assumption 3.17 we can write

$$\begin{aligned} Y_t^{N,T} &= \sum_{a=1}^{N} \int_0^t \frac{1}{\sqrt{N}}\big[f\big(x_{Tu}^a\big)g(s_u) + h(s_u) - \bar{\lambda}(s_u)\big]\, \mathrm{d}u \\ &= \sum_{a=1}^{N} \int_0^t \frac{1}{\sqrt{N}}\big[f\big(x_{Tu}^a\big) - \hat{\mu}\big]g(s_u)\, \mathrm{d}u. \end{aligned} \tag{3.38}$$

Since f is one-to-one, $(f(x_t^a))_{t\geqslant 0}$ is a semi-Markov process that has the same sojourn time structure as the underlying semi-Markov process $(x_t^a)_{t\geqslant 0}$. In particular, $f(0)$ is the state whose sojourn time distribution has heavy tails. Therefore it follows from Bayraktar et al. (2006, Theorem 4.1) that

$$\begin{aligned} &\mathcal{L}\text{-}\lim_{T\to\infty} \mathcal{L}\text{-}\lim_{N\to\infty} T^{1-H}\left\{\frac{1}{\sqrt{L(T)}} Y^{N,T}\right\}_{0\leqslant t\leqslant 1} \\ &\quad = \left\{\sigma \int_0^t g(s_u)\, \mathrm{d}B_u^H\right\}_{0\leqslant t\leqslant 1} \end{aligned} \tag{3.39}$$

for some $\sigma > 0$ because $\hat{\mu} \neq f(0)$.

(ii) Let us now define a family of stochastic processes $\widetilde{Q}^{N,T} = \{\widetilde{Q}_t^{N,T}\}_{0\leqslant t\leqslant 1}$ by

$$\widetilde{Q}_t^{N,T} \triangleq \int_0^t \bar{\lambda}'(s_u)\widetilde{Q}_u^{N,T}\, \mathrm{d}u + \frac{T^{1-H}}{\sqrt{L(T)}} Y_t^{N,T} + \frac{T^{1/2-H}}{\sqrt{L(T)}} X_t^N.$$

Since the rate functions are bounded and $H > \frac{1}{2}$

$$\lim_{T\to\infty} \sup_N \sup_{0\leqslant t\leqslant 1} \frac{T^{1/2-H}}{\sqrt{L(T)}} X_t^N = 0 \qquad \mathbb{P}\text{-a.s.}$$

almost surely, and the continuous mapping theorem along with (i) yields

$$\mathcal{L}\text{-}\lim_{T\to\infty} \mathcal{L}\text{-}\lim_{N\to\infty} \{\widetilde{Q}_t^{N,T}\}_{0\leqslant t\leqslant 1} = \{\hat{Z}_t\}_{0\leqslant t\leqslant 1}.$$

(iii) Let us put

$$Q_t^{N,T} \triangleq T^{1-H}\frac{\sqrt{N}}{\sqrt{L(T)}}\big(S_t^{N,T} - s_t\big).$$

Up to a term of the order $\frac{\log N}{\sqrt{N}}$ we obtain

$$Q_t^{N,T} = \frac{\int_0^t \sum_{a=1}^N \{\lambda(x_{Tu}^a, S_u^{N,T}) - \lambda(x_{Tu}^a, s_u)\}\,\mathrm{d}u}{\sqrt{N}} + \frac{T^{1-H}}{\sqrt{L(T)}}Y_t^N + \frac{T^{1/2-H}}{\sqrt{L(T)}}X_t^N.$$

Using the same arguments as in the proof of Theorem 3.13, we thus see that

$$\lim_{N\to\infty}\sup_{0\leqslant t\leqslant 1}|Q_t^{N,T} - \widetilde{Q}_t^{N,T}| = 0 \quad \text{in probability}$$

for all $T \in \mathbb{N}$. Hence the assertion follows from (ii).

□

Remark 3.20. In the case of Markov switching, i.e., when the process x_t is a Markov process, we obtain standard Ornstein–Uhlenbeck process, i.e., we have that

$$\mathcal{L}\text{-}\lim_{T\to\infty}\mathcal{L}\text{-}\lim_{N\to\infty}\sqrt{T}\frac{\sqrt{N}}{\sqrt{L(T)}}\big\{S_t^{N,T} - s_t\big\}_{0\leqslant t\leqslant 1} = \big\{\tilde{Z}_t\big\}_{0\leqslant t\leqslant 1},$$

where $\tilde{Z}$ denotes unique solution to the stochastic differential equation

$$\mathrm{d}\tilde{Z}_t = \bar{\lambda}'(s_t)\tilde{Z}_t\,\mathrm{d}t + \sigma g(s_t)\,\mathrm{d}B_t,$$

with B a standard Brownian motion.

4 Outlook and conclusion

We briefly outline two possible avenues of future research: microstructure models of fractional volatility and strategic interactions between "big players."

4.1 Fractional volatility

In this article, we suggested a microeconomic approach to financial price fluctuations that is capable of explaining the decay of the Hurst coefficient of the S&P 500 index in the late 1990s. We note that the evidence of long memory in stock price returns is mixed, there are several papers in the empirical finance literature providing evidence for the existence of long memory, yet there are

several other papers that contradict these empirical findings; see e.g. Bayraktar et al. (2004) for an exposition of this debate and references. However, long memory is a well accepted feature in volatility (squared and absolute returns) and trading volume (see e.g. Cont, 2001 and Ding et al., 1993). We are now going to illustrate how the mathematical results of this paper might also be seen as an intermediate step towards a microstructural foundation for this phenomenon. To ease notational complexity and to avoid unnecessary technicalities we restrict ourselves to the simplest case where the order rates do not depend on asset prices. Specifically, we assume that (after taking the N-limit) the dynamics of the asset price process can be described by a stochastic equation of the from

$$S_t^T = \frac{1}{T}\left\{\Pi_+\left(T\int_0^t \lambda_+\big(Y_u^T\big)\,\mathrm{d}u\right) - \Pi_-\left(T\int_0^t \lambda_-\big(Y_u^T\big)\,\mathrm{d}u\right)\right\}$$

where the Gaussian process Y^T defined in (3.11) converges in distribution to a fractional Brownian motion process. In view of the strong approximation of Poisson processes by Brownian motion, and because the rate functions are bounded, the evolution of prices can be described in terms of an ordinary differential equation in a random environment generated by a fractional Brownian motion:

$$\mathcal{L}\text{-}\lim_{T\to\infty}\big\{S_t^T\big\}_{0\leqslant t\leqslant 1} = \{\hat{s}_t\}_{t\leqslant 0\leqslant 1} \quad \text{where} \quad \mathrm{d}\hat{s}_t = \lambda\big(B_t^H\big)\,\mathrm{d}t.$$

The fluctuations around this first-order approximation satisfy

$$\sqrt{T}\left(S_t^T - \int_0^t \lambda\big(Y_u^T\big)\,\mathrm{d}u\right) = B_+\left(\int_0^t \lambda_+\big(Y_u^T\big)\,\mathrm{d}u\right) - B_-\left(\int_0^t \lambda_-\big(Y_u^T\big)\,\mathrm{d}u\right),$$

up to a term of the order $\frac{\log T}{\sqrt{T}}$. Convergence of the Gaussian process Y^T to fractional Brownian motion along with continuity of the rate functions yields

$$\mathcal{L}\text{-}\lim_{T\to\infty}\left\{B_\pm\left(\int_0^t \lambda_\pm\big(Y_u^T\big)\,\mathrm{d}u\right)\right\}_{0\leqslant t\leqslant 1} = \left\{\int_0^t \sqrt{\lambda_\pm\big(B_u^H\big)}\,\mathrm{d}B_u^\pm\right\}_{0\leqslant t\leqslant 1}.$$

Thus, for large T, logarithmic asset prices satisfy

$$S_t^T \overset{\mathbb{D}}{\approx} \int_0^t \lambda\big(Y_u^T\big)\,\mathrm{d}u + \frac{1}{\sqrt{T}}\int_0^t \sqrt{\lambda_+\big(Y_u^T\big)}\,\mathrm{d}B_u^+ - \frac{1}{\sqrt{T}}\int_0^t \sqrt{\lambda_-\big(Y_u^T\big)}\,\mathrm{d}B_u^-$$

$$\overset{\mathbb{D}}{\approx} \int_0^t \lambda\big(B_u^H\big)\,\mathrm{d}u + \frac{1}{\sqrt{T}}\int_0^t \sqrt{\lambda_+(B_u^H)}\,\mathrm{d}B_u^+ - \frac{1}{\sqrt{T}}\int_0^t \sqrt{\lambda_-(B_u^H)}\,\mathrm{d}B_u^-,$$

i.e., the volatility is driven by a fractional Brownian motion process which is independent of the Wiener processes B^+ and B^-. We will further elaborate on the microstructure of fractional volatility in a separate paper.

4.2 Strategic interactions

Together with the price taking small investors, it is also possible to incorporate the effects of large investors who influence the price. The existence of large agent price effects has been empirically described in several papers: Kraus and Stoll (1972), Holthausen et al. (1987) and Chan and Lakanishok (1993) describe the impacts of institutional trades on stock prices. In the presence of large agents there is limited liquidity in the market since the holdings of the stocks is concentrated in the hands of a few big traders. Trades of "big player's" also affect stock prices due to large order sizes.

4.2.1 Stochastic equations in strategically controlled environments

Horst (2004, 2005) provides a mathematical framework for analyzing linear stochastic difference equation of the form (2.3) when the dynamics of the random environment is simultaneously controlled by the actions of strategically interacting agents playing a discounted stochastic game with complete information. In Horst (2004), we considered a simple microstructure models where small investors choose their current benchmarks in reaction to the actions taken by some "big players." One may, for example, think of a central bank that tries to keep the "mood of the market" from becoming too optimistic and, if necessary, warns the market participants of emerging bubbles. One may also think of financial experts whose recommendations tempt the agents into buying or selling the stock. These market participants influence the stock price process through their impact on the behavior of small investors, but without actively trading the stock themselves. It seems natural to assume that the big players anticipate the feedback effect their actions have on the evolution of stock prices and thus interact in a strategic manner. Under a weak interaction condition, the resulting stochastic game has a homogeneous Nash equilibrium in Markovian strategies. It turns out that the main qualitative feature of the models studied in Föllmer and Schweizer (1993), Föllmer et al. (2005) and Horst (2005), namely asymptotic stability of stock prices, can be preserved even in a model of strategic interactions. However, the long run distribution of stock prices depends on the equilibrium strategy and is thus not necessarily uniquely determined. Hence, the presence of strategically interacting market participants can be an additional source of uncertainty.

4.2.2 Stochastic games in a non-Markovian setting

Bayraktar and Poor (2005) considered the strategic interaction of large investors and found an equilibrium stock price taking into account that the feedback effects of the large investors on the price. The large traders find themselves in a random environment due to the trades of small (i.e. price taking) investors. In Bayraktar and Poor (2005), the institutional investors strategically interact through the controls they exert on the coefficients of a stochastic differential equation driven by a fractional Brownian motion. Here, the fractional Brownian motion models the effect of the price taking investors on the price. It can be argued that the observed stock price is the Nash-equilibrium price that arises as a result of the strategic interaction of the institutional investors this random environment. Bayraktar and Poor carries out an analysis of stochastic differential games in a non-Markov environment using the stochastic analysis for fractional Brownian motion developed in Duncan et al. (2000). This analysis can be viewed as a first step toward incorporating the feedback effects of the large investors and the strategic interaction into the description of the stock price dynamics.

Acknowledgements

We thank W. Massey and participants of the workshops on "Microscopic Stochastic Dynamics in Economics" and "Complexity and Randomness in Economic Dynamical Systems" at Bielefeld University for valuable comments and discussions.

References

Anisimov, V.V. (2002). Diffusion approximation in overloaded switching queuing models. *Queuing Systems* 40, 143–182.

Barber, B., Odean, T. (2001). The Internet and the investor. *Journal of Economic Perspectives* 15, 41–54.

Barber, B., Odean, T. (2002). Online investors: Do the slow die first? *Review of Financial Studies* 15, 455–487.

Bauer, H. (1992). *Mass-Und Integrationstheorie*. De Gruyter, New York.

Bayraktar, E., Horst, U., Sircar, R. (2006). A limit theorem for financial markets with inert investors. *Mathematics of Operations Research* 31 (4), 798–810.

Bayraktar, E., Poor, H.V. (2005). Arbitrage in fractal modulated Black–Scholes models when the volatility is stochastic. *International Journal of Theoretical and Applied Finance* 8 (3), 1–18.

Bayraktar, E., Poor, H.V. (2005). Stochastic differential games in a non-Markovian setting. *SIAM Journal on Control and Optimization* 43, 1737–1756.

Bayraktar, E., Poor, H.V., Sircar, R. (2004). Estimating the fractal dimension of the S&P 500 index using wavelet analysis. *International Journal of Theoretical and Applied Finance* 7, 615–643.

Bianchi, S. (2005). Pathwise identification of the memory function of multifractional Brownian motion with applications to finance. *International Journal of Theoretical and Applied Finance* 8, 255–281.

Billingsley, P. (1995). Probability and Measure. In: *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York.

Böhm, V., Chiarella, C. (2005). Mean variance preferences, expectations formation, and the dynamics of random asset prices. *Mathematical Finance* 15, 61–97.

Böhm, V., Deutscher, N., Wenzelburger, J. (2000). Endogenous random asset prices in overlapping generations economies. *Mathematical Finance* 10, 23–38.
Böhm, V., Wenzelburger, J. (2005). On the performance of efficient portfolios. *Journal of Economic Dynamics and Control* 29 (4), 721–740.
Brock, W.A., Hommes, C. (1997). A rational route to randomness. *Econometrica* 65, 1059–1095.
Brock, W.A., Hommes, C. (1998). Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic Dynamics and Control* 22, 1235–1274.
Brock, W.A., Hommes, C., Wagener, F. (2005). Evolutionary dynamics in financial markets with many trader types. *Journal of Mathematical Economics* 41, 95–132.
Çinlar, E. (1975). Markov renewal theory: A survey. *Management Science* 21, 727–752.
Cetin, U., Jarrow, R., Protter, P. (2004). Liquidity risk and arbitrage pricing theory. *Finance and Stochastics* 8, 311–341.
Chan, L.K.C., Lakanishok, J. (1993). Institutional trades and intraday stock price behavior. *Journal of Financial Economics* 33, 173–199.
Chen, H., Yao, D. (2001). *Fundamentals of Queuing Networks: Performance, Asymptotics, and Optimization*. Springer.
Chiarella, C., He, X.Z. (2001). Asset pricing and wealth dynamics under heterogeneous expectations. *Quantitative Finance* 1, 509–526.
Choi, J., Laibson, D., Metrick, A. (2002). How does the Internet affect trading? Evidence from investor behavior in 401(k) plans. *Journal of Financial Economics* 64, 397–421.
Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1, 223–236.
Cont, R. (2004). Volatility clustering in financial markets: Empirical facts and agent based models. *Preprint*.
Cont, R., Bouchaud, J.P. (2000). Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics* 4, 170–196.
Davis, M., Esparragoza, J.C. (2004). A queuing network approach. *Preprint*. Department of Mathematics, Imperial College.
Day, R., Huang, W. (1990). Bull, bears, and market sheep. *Journal of Economic Behavior and Organization* 14, 299–329.
Ding, Z., Granger, C.W.J., Engle, R.F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1, 83–106.
Duffield, N.G., Whitt, W. (1998a). Network design and control using on-off and multi-level source traffic models with heavy tailed distributions. In: Park, K., Willinger, W. (Eds.), *Self-Similar Network Traffic and Performance Evaluation*. Wiley, Boston, pp. 421–445.
Duffield, N.G., Whitt, W. (1998b). A source traffic model and its transient analysis for network control. *Stochastic Models* 14, 51–78.
Duncan, T.E., Hu, Y., Pasik-Duncan, B. (2000). Stochastic calculus for fractional Brownian motion. *SIAM Journal on Control and Optimization* 38, 582–612.
Estigneev, I., Hens, T., Schenk-Hoppé, K.R. (2006). Evolutionary stabile stock markets. *Economic Theory* 27 (2), 449–468.
Ethier, S.N., Kurtz, T.G. (1986). Markov Processes: Characterization and Convergence. In: *Wiley Series in Probability and Statistics*. Wiley, New York.
Farmer, J.D., Lillo, F. (2004). On the origin of power law tails in price fluctuations. *Quantitative Finance* 4 (1), 7–11.
Farmer, J.D., Patelli, P., Zovko, I.I. (2005). The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences of the United States of America* 102 (6), 2254–2259.
Föllmer, H. (1994). Stock price fluctuations as a diffusion model in a random environment. *Philosophical Transactions of the Royal Society of London, Series A* 374, 471–483.
Föllmer, H., Horst, U. (2001). Convergence of locally and globally interacting Markov chains. *Stochastic Processes and Their Applications* 96, 99–121.
Föllmer, H., Horst, U., Kirman, A. (2005). Equilibria in financial markets with heterogeneous agents: A probabilistic perspective. *Journal of Mathematical Economics* 41, 123–155.

Föllmer, H., Schweizer, M. (1993). A microeconomic approach to diffusion models for stock prices. *Mathematical Finance* 3, 1–23.

Frankel, J.A., Froot, K. (1986). The dollar as an irrational speculative bubble: A tale of fundamentalists and chartists. *The Marcus Wallenberg Papers on International Finance* 1, 27–55.

Frankel, J.A., Froot, K.A. (1987). Using survey data to test standard propositions regarding exchange rate expectations. *American Economic Review* 77, 133–153.

Frey, R., Stremme, A. (1997). Market volatility and feedback effects from dynamic hedging. *Mathematical Finance* 7, 351–374.

Gabaix, X., Gopikrishnan, P., Plerou, V., Stanley, H.E. (2003). A theory of power law distributions in financial market fluctuations. *Nature* 423, 267–270.

Garman, M. (1976). Market microstructure. *Journal of Financial Economics* 3, 257–275.

Gaunersdorfer, A. (2000). Endogenous fluctuations in a simple asset pricing model with heterogeneous expectations. *Journal of Economic Dynamics and Control* 24, 799–831.

Grasso, R., et al. (2000). *Shareownership 2000*. http://www.nyse.com/pdfs/shareho.pdf.

Hens, T., Schenk-Hoppé, K.R. (2005). Evolutionary stability of portfolio rules in incomplete financial markets. *Journal of Mathematical Economics* 41, 123–155.

Holthausen, R., Leftwich, R., Mayers, D. (1987). The effect of large block transactions on security prices: A cross-sectional analysis. *Journal of Financial Economics* 19, 237–267.

Hommes, C. (2006). Heterogeneous agent models in economics and finance. In: Judd, K., Tesfatsion, L. (Eds.), *Handbook of Computational Economics II: Agent-Based Computational Economics*. North-Holland.

Horst, U. (2002). Asymptotics of locally interacting Markov chains with global signals. *Advances in Applied Probability* 34, 1–25.

Horst, U. (2004). Stability of linear stochastic difference equations in strategically controlled random environments. *Advances in Applied Probability* 35, 961–981.

Horst, U. (2005). Equilibria in discounted stochastic games with weakly interacting players. *Games and Economic Behavior* 52, 83–108.

Horst, U. (2005). Financial price fluctuations in a stock market model with many interacting agents. *Economic Theory* 25 (4), 917–932.

Horst, U., Wenzelburger, J. (2005). Non-ergodic price dynamics in financial markets with heterogeneous agents. *Working paper*.

Jonsson, M., Keppo, J. (2002). Option pricing for large agents. *Applied Mathematical Finance* 9, 261–272.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47, 263–291.

Karatzas, I., Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.

Kirman, A. (1992). Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6, 117–136.

Kraus, A., Stoll, H. (1972). Price impacts of block trading on the New York Stock Exchange. *Journal of Finance* 27, 569–588.

Kruk, L. (2003). Functional limit theorems for a simple auction. *Mathematics of Operations Research* 28 (4), 716–751.

Kurtz, T.G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications* 6, 223–240.

Luckock, H. (2003). A steady-state model of continuous double auction. *Quantitative Finance* 3, 385–404.

Lux, T. (1995). Herd behavior, bubbles and crashes. *Economic Journal* 105, 881–896.

Lux, T. (1997). Time variation of second moments from a noise trader/infection model. *Journal of Economic Dynamics and Control* 22, 1–38.

Lux, T. (1998). The socio-economic dynamics of speculative markets: Interacting agents, chaos, and the fat tails of return distributions. *Journal of Economic Behavior and Organization* 33, 143–165.

Lux, T., Marchesi, M. (2000). Volatility clustering in financial markets: A microsimulation of interacting agents. *International Journal of Theoretical and Applied Finance* 3, 675–702.

Madrian, B., Shea, D. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics* 116, 1149–1187.

Mandelbaum, A., Massey, W., Reiman, M. (1998). Strong approximations for Markovian service networks. *Queuing Systems* 30, 149–201.

Mandelbaum, A., Pats, G. (1998). State-dependent stochastic networks, part I: Approximations and applications with continuous diffusion limits. *The Annals of Applied Probability* 8, 569–646.

Mehra, R., Prescott, E.C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics* 15, 145–161.

Mendelson, H. (1982). Market behavior in a clearing house. *Econometrica* 50 (6), 1505–1524.

O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell, MA.

Platen, E., Schweizer, M. (1998). On feedback effects from hedging derivatives. *Mathematical Finance* 8, 67–84.

Potscher, B.M., Prucha, I.R. (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica* 57, 675–683.

Sandroni, A. (2000). Do markets favour agents able to make accurate predictions? *Econometrica* 69, 1303–1341.

Schönbucher, P., Wilmott, P. (2000). The feedback effect of hedging in illiquid markets. *SIAM Journal of Applied Mathematics* 61 (1), 232–272.

Shefrin, H., Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40, 777–790.

Sircar, K.R., Papanicolaou, G.C. (1998). General Black–Scholes models accounting for increased market volatility from hedging strategies. *Applied Mathematical Finance* 5 (1), 45–82.

Smith, E., Farmer, J.D., Gillemot, L., Krishnamurthy, S. (2003). Statistical theory of continuous double auctions. *Quantitative Finance* 3, 481–514.

Taqqu, M.S., Willinger, W., Sherman, R. (1997). Proof of a fundamental result in self-similar traffic modeling. *Computer Communications Review* 27, 5–23.

Taylor, M.P., Allen, H. (1992). The use of technical analysis in the foreign exchange market. *Journal of International Money and Finance* 11, 304–314.

Wenzelburger, J. (2004). Learning to predict rationally when beliefs are heterogeneous. *Journal of Economic Dynamics and Control* 28, 2075–2104.

Whitt, W. (2002). *Stochastic Process Limits*. In: Springer Series in Operations Research. New York.

This page intentionally left blank

PART V

Risk Management

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15016-7

Chapter 16

Economic Credit Capital Allocation and Risk Contributions

Helmut Mausser
Algorithmics Incorporated, Toronto, Canada
E-mail: hmausser@algorithmics.com

Dan Rosen
R^2 Financial Technologies and the Fields Institute, Toronto, Canada
E-mail: drosen@fields.utoronto.ca

Abstract

Economic capital (EC) acts as a buffer for financial institutions to absorb large unexpected losses, thereby protecting depositors and other claim holders and providing confidence to external investors and rating agencies on the financial health of the firm. Once the amount of capital has been determined, it must be allocated equitably among the various components of a portfolio (e.g., activities, business units, obligors or individual transactions). Capital allocation is an important management decision support and business planning tool, required for pricing, profitability assessment and limits, building optimal risk-return portfolios and strategies, performance measurement and risk based compensation.

This chapter provides a practical overview of the measurement of economic credit capital contributions and their application to capital allocation. We discuss the advantages and disadvantages of various risk measures and models, the interpretation of various allocation strategies as well as the numerical issues associated with this task. We stress four key points. First, marginal risk contributions provide a useful basis for allocating EC since they are additive and reflect the benefits of diversification within a portfolio. Second, the choice of the risk measure can have a substantial impact on capital allocation. In particular, Value at Risk (VaR) and expected shortfall (ES) contributions avoid the inconsistencies, and potentially inefficient allocations, associated with the widely-used volatility-based methods. The quantile level chosen for measuring risk can also have a significant impact on the relative amount of capital allocated to portfolio components. Third, VaR and ES contributions can be calculated analytically under certain simple models. These methods provide fast calculations and can be used to understand capital allocation strategies better, but they present important practical limitations, as well. Finally, Monte Carlo methods may be required to compute risk contributions in more realistic credit models. Computing VaR and ES contributions is challenging, especially at the extreme quantiles typically used for credit capital definition. The quality of contribution estimates can be improved by exploiting the conditional independence framework underlying the most common

models, through the use of more sophisticated quantile estimators (especially for VaR) and through the use of variance reduction techniques, such as Importance Sampling.

1 Introduction

In financial institutions, *economic capital* (EC) acts as a buffer to absorb large unexpected losses, thereby protecting depositors and other claim holders and providing confidence to external investors and rating agencies on the financial health of the firm. In contrast, regulatory capital refers to the minimum capital requirements which banks are required to hold, based on regulations established by the banking supervisory authorities. From the perspective of the regulator, the objectives of capital adequacy requirements are to safeguard the security of the financial institutions and to ensure their ongoing viability, as well as to create a level playing field. For example, regulations for internationally active banks are given by the *Basel Accord*, the framework created by the Basel Committee on Banking Supervision (BCBS), which is now the basis for banking regulation around the world (BCBS, 1988; BCBS, 2004).

Economic capital covers all the risks (e.g., market, credit, operational and business) that may force a financial institution into insolvency. While most of the concepts and methodologies in this chapter have broader applicability, we focus on *economic credit capital* – the buffer against those risks specifically associated with obligor credit events such as *default*, *credit migrations* (downgrades or upgrades) and *credit spread* changes.

Traditionally, capital is designed to absorb *unexpected* losses, at a specified confidence level, while credit reserves are set aside to cover *expected* losses. Thus, EC is commonly defined as the difference between the portfolio's value-at-risk (VaR) and the expected loss of the portfolio. The VaR level (i.e., quantile) is chosen in a way that trades off providing a high return on capital for shareholders with protecting debt holders and depositors (and maintaining a desired credit rating).

Once the amount of capital has been determined, it must be allocated equitably among the various components of the credit portfolio (e.g., activities, business units, obligors or individual transactions). This is vital for management decision support and business planning, performance measurement and risk based compensation, pricing, profitability assessment and limits, as well as building optimal risk-return portfolios and strategies.

There is no unique method to allocate EC; each methodology has its advantages and disadvantages, and might be appropriate for a given managerial application. In particular, marginal risk contributions yield an additive decomposition of EC that accounts for the effects of diversification within

the portfolio. An EC allocation based on the marginal contribution to the volatility (or standard deviation) of the portfolio losses is the most common approach used today by practitioners. However, this allocation scheme is ineffective if the loss distribution is not Normal, as is typical of credit losses. This can produce inconsistent capital charges, and in some cases a loan's capital charge can even exceed its exposure (see Praschnik et al., 2001; Kalkbrener et al., 2004).

Given the definition of EC, it is more natural to allocate capital based on contributions to VaR. However, VaR has several shortcomings since it is not a coherent risk measure (Artzner et al., 1999). Specifically, while VaR is subadditive for Normal distributions, this is not true in general. This limitation is especially relevant for credit loss distributions, which may be far from Normal and not even smooth. Furthermore, VaR refers to one particular loss (i.e., one point in the loss distribution), which makes it difficult to obtain accurate and stable risk contributions with Monte Carlo simulation.

Recently, several authors have proposed using expected shortfall (ES) for allocating EC (see, for example, Kalkbrener et al., 2004). As a coherent risk measure, ES represents a good alternative both for measuring and allocating capital. In particular, Kalkbrener et al. (2004) show that ES yields a linear (or additive), diversifying capital allocation. This requires a slight modification of the standard interpretation of economic capital, namely that it offsets an expected loss conditional on exceeding a certain quantile.

The marginal risk contribution of a position in a portfolio is based on the derivative of the risk measure with respect to the size of that position. While this derivative may not always exist, it has been shown (see Gouriéroux et al., 2000; Tasche, 2000, 2002) that the derivatives of quantile measures (e.g. VaR, or ES) can be expressed as conditional expectations. Several semi-analytical approaches have been proposed for calculating VaR or ES contributions (e.g. Martin et al., 2001; Kurth and Tasche, 2003) and, for certain simple models, risk contributions can be obtained analytically (e.g. Gordy, 2003a; Emmer and Tasche, 2005; Garcia Cespedes et al., 2006).

More realistic credit models typically entail the use of Monte Carlo (MC) simulation, which readily supports diversification through multiple factors and more flexible co-dependence structures, multiple asset classes and default models, as well as stochastic (correlated) modeling of exposures and loss given default. However, computing conditional expectations with simulation is computationally challenging, primarily due to the effects of random noise in the data and the discrete nature of individual credit losses.

Various numerical methods have been proposed to improve the quality of simulation-based risk contributions. For instance, the standard quantile estimator, which considers only a single observation, is known to be unreliable for this purpose (Mausser and Rosen, 1998) and several authors have suggested estimating VaR contributions from multiple observations (Praschnik et al., 2001; Hallerbach, 2003; Mausser, 2003; Mausser and Rosen, 2005). Alternatively, importance sampling generates a greater number of observations from the tail

of the loss distribution in order to obtain more stable estimates (Kalkbrener et al., 2004; Merino and Nyfeler, 2004; Glasserman, 2005).

This chapter provides a practical overview of the measurement of economic credit capital contributions and their application to capital allocation. We discuss the advantages and disadvantages of various risk measures and models, the implications of various allocation strategies, as well as the numerical issues associated with measuring risk contributions.

Various methods of calculating VaR and ES contributions are presented, including both analytical and simulation-based approaches. For the latter, we explain the difficulties associated with estimating the required conditional expectations, and describe numerical techniques, such as semi-analytical convolutions, L-estimators and Importance Sampling, that can be used to address these problems. Several examples are provided to illustrate concepts relevant to using marginal risk contributions for allocating EC and, in particular, stress the shortcomings of volatility contributions relative to quantile-based measures such as VaR and ES.

The reader is further referred to Martin (2004) and Glasserman (this volume) for comprehensive presentations on credit portfolio modeling and computational methodologies for calculating portfolio credit risk. For basic presentations on economic capital and regulatory capital see Aziz and Rosen (2004) and Rosen (2004) and the references cited therein. Finally, for credit portfolio optimization see Mausser and Rosen (2000, 2001).

The rest of this chapter is organized as follows. Section 2 briefly reviews the general framework for credit portfolio models and describes the popular Normal-copula model. Section 3 first introduces capital allocation and then focuses on marginal risk contributions as a way of accomplishing this task. Section 4 describes three models where marginal risk contributions can be obtained analytically: the one-factor credit model in Basel II, the granularity adjustment and a multi-factor extension to Basel II. The more general problem of computing risk contributions with simulation is considered in Section 5. Illustrative examples are presented in Section 6, and Section 7 offers concluding remarks and suggestions for further research.

2 Credit portfolio models and general framework

Over the last decade, several credit portfolio models have been developed for measuring economic credit capital. Some popular industry models include CreditMetrics (Gupton et al., 1997), CreditRisk+ (Credite Suisse Financial Products, 1997), Credit Portfolio View (Wilson, 1997a, 1997b), KMV's Portfolio Manager (Crosbie, 1999). Although the models appear quite different on the surface, they all share an underlying mathematical equivalence among them (Koyluoglu and Hickman, 1998; Gordy, 2000; Frey and McNeil, 2003). The models differ in their distributional assumptions, restrictions, calibration

and solution, but can be calibrated to yield similar results if the input data is consistent.

All of the above are single-period models and generally assume that market risk factors, such as interest rates, are constant. While these assumptions are appropriate for portfolios of simple loans or bonds, they may lead to significant errors when a portfolio contains derivatives or instruments with embedded optionality (such as credit lines), or when exposures vary over time (and hence the timing of the default affects the portfolio losses). An example of a multi-step integrated market and credit risk model that overcomes these limitations is given in Iscoe et al. (1999).

The credit portfolio modeling framework which encompasses these models is referred to as the *conditional independence framework*. In general terms, it consists of five parts:

1. *Systemic Scenarios* ("*states of the world*"). This models the evolution of the relevant "systemic" or sector-specific credit drivers that affect credit events, as well as market factors driving obligor exposures, over the period of analysis.
2. *Conditional default and credit migration probabilities*. Default and migration probabilities vary as a result of changing economic conditions. At each point in time, an obligor's default/migration probabilities are conditioned on the state of the world. Default correlations among obligors are determined from a correlated default/migration model, which describes how changes in the credit drivers affect the conditional default/migration probabilities.
3. *Conditional obligor exposures, recoveries and losses*. The credit exposure to an obligor is the amount that the institution stands to lose should the obligor default. Recovery rates are generally expressed as the percentage of the exposure that is recovered through such processes as bankruptcy proceedings, the sale of assets or direct sale to default markets. Exposures can be assumed to be constant in all scenarios for banking instruments without optionality as well as bonds, but not for other instruments such as derivatives, lines of credit, collateral or unhedged exposures in various currencies.
4. *Conditional portfolio loss distribution*. Conditional on a scenario, obligor defaults (and migrations) are independent. This property facilitates obtaining the conditional portfolio loss distribution, as the conditional portfolio loss is the sum of independent random variables (i.e., obligor losses).
5. *Unconditional portfolio loss distribution*. The unconditional distribution of portfolio credit losses is obtained by averaging the conditional loss distributions over all scenarios.

We now illustrate the framework with the so-called Normal copula model, originally popularized by the CreditMetrics and KMV portfolio models.

2.1 Multi-factor Normal copula model

Consider a portfolio of N obligors and a single-step model. Without loss of generality, assume that each obligor j has a single loan with loss given default and exposure at default given by LGD_j, EAD_j respectively.

For each obligor, we define a continuous variable, the *creditworthiness index* (*CWI*), that represents its financial health, and assume that it has a standard Normal distribution. The CWI of each obligor j depends on d systemic factors (also assumed to be independent and standard Normal) through the multi-factor model

$$\begin{aligned} Y_j &= \beta_{j1}Z_1 + \beta_{j2}Z_2 + \cdots + \beta_{jd}Z_d + \sigma_j\varepsilon_j, \\ \sigma_j &= \sqrt{1 - (\beta_{j1}^2 + \beta_{j2}^2 + \cdots + \beta_{jd}^2)} \end{aligned} \tag{1}$$

where

- $Z_1, \ldots, Z_d$ are independent, standard Normal distributed systemic risk factors;
- ε_j is a standard Normal distributed specific risk factor for obligor j;
- $\beta_{j1}, \ldots, \beta_{jd}$ are the sensitivities of obligor j to the systemic risk factors.

Suppose that obligors migrate into one of R possible credit states, ordered by increasing credit quality (i.e., in a default-only model, $R = 2$ and $r = 1$ is the default state), and let p_{jr} denote the unconditional probability that obligor j transitions into state r.

Assume that defaults for each obligor are driven by a Merton-type model, so that obligor j defaults when its CWI, Y_j, falls below a given threshold at the horizon. If PD_j denotes the obligor's (unconditional) default probability ($PD_j = p_{j1}$), we can express the default threshold by $\Phi^{-1}(PD_j)$, with $\Phi^{-1}(\cdot)$ the cumulative standard Normal distribution. Also in a given scenario given by the outcomes of the risk factors $\boldsymbol{Z}$, the conditional default probability of obligor j is

$$\begin{aligned} p_{j1}(Z) &= \Pr\left[Y_j < \Phi^{-1}(PD_j) \mid Z\right] \\ &= \Phi^{-1}\left(\frac{\Phi^{-1}(PD_j) - (\beta_{j1}Z_1 + \beta_{j2}Z_2 + \cdots + \beta_{jd}Z_d)}{\sigma_j}\right). \end{aligned} \tag{2}$$

Similar formulae are obtained for the conditional transition probabilities to each credit state r (see, for example, Gupton et al., 1997).[1]

Other credit models can be used to obtain different functional forms similar to Eq. (2). For example in a *logit model* the expression for conditional default

[1] This model is also referred to as an *ordered probit* model.

probabilities is given by[2]

$$p_{j1}(Z) = \left[1 + a \cdot \exp\left(b \cdot \left(\sum_{j=1}^{d} \beta_{ij} Z_j\right)\right)\right]^{-1}.$$

Credit migrations can also be handled similarly in such a model (this is referred to as an *ordered logit* model).

Let c_{jr} denote the loss that is incurred if obligor j transitions into state r.[3] Since an integrated market/credit model recognizes that credit losses can be dependent on the systemic factors, one can denote by $c_{jr}(\boldsymbol{Z})$ the loss that results if obligor j transitions into state r, given the factors $\boldsymbol{Z}$ (e.g. Iscoe et al., 1999). However, for simplicity, we ignore such dependence in the sequel.

To understand conditional independence models, consider sampling randomly from the systemic risk factors, $\boldsymbol{Z}$. We refer to each sample Z^m, $m = 1, \ldots, M$, as a scenario. For ease of notation, define the random variable L_j^m to be the loss of obligor j conditional on Z^m, and denote the conditional transition probability and (known) loss of obligor j in state r as p_{jr}^m and c_{jr}^m, respectively. For each obligor j, we have a discrete conditional credit loss distribution

$$L_j^m \sim D_j^m = \left\{\left(c_{jr}^m, p_{jr}^m\right) \mid r = 1, \ldots, R\right\}. \tag{3}$$

Since obligors are conditionally independent, the portfolio loss in scenario m, denoted L^m, is a sum of independent random variables and its distribution is the convolution of the obligor loss distributions

$$L^m \equiv \sum_{j=1}^{N} L_j^m \sim D^m = D_1^m \otimes \cdots \otimes D_N^m \tag{4}$$

D^m has finite support C^m, comprising up to r^N elements (in practice, this number is often less because various combinations of obligor losses sum to the same portfolio loss).

Computing the conditional portfolio loss distribution (Eq. (4)) exactly is computationally challenging if the number of possible portfolio losses (i.e., C^m) is large. However, any method for obtaining the distribution of a sum of independent random variables can be applied in this case. In practice, various techniques that have been used to approximate this convolution efficiently include[4]:

- Fast Fourier Transform methods (in conjunction with a discretization scheme, where losses are assigned to "buckets" in a common grid).

[2] See Wilson (1997a, 1997b), Bucay and Rosen (2000).

[3] Losses are net of recovery in the case of default, $r = 1$.

[4] See, for example, Finger (1999), Martin (2004), Glasserman (this volume), and the references cited therein.

- Saddle Point methods, based on analytical approximations of the distribution around the quantile of interest.
- Analytical approximations of the conditional portfolio loss distributions by simpler distributions. For example, if the portfolio is large and granular, the Law of Large Numbers suggests that the conditional portfolio loss distribution is effectively a point mass at the sum of the mean obligor losses (all other moments vanish). Alternatively, the Central Limit Theorem can be applied under the assumption that all conditional losses are Normally distributed.
- Direct sampling from each obligor's conditional loss distribution (D_j^m) and aggregation to get a sample observation from D^m.

The unconditional loss distribution requires integrating the conditional loss distributions across all scenarios (the joint distribution of the systemic factors). Thus, it is the average of the M conditional portfolio loss distributions (i.e., its support is the union of the C^m and all probabilities are multiplied by $1/M$). In the case of a one-factor model, the integration might be done analytically, but with multi-factor models, simulation is typically required (although some semi-analytical approximations are available, e.g. Pykhtin, 2004).

Credit capital is commonly defined in terms of a high quantile (e.g. in Basel II it is defined at the 99.9% level). Therefore, in practice, it might be difficult to get numerically stable and accurate VaR or expected shortfall estimates for realistic credit portfolios using standard MC methods. This problem is further amplified when calculating risk contributions (see next section). Variance reduction techniques, such as Importance Sampling (IS) and Control Variates, can be applied which reduce significantly the variance of Monte Carlo simulation. For example, IS increases the number of extreme observations, thereby improving accuracy in the tail of the portfolio loss distribution (e.g. Glasserman and Li, 2005). Tchistiakov (2004) applies a control variate technique to estimate portfolio risk, where the control variable is derived from the limiting distribution of a homogeneous portfolio (Vasicek loss distribution) that approximates the portfolio.

3 Capital allocation and risk contributions

In addition to computing the total EC for a portfolio, it is important to develop methodologies to *attribute* this capital *a posteriori* to various subportfolios such as the firm's activities, business units, counterparties or even individual transactions, and to *allocate* it *a priori* in an optimal fashion, to maximize risk-adjusted returns.

In general, the sum of the stand-alone EC for each sub-portfolio or asset is higher than the total portfolio EC due to the benefits of diversification. There is no unique method to allocate EC down a portfolio, and we can classify the

capital allocation methodologies that are currently used in practice into three broad categories[5]:

- *Stand-alone* capital contributions – a sub-portfolio is assigned the amount of capital that it would consume on a stand-alone basis. As such, it does not reflect the beneficial effect of diversification; the sum of stand-alone capital for the individual sub-portfolios may exceed the total EC for the portfolio.
- *Incremental* capital contributions (or *discrete marginal* capital contributions) – calculated by taking the EC for the entire portfolio and subtracting from it the EC for the portfolio without the sub-portfolio. This method captures the amount of capital that would be released if the sub-portfolio were sold or added. Thus it is a natural measure for evaluating the risk of acquisitions or divestitures. A disadvantage of this method is that it does not yield an additive risk decomposition.
- *Marginal* capital contributions (or *diversified* capital contributions) – measures of risk contributions that are additive. By construction, the sum of diversified capital for all the sub-portfolios is equal to total EC for the portfolio. Marginal contributions are specifically designed to allocate the diversification benefit among the sub-portfolios, by capturing the amount of the portfolio's capital that should be allocated to each sub-portfolio, on a marginal basis, when viewed as part of the portfolio.

Several alternative methods for additive risk contributions have been proposed from game theory (see Denault, 2001; Koyluoglu and Stoker, 2002). For example, the Shapley method is based on the formation of coalitions so that a group of units benefits more as a group than if each works separately. This method is computationally intensive, and may be impractical for problems with even a small number of sub-portfolios. A variant called the Aumann–Shapley method requires less computation and is, thus, potentially more practical. Under most (but not all) conditions, these methods yield similar results to marginal contributions. While these methods are today receiving some academic attention, they are mostly not yet used in practice by financial institutions.

We now describe the methodology for capital allocation based on marginal risk contributions, leading to an additive decomposition of a portfolio's risk.

3.1 Definitions

Consider a credit portfolio that contains positions in N obligors (while our discussion assumes that obligors are the components of interest, one could alternatively consider individual loans or transactions). For each obligor j, define x_j to be the size of the position (number of units), and let the random variable

[5] There is no unique terminology for risk contributions, and here we follow Aziz and Rosen (2004).

l_j denote the credit loss per unit position. Credit losses can arise from default events, credit migration or more general spread movements.

If L_j denotes the credit loss due to the jth obligor then the loss of the portfolio is

$$L = \sum_{j=1}^{N} L_j = \sum_{j=1}^{N} l_j x_j. \tag{5}$$

Let F denote the portfolio loss probability distribution, which may or may not be available in closed form (e.g., F may be defined empirically by the results of a Monte Carlo simulation). Let $\rho(L)$ denote a measure of the risk of the portfolio, as implied by the distribution F. An additive decomposition of the risk $\rho(L)$ satisfies

$$\rho(L) = \sum_{j=1}^{N} C_j^{\rho} \tag{6}$$

where C_j^{ρ} represents the risk contribution of obligor j. The relative risk contribution of obligor j is defined to be its proportion of the total risk:

$$R_j^{\rho} = \frac{C_j^{\rho}}{\rho(L)}.$$

If $\rho(L)$ is homogeneous of degree one and differentiable, then Euler's Theorem implies that

$$C_j^{\rho} = x_j \frac{\partial \rho}{\partial x_j}. \tag{7}$$

From Eq. (7), the marginal risk contribution of an entity can be loosely interpreted as the rate of change of the portfolio due to a 1% change in the positions of that entity.

3.2 Risk measures and coherent capital allocation

Economic capital is designed in practice to absorb unexpected losses up to a certain confidence level, α, while credit reserves are set aside to absorb expected losses (EL). Thus, economic capital is typically estimated as the α-quantile of the portfolio loss distribution (VaR_α) minus the expected losses over a specified time horizon[6]:

$$EC_\alpha = VaR_\alpha - EL. \tag{8}$$

[6] The regulatory proposal in Basel II is based on the 99.9% VaR (BCBS, 2004).

This is the approach commonly taken by practitioners, and generally leads to conservative estimates of EC. More formally Eq. (8) represents only a simplifying approximation to the true EC (see Kupiec, 2002; Aziz and Rosen, 2004). The rationale for subtracting *EL* is that credit products are already priced such that net interest margins less non-interest expenses are sufficient to cover estimated *EL* (and also a desired return to capital). More precisely, the credit VaR measure appropriate for EC should consider losses relative to the portfolio's initial mark-to-market (MtM) value and not relative to the *EL* in its end-of-period distribution. Also, credit VaR normally ignores the interest payments that must be made on the funding debt. These payments must be added explicitly to the EC.

Three risk measures are often used for allocating economic capital among a portfolio's constituent positions: *volatility*, *VaR* and *expected shortfall* (sometimes called CVaR or conditional tail expectation). All three measures are homogeneous and hence lead to additive marginal risk contributions.

In current practice, the most common approach for assigning capital on a diversified basis computes a component's marginal contribution to the volatility of the portfolio loss distribution and scales it to correspond to the economic capital (e.g. Smithson, 2003). Specifically, if $\rho(L) \equiv \sigma(L)$ then Eq. (7) leads to the well-known formula

$$C_j^{\sigma} = \frac{\operatorname{cov}(L_j, L)}{\sigma(L)} \tag{9}$$

and the capital charged to obligor j is

$$C_j^{EC_\alpha} = R_j^{\sigma} \times EC_\alpha.$$

This approach works well if losses are normally distributed, since quantiles are constant multiples of the volatility in this case.[7] Due to the non-normality of credit loss distributions, however, volatility allocation often produces inconsistent capital charges (see Praschnik et al., 2001). In particular, Kalkbrener et al. (2004) show that a loan's capital charge can exceed its exposure.

The VaR contribution of obligor j is

$$C_j^{VaR_\alpha} = E[L_j \mid L = VaR_\alpha]. \tag{10}$$

This follows from the relation between partial derivatives and conditional expectations (e.g., Tasche, 1999; Gouriéroux et al., 2000). The capital charged to obligor j is then

$$C_j^{EC_\alpha} = C_j^{VaR_\alpha} - E[L_j]. \tag{11}$$

An obligor's contribution to *EL* is simply its expected loss, which is easy to compute analytically. Thus, capital allocation essentially reduces to the more difficult task of measuring VaR contributions (i.e., Eq. (10)).

[7] More precisely, this is the case for elliptic distributions in general.

It is widely recognized that VaR has several shortcomings since it is not a coherent risk measure (in the sense of Artzner et al., 1999). Specifically, while VaR is sub-additive (or diversifying) for normal distributions, this is not true in general. This limitation is relevant for credit loss distributions, which may be far from normal and not even smooth. In particular, the discreteness of individual credit losses leads to non-smooth profiles and marginal contributions.

Expected shortfall (ES) is a coherent risk measure and presents a good alternative to VaR and volatility both for measuring and allocating capital. As with VaR, ES contributions represent conditional expectations (Tasche, 2002; Scaillet, 2004)

$$C_j^{ES_\alpha} = E[L_j \mid L \geqslant VaR_\alpha]. \tag{12}$$

Since ES acts as a buffer for an expected loss conditional on exceeding a certain quantile, its use in allocating economic capital requires a rescaling similar to that of volatility. That is, the capital charged to obligor j equals

$$C_j^{EC_\alpha} = R_j^{ES_\alpha} \times VaR_\alpha - E[L_j]. \tag{13}$$

Although VaR and ES may not be differentiable in some cases,[8] it is reasonable to define the risk contributions by Eqs. (10) and (12) in general (e.g., Kurth and Tasche, 2003; Hallerbach, 2003).

Kalkbrener et al. (2004) formally introduce an axiomatic approach to define the concept of a *coherent capital allocation*. The three axioms can be summarized as follows:

- *Linear (or additive) allocation*: the capital allocated to a union of sub-portfolios is equal to the sum of the capital amounts allocated to the individual sub-portfolios.
- *Diversifying allocation*: the capital allocated to a sub-portfolio X of a larger portfolio Y never exceeds the risk capital of X considered as a stand-alone portfolio.
- *Continuous allocation*: a small increase in a position only has a small effect on the risk capital allocated to that position.

The authors show that these three axioms uniquely determine a capital allocation scheme – which is essentially a marginal capital allocation. Also, they show that any allocation satisfying these axioms is associated with a coherent risk measure. Notably, ES yields a linear (or additive), diversifying and continuous capital allocation, while VaR yields an additive but not a diversifying allocation.

[8] Laurent (2003) discusses the differentiability of risk measures when the loss distribution is discrete.

4 Credit risk contributions in analytical models

In the presence of diversification, the marginal capital required for a counterparty or loan may depend on the overall portfolio composition. If capital charges are based on marginal portfolio contributions, these charges are not, in general, *portfolio-invariant*, and are different from their stand-alone capital. Thus, an interesting question is: under what circumstances do portfolio models yield portfolio-invariant capital contributions?

If economic capital is defined in terms of a VaR measure, Gordy (2003a) shows that two conditions are necessary and sufficient to guarantee portfolio-invariant contributions:

- The portfolio must be asymptotically fine-grained; i.e. no single exposure can account for more than an arbitrarily small share of total portfolio exposure.
- There must be only a single systematic risk factor.

The "single-factor, asymptotically-fine-grained" portfolio model is at the heart of the new Basel II banking credit regulation (BCBS, 2004). In this context, capital only covers *systemic credit risk*; it does not account for the idiosyncratic risk that exists in non-granular portfolios, leading to counterparty (or name) concentrations. Gordy (2003a, 2003b) and Martin and Wilde (2002) further present an asymptotic approximation for the idiosyncratic credit risk when portfolios are not sufficiently granular (the so called "granularity" adjustment). Of course, in the presence of idiosyncratic credit risk, marginal capital contributions are dependent on the portfolio composition (specifically, the level of name concentration risk in the portfolio).[9]

We now briefly describe the analytical formulae for credit risk contributions in the one-factor, Basel II model and its extension to account for idiosyncratic risk (the so-called granularity adjustment). Finally, we discuss capital allocations in the context of a simple multi-factor extension of the Basel II model. By introducing explicitly the concept of a *diversification factor* at both the portfolio and obligor or sector levels, the multi-factor extended model provides useful intuition on capital contributions, and their sources.

4.1 Capital contributions in the Basel II model

Consider a portfolio with N obligors and a single-step model. Without loss of generality, assume that each obligor j has (unconditional) default probability PD_j, and a single loan with loss given default and exposure at default given by LGD_j, EAD_j respectively.[10]

[9] See for example Emmer and Tasche (2005) for a discussion of risk contributions in a one factor model with the granularity adjustment.

[10] We use here the notation commonly used in the Basel accord, where now the product $EAD_j \cdot LGD_j = c_{j1}$, as given in Section 2. As they are assumed deterministic, they are the same in each scenario m.

For each obligor j, the credit losses at the end of the horizon (e.g., one year) are driven by a Merton model, as given in Section 2, but in this case with one, single, systemic factor. Obligor j defaults when its creditworthiness index falls below a given threshold, given by $\Phi^{-1}(PD_j)$.

The creditworthiness of obligor j is driven by a single systemic factor:

$$Y_j = b_j Z + \sqrt{1 - b_j^2}\varepsilon_j \tag{14}$$

with Z is a standard Normal variable representing the single systemic, economy-wide factor, and the ε_j are independent standard Normal variables representing the idiosyncratic movement of obligors' creditworthiness. We commonly refer to b_j^2 as the asset correlation of obligor j.

Gordy (2003a) shows that the α-percentile systemic portfolio loss (i.e. the loss assuming the portfolio is asymptotically fine-grained), VaR_α, is given by the sum of individual obligor losses, when an α-percentile move occurs in the systemic sector factor Z:

$$VaR_\alpha = \sum_j LGD_j \cdot EAD_j \cdot \Phi\left(\frac{\Phi^{-1}(PD_j) - b_j z^\alpha}{\sqrt{1 - b_j^2}}\right) \tag{15}$$

where z^α denotes the α-percentile of a standard normal variable.

EC is defined to cover only the *unexpected losses* (i.e., Eq. (8)), where the expected losses are $E[L] = \sum_{j=1}^{N} LGD_j \cdot EAD_j \cdot PD_j$.[11] Thus, the capital for the portfolio can be written as

$$EC_\alpha = \sum_{j=1}^{N} C_j^{EC_\alpha} \tag{16}$$

where $C_j^{EC_\alpha}$ denotes the capital contribution of counterparty j:

$$C_j^{EC_\alpha} = LGD_j \cdot EAD_j \cdot \left[\Phi\left(\frac{\Phi^{-1}(PD_j) - b_j z^\alpha}{\sqrt{1 - b_j^2}}\right) - PD_j\right]. \tag{17}$$

The capital contribution in Eq. (17) does not depend on the composition of the rest of the portfolio. In Section 4 we present an example of the allocation produced by this model and discuss the impact of the quantile chosen on the capital allocation.

[11] The following discussion still holds if capital is defined by VaR, by simply adding back the EL at the end of the analysis.

4.2 Capital contributions with idiosyncratic risk (the granularity adjustment)

When there is one systemic factor and the portfolio is (infinitely) granular, the credit portfolio loss distribution is obtained analytically and risk contributions are portfolio-invariant. Idiosyncratic risk arises when the portfolio is of finite size and not homogeneous (i.e. with some counterparty or name concentrations). In this case, even with a one-factor model, a general analytical solution might not be available, and various methods can be used to approximate the loss distribution.

Risk measures (including VaR and ES) can be decomposed into their systemic risk and idiosyncratic contributions. For example, the variance of portfolio losses can be written as the sum of the variance of the conditional expected losses and the expected conditional variance of losses:

$$V[L] = V[E[L|Z]] + E\big[V[L|Z]\big].$$

The first term is the contribution of systemic risk, and the second can be interpreted as the idiosyncratic risk, which vanishes as the number of obligors in a portfolio goes to infinity (and the idiosyncratic risk is diversified away). In a moderately large portfolio, the systemic component may be much larger than the idiosyncratic risk, but the latter may be too large to be neglected.

VaR and ES can be decomposed in a similar manner, although their idiosyncratic components do not have general closed-form expressions. When the conditional variance of losses is small, we can obtain analytical approximations of VaR and ES for non-granular portfolios as "small adjustments" to the infinitely granular portfolio. This *granularity adjustment* method is essentially a second order Taylor series expansion of the quantile (around the "infinitely granular" portfolio).[12]

Denote by VaR_α and $VaR_\alpha^S = VaR_\alpha(E[L|Z])$ the VaR of the (non-granular) portfolio and the systemic VaR of the portfolio (the VaR of the portfolio assuming it is infinitely granular), respectively. The VaR of the portfolio is approximated by:

$$VaR_\alpha \approx VaR_\alpha^S + GA_\alpha. \tag{18}$$

The general formula for the granularity adjustment is

$$GA_\alpha = -\frac{1}{2f(y)}\frac{\partial}{\partial y}\big[\sigma^2(z^\alpha)f(y)\big]\big|_{y=VaR_\alpha^S} \tag{19}$$

where $f(y)$ denotes the density function of the infinitely granular portfolio's loss, and $\sigma^2(z^\alpha)$ is the (idiosyncratic) variance of the portfolio losses conditional on the systemic factor level corresponding to the systemic portfolio losses being equal to VaR_α^S. A similar expression is available for ES.

[12] Gordy (2003a, 2003b) presented this approach first and has then been refined (see for example Martin and Wilde, 2002). Pykhtin (2004) further extended the method to multiple factors.

By applying Eqs. (18) and (19) directly to a given (one-factor) portfolio model, we can obtain closed form expressions to approximate the portfolio VaR and the risk contributions. For the one-factor Merton model, the systemic VaR, VaR_α^S, is given by expression (15) and the granularity adjustment is[13]:

$$GA_\alpha = \sum_{j=1}^{N} C_j^{GA_\alpha} \tag{20}$$

with

$$C_j^{GA_\alpha} = \frac{EAD_j^2 LGD_j^2}{2(VaR_\alpha^S)'}\left[\left(\sqrt{\frac{b_j^2}{1-b_j^2}}\cdot\phi\left(PD_j^\alpha\right)\cdot\left(1-2\Phi\left(PD_j^\alpha\right)\right)\right) + \left(z^\alpha + \frac{(VaR_\alpha^S)''}{(VaR_\alpha^S)'}\right)\cdot\left(\Phi\left(PD_j^\alpha\right)-\Phi\left(PD_j^\alpha\right)^2\right)\right] \tag{21}$$

where $PD_j^\alpha = (\Phi^{-1}(PD_j) - b_j z^\alpha)/\sqrt{1-b_j^2}$, and $(VaR_\alpha^S)'$ and $(VaR_\alpha^S)''$ denote the first and second derivatives of expression (15).

The terms $C_j^{GA_\alpha}$ can be interpreted as the idiosyncratic risk contribution of each obligor j. In this case, contributions are not portfolio invariant, since they depend on the total portfolio composition through the terms $(VaR_\alpha^S)'$ and $(VaR_\alpha^S)''$.

4.3 Credit risk contributions in an extended multi-factor model

A model that yields portfolio-invariant capital contributions is desirable for regulatory purposes, management transparency and computational tractability. However, such a model does not fully recognize diversification and may not be useful for capital allocation. We thus require also tools to understand and measure diversification in a multi-factor portfolio setting. Pykhtin (2004) recently obtains an elegant, analytical multi-factor adjustment to the Basel II one-factor model. This method can also be used effectively to compute capital contributions numerically (given its closed form solution to compute portfolio capital). However, closed-form expressions for capital contributions are quite intricate.

Garcia Cespedes et al. (2006) present a simple model that recognizes the diversification obtained from a multi-factor credit setting. The authors introduce the concept of a *diversification factor* at the portfolio level and also at the obligor or sub-portfolio level to account for diversification contributions to the portfolio (*marginal diversification factors*). Tasche (2006) further presents a

[13] See Emmer and Tasche (2005).

mathematical foundation for the diversification factor and analytical formulae for computing diversification contributions.[14]

To illustrate the Garcia Cespedes et al. model, consider a single-step model with K homogeneous sectors (each of these sectors can represent an asset class or geography, etc.). Similar to the Basel II model, for each obligor j in a given sector k, the credit losses at the end of the horizon are driven by a single-factor Merton model.[15] In this case, however, the creditworthiness of obligor j, in sector k, is driven by a single systemic factor:

$$Y_j = b_k Z_k + \sqrt{1 - b_k^2}\,\varepsilon_j \tag{22}$$

where Z_k is a standard Normal variable representing the systemic factor for sector k, and the ε_j are independent standard Normal variables representing the idiosyncratic movement of an obligor's creditworthiness. While in the Basel II model all sectors are driven by the same systemic factor Z, here each sector can be driven by a different factor.

Assume further that the systemic factors are correlated through a single macro-factor, Z,

$$Z_k = B_k Z + \sqrt{1 - B_k^2}\,\eta_k, \quad k = 1, \ldots, K \tag{23}$$

where η_k are independent standard Normals, and each sector has a different correlation level B_k to the systemic, economy-wide factor, Z.

Assume, as before, that each obligor j has a single loan with loss given default and exposure at default given by LGD_j and EAD_j, respectively. Since credit losses within each sector are driven by a one-factor model, for asymptotically fine-grained sector portfolios, the stand-alone α-percentile capital for a given sector k, $EC_{\alpha,k}$, is given by

$$EC_{\alpha,k} = \sum_{j \in Sector\ k} LGD_j \cdot EAD_j \cdot \left[\Phi\left(\frac{\Phi^{-1}(PD_j) - b_k z^{\alpha}}{\sqrt{1 - b_k^2}} \right) - PD_j \right]. \tag{24}$$

Under Basel II, or equivalently assuming perfect correlation between all the sectors, the overall capital is simply the sum of the stand-alone capital for all individual sectors (for simplicity, we omit the parameter α hereafter from the notation):

$$EC^{1f} = \sum_{k=1}^{K} EC_k. \tag{25}$$

[14] The paper presents a two-dimensional example which has an analytical solution. Problems of dimension N require numerical integration of dimension $N - 1$.

[15] We focus the discussion on a one-period Merton model for default losses. The methodology and results are quite general and can be used with other credit models, and can also incorporate losses due to credit migration, in addition to default.

We define the *capital diversification factor*, DF, as the ratio of the actual capital computed using the multi-factor model and the stand-alone capital, $DF \leqslant 1$. This allows us to express the (diversified) economic capital as:

$$EC = DF \cdot EC^{1f}. \tag{26}$$

Economic capital is thus a function of

- the "additive" bottom-up capital from the one-factor (Basel II) model, EC^{1f}, and
- DF, a "factor adjustment" which represents the diversification of the portfolio.

The basic idea behind the model is to approximate DF, by a scalar function of a small number of parameters, which leads to a reasonable approximation of the true, multi-factor, economic credit capital, and which can be tabulated.

We can think of diversification basically being a result of two sources:

- The relative size of various sector portfolios; clearly a portfolio with one dominating, very large, sector results in high concentration risk and limited diversification. So we seek a parameter representing essentially an "effective number of sectors" accounting for their sizes.
- The cross-sector correlations. Hence a natural choice for a parameter in our model is some form of average cross-sector correlation.

Ideally, the "concentration index" representing the first source of diversification should account for the size of the exposures and also the differences in credit characteristics as they affect capital. Thus, a sector with a very large exposure on highly rated obligors, might not necessarily represent a large contribution from a capital perspective.

The Garcia Cespedes et al. model expresses the economic capital in Eq. (26), for a given confidence level, as

$$EC = DF\big(CDI, \bar{B}^2\big) \cdot \sum_{k=1}^{K} EC_k \tag{27}$$

where the two parameters in the diversification factor are:

- The *capital diversification index*, CDI, given by the sum of squares of the *capital weights* in each sector

$$CDI = \frac{\sum_k EC_k^2}{(EC^{1f})^2} = \sum_k w_k^2 \tag{28}$$

  with $w_k = EC_k / EC^{1f}$ the contribution to one-factor capital of sector k.
- The (capital weighted) average cross-sector correlation: $\bar{B}^2$.

The *CDI* is the well-known Herfindahl concentration index applied to the *stand-alone capital* of each sector (instead of the exposures, *EADs*, as is commonly used). Intuitively, it gives an indication of the portfolio diversification across sectors (not accounting for the correlation between them). For example, in the two-factor case, the *CDI* ranges between 0.5 (maximum diversification) and one (maximum concentration). The inverse of the *CDI* can be interpreted as an "effective number of sectors" in the portfolio, from a capital perspective.

In a similar way, the average correlation parameter is also capital weighted, to account better for the "contributions" of each sector (and accounting for the credit quality in addition to size). From the various possible definitions for an average sector correlation, we choose the following one. Assume a general sector factor correlation matrix, Q (which can be more general than that resulting from Eq. (23), where $Q_{ij} = \beta_i \beta_j,\ j \neq i$), and a vector of portfolio weights $W = (w_1 \dots w_S)^T$. We define the average sector factor correlation as

$$\bar{B}^2 = \frac{\sum_i \sum_{j \neq i} Q_{ij} w_i w_j}{\sum_i \sum_{j \neq i} w_i w_j} = \frac{\sigma^2 - \delta^2}{\vartheta^2 - \delta^2}$$

where $\sigma^2 = W^T Q W$ is the variance of the random variable given by the weighted sum of the factors, $\delta^2 = \sum_i w_i^2$ and $\vartheta^2 = (\sum_i w_i)^2$. $\bar{B}^2$ is an average correlation in the sense that $W^T B W = W^T Q W = \sigma^2$, with B the correlation matrix with all the non-diagonal entries equal to $\bar{B}^2$. For our specific case, we chose the portfolio weights to be the stand alone capital for each sector. Therefore, $\delta^2 = \sum_i EC_i^2$ and $\vartheta^2 = (\sum_i EC_i)^2 = (EC^{sf})^2$.

Garcia Cespedes et al. (2006) calibrate the model (27) through Monte Carlo simulation, and tabulate the diversification factor for several levels of correlation and *CDI* (see Fig. 1).

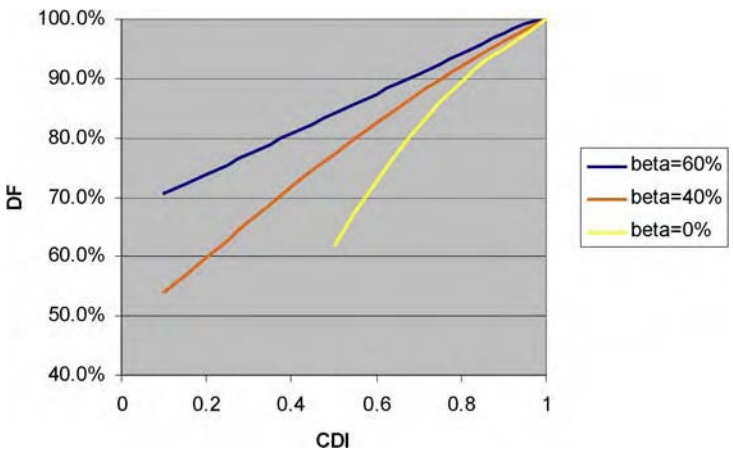


Fig. 1. Calibrated *DF* model in Garcia Cespedes et al. (2006).

Using the diversification factor model, one might be tempted simply to allocate back the diversification effect evenly across sectors, so that the total capital contributed by a given sector is $DF \cdot C_k$. We refer to these as the *unadjusted capital contributions*. This does not account, however, for the fact that each sector contributes differently to the overall portfolio diversification. Instead, we seek a capital decomposition of the form

$$EC^{mf} = \sum_{k=1}^{K} DF_k \cdot EC_k. \tag{29}$$

We refer to the factors DF_k in (29) as the *marginal sector diversification factors*.

For a general model of the form (26), if DF a homogeneous function of degree zero in the EC_k's, Euler's theorem leads to the additive marginal capital decomposition (29) with

$$DF_k = \frac{\partial EC^{mf}}{\partial EC_k}, \quad k = 1, \ldots, K. \tag{30}$$

In the specific model (27), the DF only depends on CDI and $\bar{\beta}$, which are both homogeneous of degree zero. By solving for the derivatives in expression (30), the marginal sector diversification factors are given by

$$\begin{aligned} DF_k = DF &+ 2\frac{\partial DF}{\partial CDI} \cdot \left[\frac{EC_k}{EC^{sf}} - CDI\right] \\ &+ 2\frac{\partial DF}{\partial \bar{B}^2} \cdot \frac{1 - (EC_k/EC^{sf})}{1 - CDI} \cdot \left[\bar{Q}_k - \bar{B}^2\right] \end{aligned} \tag{31}$$

where

$$\bar{Q}_k = \frac{\sum_{j \neq k} Q_{kj} EC_j}{\sum_{j \neq k} EC_j}$$

can be interpreted as the average correlation of sector factor k to the rest of the systemic sector factors in the portfolio.

The marginal capital allocation resulting from the model leads to an intuitive decomposition of diversification effects (or concentration risk) into three components: overall portfolio diversification, sector size and sector correlation:

$$DF_k = DF + \Delta DF_{\text{Size}} + \Delta DF_{\text{Corr}}.$$

The three components represent:

- The overall portfolio DF;
- An adjustment due to the "*relative size*" of the sector to the overall portfolio. Intuitively, for $DF > 0$ and all sectors having the same correlation B^2, a sector with small stand-alone capital ($w_k < CDI$) contributes, on the margin, less to the overall portfolio capital; thus, it gets a higher diversification benefit DF_k;

- An adjustment due to cross-sector correlations. Sectors with lower than average correlation get a higher diversification benefit, as one would expect.

5 Numerical methods to compute risk contributions

Simulation may be required to obtain portfolio loss distributions and calculate risk contributions when the underlying credit model presents a richer co-dependence structure described by multiple systemic factors, when the portfolio contains name concentrations (i.e. it is not granular), when credit losses account for migration and spread risk, or when exposures and LGDs are stochastic (and correlated). We can divide the simulation methods for calculating risk contributions into two broad classes:

- Full Monte Carlo (MC) simulation, with direct sampling of credit events and losses. In this case, the output of the simulation is an independent, identically-distributed sample of size M, where each observation comprises losses for all obligors (and the portfolio loss which is the sum of obligor losses). We make no assumptions about the model that underlies the sample.
- Two stage numerical solution based on the conditional independence framework for credit portfolio models (Section 2). In this case, it is possible to simulate first the systemic factors and then employ various numerical methods to obtain the unconditional portfolio loss distribution. Each systemic scenario comprises the conditional loss distributions for all obligors, with the conditional portfolio loss distribution as the convolution of these losses. As noted earlier, conditional portfolio loss distributions may be obtained using various techniques.

We now briefly summarize the application of these methods to compute credit risk contributions for VaR and ES.

5.1 Monte Carlo simulation with direct sampling of credit events

In a direct simulation approach, VaR and ES are estimated from the order statistics of the sampled portfolio losses.[16] Given the extreme quantiles typically used to measure credit risk, obtaining accurate risk contributions is a challenging task since the conditional expectations (Eqs. (10) and (12)) depend on rare events. This is of particular concern for VaR since the contribution is conditional on a single level of loss, while the ES contribution is conditional on a range of losses. Thus the accuracy of the VaR contributions depends critically on the chosen quantile estimator. In particular, the sample quantile, which is

[16] Order statistic k is the kth smallest loss.

frequently used in practice, is poorly suited for this task since it relies on a single order statistic. In contrast, L-estimators yield more robust estimates of VaR contributions.

5.1.1 Sample quantile estimators

Consider estimating a portfolio's VaR and ES at the 95% level, from an independent MC sample of size 100. In practice, the 95% VaR is often taken to equal the sample quantile (i.e., the 96th order statistic $L^{(96)}$), since $\mathrm{P}(L \geqslant L^{(96)}) = 0.05$ for the sample. A corresponding estimate of the 95% ES is given by the arithmetic average of order statistics 96 through 100.

The sample quantile, as defined in the example above, corresponds to an estimator known as the upper empirical cumulative distribution value (UECV). More generally, in a sample of size M, the UECV estimator estimates the α-quantile of the loss distribution by

$$\overline{VaR}_\alpha = L^{(\lfloor M\alpha \rfloor + 1)},$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x, and the ES at level α by

$$\overline{ES}_\alpha = \frac{1}{M(1-\alpha)}\left[\big(\lfloor M\alpha \rfloor + 1 - M\alpha\big) L^{(\lfloor M\alpha \rfloor + 1)} + \sum_{k=\lfloor M\alpha \rfloor + 2}^{M} L^{(k)} \right].$$

Since the portfolio loss in any particular scenario is given by the sum of the obligor losses, in the example above, the contribution of obligor j to the 95% VaR is estimated by $L_j^{(96)}$ (i.e., the loss of obligor j that occurs with the 5th largest portfolio loss). Its contribution to the 95% ES is estimated by averaging the losses of obligor j that occur with the five largest portfolio losses.

From Eq. (10), the VaR contribution of obligor j equals the average loss of obligor j when the sampled portfolio loss equals the VaR estimate. Formally, if $L^{(k)} = \overline{VaR}_\alpha$ for $k_\alpha^{\min} \leqslant k \leqslant k_\alpha^{\max}$ then

$$\overline{C}_j^{VaR_\alpha} = E\big[L_j \mid L = \overline{VaR}_\alpha\big] = \frac{1}{k_\alpha^{\max} - k_\alpha^{\min} + 1} \sum_{k=k_\alpha^{\min}}^{k_\alpha^{\max}} L_j^{(k)}.$$

If only one sampled portfolio loss equals the estimated VaR (as is typically the case in a MC simulation) then $k_\alpha^{\min} = k_\alpha^{\max} = \lfloor M\alpha \rfloor + 1$ and the VaR contribution is estimated from a single observation, namely

$$\overline{C}_j^{VaR_\alpha} = L_j^{(\lfloor M\alpha \rfloor + 1)}.$$

Since a given portfolio loss may result from numerous combinations of obligor losses, estimates of VaR contributions provided by the sample quantile are often unreliable (e.g., Mausser and Rosen, 1998; Mausser, 2003). Two approaches for improving these estimates are:

- use quantile estimators that average multiple order statistics (e.g., L-estimators);
- use sampling strategies that increase the number of observations with a portfolio loss of $\overline{VaR}_\alpha$ (e.g., importance sampling).[17]

In contrast, the estimated ES contributions

$$\overline{C}_j^{ES_\alpha} = \frac{1}{M(1-\alpha)}\left[\left(k_\alpha^{\max} - M\alpha\right)\overline{C}_j^{VaR_\alpha} + \sum_{k=k_\alpha^{\max}+1}^{M} L_j^{(k)}\right]$$

tend to be more robust because ES is, by definition, an average of multiple order statistics. Nevertheless, both L-estimators and importance sampling can be applied to refine estimated ES contributions.

5.1.2 *L-estimators*

To improve the quality of the VaR contributions, several authors have proposed computing a weighted average of the losses over a range of order statistics around the sample quantile (e.g., Praschnik et al., 2001; Hallerbach, 2003). Conceptually, this is consistent with a more general class of quantile estimators known as L-estimators.

An L-estimator (e.g. Sheather and Marron, 1990) computes a quantile estimate as a weighted average of multiple order statistics.[18] Specifically, in a sample of size M, VaR_α is estimated as

$$\overline{VaR}_\alpha = \sum_{k=1}^{M} w_{\alpha,M,k}^{VaR} L^{(k)}, \tag{32}$$

where the weights depend only on the VaR level and the sample size. The VaR contribution for obligor j is then estimated as[19]

$$\overline{C}_j^{VaR_\alpha} = \sum_{k=1}^{M} w_{\alpha,M,k}^{VaR} L_j^{(k)}. \tag{33}$$

To estimate ES_α, observe that

$$ES_\alpha = E\left[F^{-1}(p) \mid p \geqslant \alpha\right]$$

[17] As pointed out in Glasserman (2005), it may be necessary to define a small "window" around $\overline{VaR}_\alpha$ in order to obtain a sufficiently large sample for estimating the conditional expectation.

[18] It is interesting to note the similarities between L-estimators and spectral risk measures (Acerbi, 2002). The quantity computed by the L-estimator is a spectral risk measure (i.e., coherent) if the weights are non-negative, non-decreasing (with respect to loss size) and sum to one.

[19] If a loss occurs multiple times in the sample, then its total weight is distributed equally among all relevant order statistics. For example, if $L_{(j)} = L_{(j+1)}$ for some j, then we set the weights for both order statistics equal to $\frac{1}{2}(w_{\alpha,S,j}^{VaR} + w_{\alpha,S,j+1}^{VaR})$ in Eq. (33).

$$= \frac{1}{1-\alpha} \int_{\alpha}^{1} F^{-1}(p)\, \mathrm{d}p. \tag{34}$$

Replacing $F^{-1}(p)$ in Eq. (34) by its estimate from Eq. (32) yields

$$\begin{aligned} \overline{ES}_{\alpha} &= \frac{1}{1-\alpha} \int_{\alpha}^{1} \left(\sum_{k=1}^{M} w_{p,M,k}^{VaR} L^{(k)} \right) \mathrm{d}p \\ &= \sum_{k=1}^{M} \left(\frac{1}{1-\alpha} \int_{\alpha}^{1} w_{p,M,k}^{VaR}\, \mathrm{d}p \right) L^{(k)}. \end{aligned} \tag{35}$$

Equation (35) defines an L-estimator for expected shortfall with weights

$$w_{\alpha,M,k}^{ES} = \frac{1}{1-\alpha} \int_{\alpha}^{1} w_{p,M,k}^{VaR}\, \mathrm{d}p.$$

The ES contribution for obligor j is estimated as

$$\overline{C}_{j}^{ES_{\alpha}} = \sum_{k=1}^{M} w_{\alpha,M,k}^{ES} L_{j}^{(k)},$$

with the weights again adjusted for duplicate losses, if any (see footnote 19).

An L-estimator that has been found to perform well in practice is due to Harrell and Davis (1982). Empirical evidence suggests that the Harrell–Davis (HD) estimator outperforms the sample quantile for VaR contributions (Mausser, 2003; Mausser and Rosen, 2005). Appendix A derives the HD estimator weights for VaR and ES.

It is useful to compare Eq. (10), which defines a VaR contribution as a conditional expectation, and Eq. (33), which expresses it as weighted sum of ordered statistics. This suggests an intuitive interpretation of the weights in the L-estimator: for a desired quantile, they reflect the estimated probabilities that each order statistic equals the actual VaR, i.e.,

$$\overline{C}_{j}^{VaR_{\alpha}} = \sum_{k=1}^{M} \Pr\left[L^{(k)} = VaR_{\alpha}\right] E\left[L_{j} \mid L = L^{(k)}\right].$$

In fact, Sheather and Marron (1990) point out that the HD estimator is actually the bootstrap estimator of $\mathrm{E}[L^{((M+1)\alpha)}]$, where the expectation is computed analytically rather than by resampling.

For extreme quantiles of the portfolio loss distribution, standard MC may not generate enough observations in the tail to estimate VaR or ES contributions accurately, regardless of the quantile estimator used. Further improvements in the accuracy of risk contributions can be achieved by a combination of

- Exploiting the structure of the model and taking advantage of the underlying conditional independence framework;
- Applying variance reduction techniques such as Importance Sampling, Control Variates or Quasi MC methods.

5.2 Risk contributions in conditional independence models

Consider the conditional independence framework described in Section 2. We can express risk contributions as follows. Let the random variable

$$\bar{L}_j^m = L^m - L_j^m$$

denote the combined loss of all obligors other than j in scenario m. Since L_j^m and $\bar{L}_j^m$ are independent, the VaR contribution of obligor j in scenario m is

$$E\big[L_j^m \mid L^m = VaR_\alpha\big] = \frac{\sum_{r=1}^{R} c_{jr}^m p_{jr}^m \Pr[\bar{L}_j^m = VaR_\alpha - c_{jr}^m]}{\Pr[L^m = VaR_\alpha]}. \tag{36}$$

The unconditional contribution of obligor j is then computed as follows

$$E[L_j \mid L = VaR_\alpha] = \sum_{m=1}^{M} E\big[L_j^m \mid L^m = VaR_\alpha\big] \Pr\big[Z^m \mid L = VaR_\alpha\big]. \tag{37}$$

Expected shortfall contributions are obtained similarly by conditioning on the loss being greater than or equal to the VaR in Eqs. (36) and (37). That is, the probabilities in the numerator and denominator of Eq. (36) are substituted by

$$\Pr\big[\bar{L}_j^m \geqslant VaR_\alpha - c_{jr}^m\big] \quad \text{and} \quad \Pr\big[L^m \geqslant VaR_\alpha\big]. \tag{38}$$

Equation (36) shows that the conditional contribution for obligor j essentially entails convoluting the distributions of L_j^m and $\bar{L}_j^m$. As with the computation of the conditional portfolio loss distribution (Eq. (4)), various numerical methods may be used to approximate this convolution efficiently. For example, Saddle Point methods provide semi-analytical expressions for VaR and ES contributions (see Martin et al., 2001). Also, by assuming that conditional portfolio losses are roughly Normal and applying the Central Limit Theorem, we can obtain analytical formulae for risk contributions. Appendix A further presents the analytical expressions for VaR and ES contributions under the CLT.

It is important to emphasize that the objective of these methods is essentially to capture the idiosyncratic risk (which arises from the conditionally independent obligor losses in each scenario). On their own, these methods are generally not effective for risk contributions of very large and granular portfolios (where the systemic scenarios are largely driving the portfolio losses) or when capital is calculated at high quantiles in the tail. This requires a greater emphasis on generating "relevant" scenarios on the systemic factors, and hence MC variance reduction methods can provide significant improvements.

5.3 Variance reduction techniques

Variance reduction techniques can be used to improve significantly the quality of risk contribution estimates, particularly for extreme quantiles. In particular, Importance Sampling can be used for simulating both systemic and specific risk factors to have more relevant scenarios in the tail of the distribution. The following list represents several examples of its application to estimating risk contributions:

- Merino and Nyfeler (2004) compute ES contributions in a default-only model, using importance sampling to estimate the probabilities in Eq. (38). Their approach requires first obtaining VaR_α for the desired α-quantile of the unconditional portfolio loss distribution. Then, if VaR_α exceeds $E[L^m]$, they adjust each conditional default probability p_{j1}^m (by means of a so-called "exponential twisting") so that the expected value of L^m under the adjusted probability measure equals VaR_α.[20] Thus, importance sampling is applied to the specific risk factors in each systemic scenario.
- Kalkbrener et al. (2004) also compute ES contributions in a default-only model, but apply importance sampling to the systemic risk factors. That is, they sample $Z_1, \ldots, Z_d$ from Normal distributions whose means are shifted to increase the likelihood of an extreme loss. The conditional expectation (see Eq. (36)) is effectively estimated based on a single sample from the conditional portfolio loss distribution associated with each systemic scenario (i.e., the entire simulation consists of M samples, where M is the number of systemic scenarios).
- Glasserman (2005) applies IS jointly to the systemic factors (shifting both their means and covariances) and specific factors (exponential tilting) to compute VaR and ES contributions. He also derives an analytical approximation that, instead of sampling from the shifted distributions, computes the conditional expectation directly.

6 Case studies

We now present several examples that demonstrate the use of marginal risk contributions for capital allocation and highlight some of the practical issues involved. Specifically, the intent is to illustrate the key properties of risk contributions, the management implications of using various risk measures, and the related numerical issues. We consider the following cases:

[20] Although they consider a default only model, the exponential twisting can be generalized for credit migration losses as well.

- The first example shows the behavior of VaR and ES contributions obtained using a one-factor model and a granular portfolio. It demonstrates that the choice of the quantile can have a significant impact on the capital allocation.
- The second example analyzes the impact of diversification from a multi-factor model on the portfolio capital and the capital contributions. It shows the sensitivity of the marginal allocations to the size of their components and the level of diversification.
- The third example analyzes an international bond portfolio with simulation. It shows how the discrete nature of the issuer credit losses makes it difficult to compute risk contributions and uses L-estimators to mitigate these effects.
- The final example compares the risk contributions of the bond portfolio based on volatility, VaR and ES measures. We show how volatility-based contributions can lead to an inefficient allocation of capital and discuss management implications.

6.1 Risk contributions in a one-factor credit model – impact of quantile

The risk measure used to define capital and measure risk contributions can have a significant impact on capital allocation decisions. We now illustrate the sensitivity of the capital allocation to the confidence level (quantile) when using VaR-type measures. Similar effects are observed when using expected shortfall.

Consider, as first example, the credit portfolio described in Table 1. It consists of ten homogeneous pools or sectors, each containing a very large number of obligors. The portfolio weights are uniform, with each sector contributing to 10% of the total exposure. Without loss of generality, we apply 100% LGD to all sectors. We model portfolio losses using a one-factor Merton model, and assign uniformly an asset correlation of 15% (this is consistent, for example, with mortgage portfolios in Basel II). For modeling purposes we assume that the portfolio is infinitely granular, and only susceptible to systemic risk.

The expected losses in the portfolio are 3.5% of the total exposure and, given that sectors are all the same size, their *EL* contributions are proportional to their PD. The VaR losses of this portfolio are obtained through the closed form expression (15) and the sector contributions are portfolio invariant. The 99.9% portfolio losses are over 19% and the total capital is just short of 16%. The first four sectors contribute to almost 80% of the VaR.[21] Three sectors contribute to almost 86% of EL and 72% of VaR.

Figure 2 shows the portfolio losses in the tail of the distribution. The maximum loss at a confidence level of 100% is the total exposure of 100. The figure

[21] We focus hereon on the VaR contributions, which also include *EL* contributions, but similar conclusions can be drawn for capital (as defined by unexpected losses only) contributions.

Table 1.
Portfolio Description (Uniform Exposures)

Sector	EAD	LGD	PD	Corr	EL	VaR (99.9%)
1	10	100%	11.00%	0.15	31.3%	25.2%
2	10	100%	10.00%	0.15	28.4%	24.0%
3	10	100%	9.00%	0.15	25.6%	22.7%
4	10	100%	2.00%	0.15	5.7%	9.1%
5	10	100%	1.50%	0.15	4.3%	7.5%
6	10	100%	1.00%	0.15	2.8%	5.7%
7	10	100%	0.30%	0.15	0.9%	2.4%
8	10	100%	0.20%	0.15	0.6%	1.8%
9	10	100%	0.10%	0.15	0.3%	1.0%
10	10	100%	0.05%	0.15	0.1%	0.6%
Total	100				3.5	19.3

Capital = 15.8.

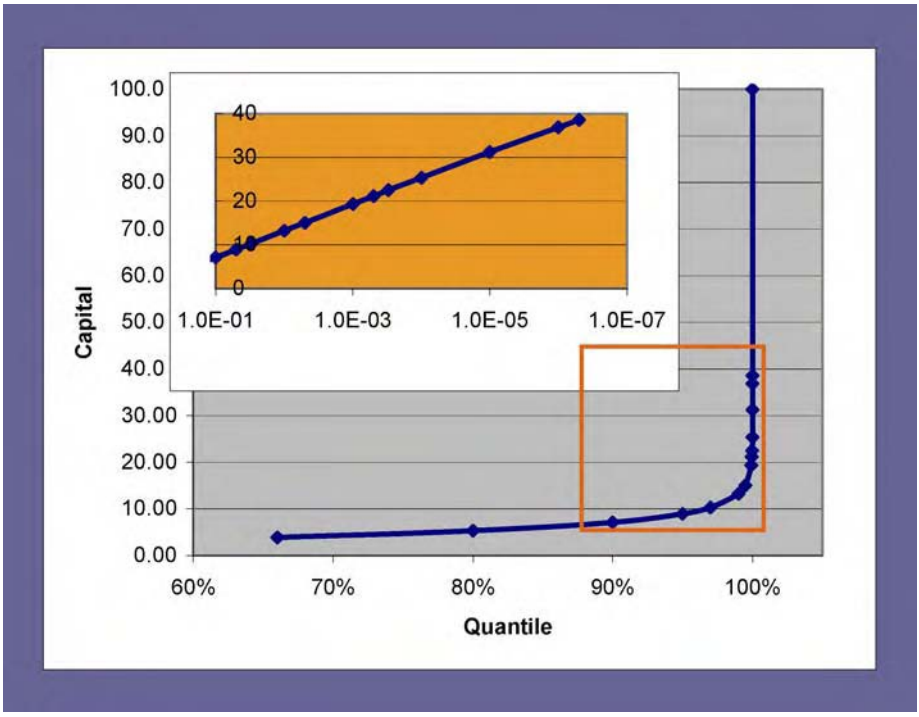


Fig. 2. Tail of Portfolio Loss Distribution (uniform exposures).

further zooms in on the tail from the 90%–99.99995%, on a log scale (where an exponential law describes well the losses in that range of the tail).

The quantile chosen can have a substantial impact on the capital allocation. This is illustrated in Fig. 3, which gives the VaR contributions as functions of the quantile (and tabulates these contributions for several quantiles). We can make the following observations from Fig. 3:

- Since all counterparties have equal exposure, at a confidence level of 100%, every sector contributes to one tenth of the losses, regardless of their credit quality. PD influences loss exposures however at all other confidence levels.

	Quantile					
	90%	99%	99.9%	99.99%	99.999%	100%
1	30.1%	27.4%	25.2%	23.3%	21.7%	10%
2	27.7%	25.7%	24.0%	22.4%	21.0%	10%
3	25.3%	24.0%	22.7%	21.4%	20.2%	10%
4	6.4%	8.0%	9.1%	10.0%	10.6%	10%
5	4.9%	6.4%	7.5%	8.5%	9.2%	10%
6	3.3%	4.6%	5.7%	6.6%	7.4%	10%
7	1.0%	1.7%	2.4%	3.1%	3.8%	10%
8	0.7%	1.2%	1.8%	2.3%	2.9%	10%
9	0.3%	0.7%	1.0%	1.5%	1.9%	10%
10	0.2%	0.4%	0.6%	0.9%	1.2%	10%
1/CDI	4.2	4.7	5.2	5.8	6.2	10.0

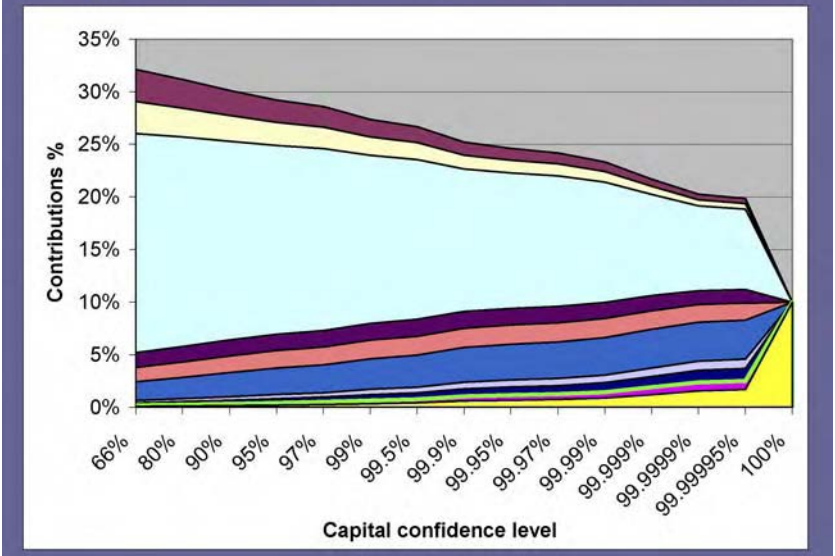


Fig. 3. Risk Contributions in the Tail of the Distribution (uniform exposures).

- At the 66% level, the three sectors with the highest PD (lowest credit quality) contribute 87% of the VaR. This goes down to 72% at a 99.9% level (the one used in Basel II) and to less than 63% for a 99.999% level.
- In general, as the quantile increases, the capital attributed to the low quality sectors is shifted to high quality sectors. For example, the shift in quantile from 99% to 99.99% results almost an almost 10% capital reallocation from low to high quality sectors.

The table also gives the inverse of the CDI (as defined in Eq. (28)), which essentially gives an "effective number of sectors" as accounted for their capital contributions. This is a simple summary measure that allows us to see the overall impact of the quantile on the capital allocation. At a 90% level the portfolio shows 4.2 effective sectors. This increases to 5.2 at the 99.9% level, 6.2 at 9.999% and 10 at 100% (which is the number of effective sectors as seen from an exposure perspective).

In this example, only the credit quality (PD) was varied across sectors. While the risk attributed to different sectors changed with the quantile, the ranking of sectors by risk contribution remains the same under all quantiles. This is not the case, however, when sectors vary across other dimensions as well. Consider now the portfolio in Table 2.

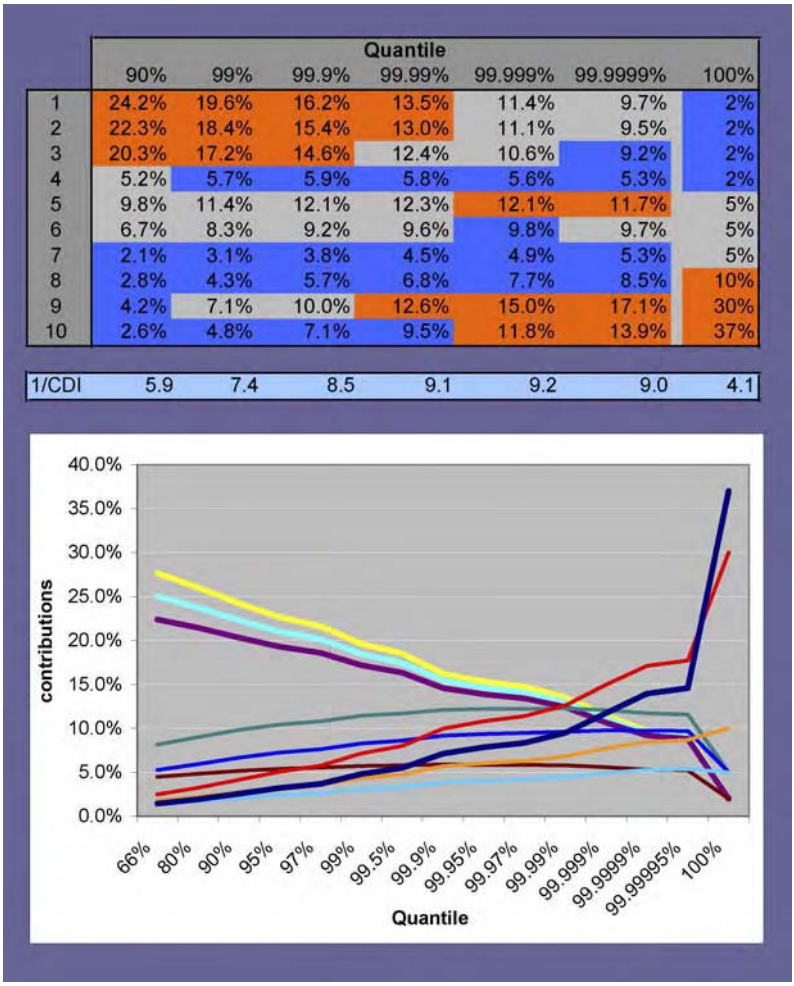
The total exposure and the distribution of PDs are the same as in the previous case (as are the losses at 100% level). However, both EL and capital at the 99.9% level are much smaller, since the exposures are in this case distributed proportionately to the credit quality (as is often the case in balanced portfolios).

The impact of the quantile chosen on the capital allocation is more complex in this case, due to the opposing effects of the distributions of credit quality and

Table 2.
Portfolio Description (Non-uniform Exposures)

Sector	EAD	LGD	PD	Corr	EL	VaR (99.9%)
1	2	100%	11.00%	0.15	25.9%	16.2%
2	2	100%	10.00%	0.15	23.6%	15.4%
3	2	100%	9.00%	0.15	21.2%	14.6%
4	2	100%	2.00%	0.15	4.7%	5.9%
5	5	100%	1.50%	0.15	8.8%	12.1%
6	5	100%	1.00%	0.15	5.9%	9.2%
7	5	100%	0.30%	0.15	1.8%	3.8%
8	10	100%	0.20%	0.15	2.4%	5.7%
9	30	100%	0.10%	0.15	3.5%	10.0%
10	37	100%	0.05%	0.15	2.2%	7.1%
Total	100				0.85	6.0

Capital = 5.2.



	Quantile						
	90%	99%	99.9%	99.99%	99.999%	99.9999%	100%
1	24.2%	19.6%	16.2%	13.5%	11.4%	9.7%	2%
2	22.3%	18.4%	15.4%	13.0%	11.1%	9.5%	2%
3	20.3%	17.2%	14.6%	12.4%	10.6%	9.2%	2%
4	5.2%	5.7%	5.9%	5.8%	5.6%	5.3%	2%
5	9.8%	11.4%	12.1%	12.3%	12.1%	11.7%	5%
6	6.7%	8.3%	9.2%	9.6%	9.8%	9.7%	5%
7	2.1%	3.1%	3.8%	4.5%	4.9%	5.3%	5%
8	2.8%	4.3%	5.7%	6.8%	7.7%	8.5%	10%
9	4.2%	7.1%	10.0%	12.6%	15.0%	17.1%	30%
10	2.6%	4.8%	7.1%	9.5%	11.8%	13.9%	37%
1/CDI	5.9	7.4	8.5	9.1	9.2	9.0	4.1

Fig. 4. Risk Contributions in the Tail of the Distribution (non-uniform exposures).

exposure sizes. This is shown in Fig. 4. We can make the following observations on this portfolio:

- The three sectors with the lowest creditworthiness are the biggest consumers of capital at lower confidence levels. At a 66% level they account for 75% of losses. This number goes down to 55% and then to 39%, at 99% and 99.99% levels respectively. At high confidence levels, these sectors are not even the highest contributors. At a 100% level, the three biggest sectors (which are also the best credits) account for 77% of the capital.

- The ranking of the sectors based on their capital consumption changes with the quantile. This can be seen by the lines intersecting in the graph in Fig. 4 and also from the adjacent table.
- The effective number of sectors (inverse of the CDI) is not a monotonic function of the quantile. The effective number of sectors reaches its peak around the 99.999% level and then goes down again to slightly over four at the 100% level.

We can think of the CDI as a measure of the "dispersion" of the risk contribution profiles in the plot at a given quantile. Thus the minimum occurs at the quantile where the dispersion is the smallest. This is basically the quantile at which the portfolio looks the most diversified. If all the capital contributions are the same for each sector, the effective number of sectors is the maximum and coincides with the actual number of sectors (in this case 10).

6.2 The diversification factor and capital contributions

Consider again the portfolio in Table 1, in the previous section, consisting of ten homogeneous, granular, sectors with uniform weights (each contributing 10% of the total exposure). Using a one-factor credit model, the first three sectors (with high PDs) contribute over 72% of the losses at the 99.9% level, and almost 69% of the capital (given their high *EL* contributions).

Now, assume that the portfolio is driven by the multi-factor model given by Eqs. (22) and (23). Each sector is driven by a single factor, with inter-sector correlation of 15% as in the previous example. Furthermore, assume that all sectors have the same correlation level to the single systemic factor (Eq. (23)). Figure 5 summarizes the capital and allocations, assuming intra-sector correlation levels of 100%, 60% and 40%.

The capital diversification index (*CDI*) for the portfolio is 0.18, which implies 5.6 effective sectors (the Herfindahl index on the exposures is 0.1). The one factor model corresponds to the case when all the sectors have correlation of 100% and hence the diversification factor (*DF*) is 100%. Correlations of 60% and 40%, results in 27% and 60% lower capital, respectively (*DF*s of 73.2% and 40%).

The last three columns (and the graph) give the capital allocations for the different correlation values. In the one factor model (100% correlation), each sector contributes its stand-alone capital. In the presence of diversification, Eq. (27) shows that each sector's *marginal diversification factor*, DF_k, depends also on the relative size of the sector (in terms of its stand-alone capital) and the relative intra-sector correlation (which in this example is the same for all sectors). Smaller sectors contribute more to the overall diversification and get percentage capital allocations smaller than their corresponding stand-alone contributions. The bigger portfolios, in contrast, get bigger contributions (percent-wise). This effect grows with the level of diversification, as can be seen from the figure (i.e. the smaller the correlation the higher this effect). Thus,

Positions	EAD	EL%	VaR %	Capital % Corr=100%	Corr=60%	Corr=40%
1	10%	31.3%	25.2%	23.9%	25.1%	26.7%
2	10%	28.4%	24.0%	23.0%	24.0%	25.3%
3	10%	25.6%	22.7%	22.0%	22.8%	23.8%
4	10%	5.7%	9.1%	9.9%	9.2%	8.3%
5	10%	4.3%	7.5%	8.3%	7.5%	6.6%
6	10%	2.8%	5.7%	6.3%	5.7%	4.8%
7	10%	0.9%	2.4%	2.7%	2.4%	1.9%
8	10%	0.6%	1.8%	2.0%	1.7%	1.4%
9	10%	0.3%	1.0%	1.2%	1.0%	0.8%
10	10%	0.1%	0.6%	0.7%	0.6%	0.5%
Total	100	3.52	19.33	15.81	11.58	9.28

CDI	0.180
Sectors	5.6

DF	100%	73.2%	40.0%
Slope		0.34	0.59

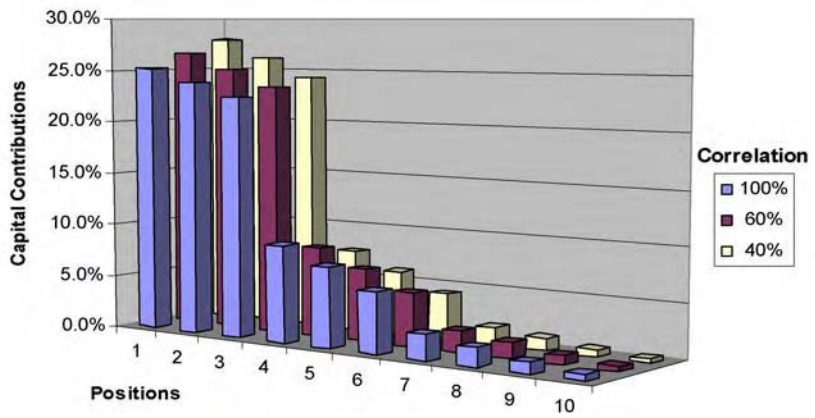


Fig. 5. Capital Contributions in a Multi-Factor Model (uniform exposures portfolio).

while the three largest portfolios contribute 69% of the capital in a one-factor model, their percent contribution grows to almost 76%.

6.3 VaR contributions, discreteness of loss distributions and L-estimators

We now analyze the credit risk of a portfolio of emerging markets debt under a multi-factor model. The portfolio, comprising 197 long-dated corporate and sovereign bonds issued by 86 obligors, has a mark-to-market value of 8.3 billion USD and a duration of approximately five years. The credit model considers both default and MtM losses. Credit migrations for each obligor occur among eight possible credit states, including a terminal default state, with transition probabilities specified by a Standard & Poor's transition matrix. The

co-dependence structure is defined by a multi-factor model of asset returns (for details, see Bucay and Rosen, 1999). For illustration purposes only, and to keep the example simple, we compute the portfolio credit loss distribution using a Monte Carlo simulation with 20,000 scenarios.[22]

Figure 6 shows the VaR and ES for a range of quantile levels, as obtained by the UECV and HD estimators.[23] Both estimators yield virtually identical results.

When applied to risk contributions, the UECV and HD estimators give similar results for ES but not for VaR. For example, Fig. 7 shows the risk attributed to Brazilian debt. While the HD estimator consistently identifies Brazil as a significant source of risk under both measures, the VaR contributions produced by the UECV estimator are erratic and, in fact, frequently fail to attribute any risk to Brazil.

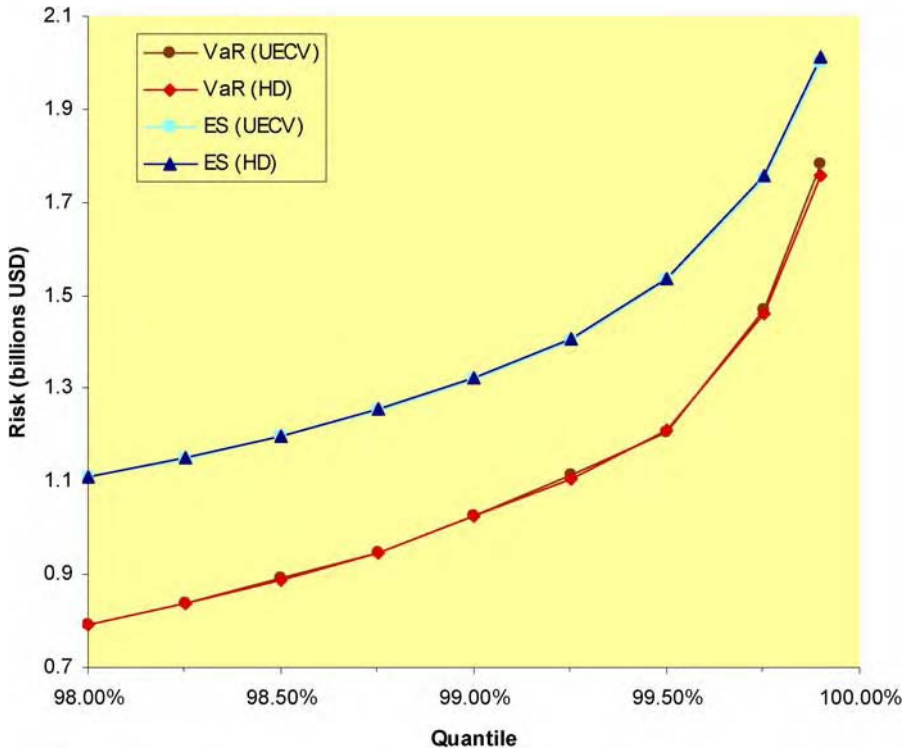


Fig. 6. Risk of Bond Portfolio.

[22] The scenario set might be relatively small for estimating accurately extreme tails (e.g. 99.9%). In practice, one would use a larger number of scenarios, or enhanced techniques such as Quasi-MC methods or importance sampling.

[23] Weights less than 10^{-6} are set to zero in our analysis.

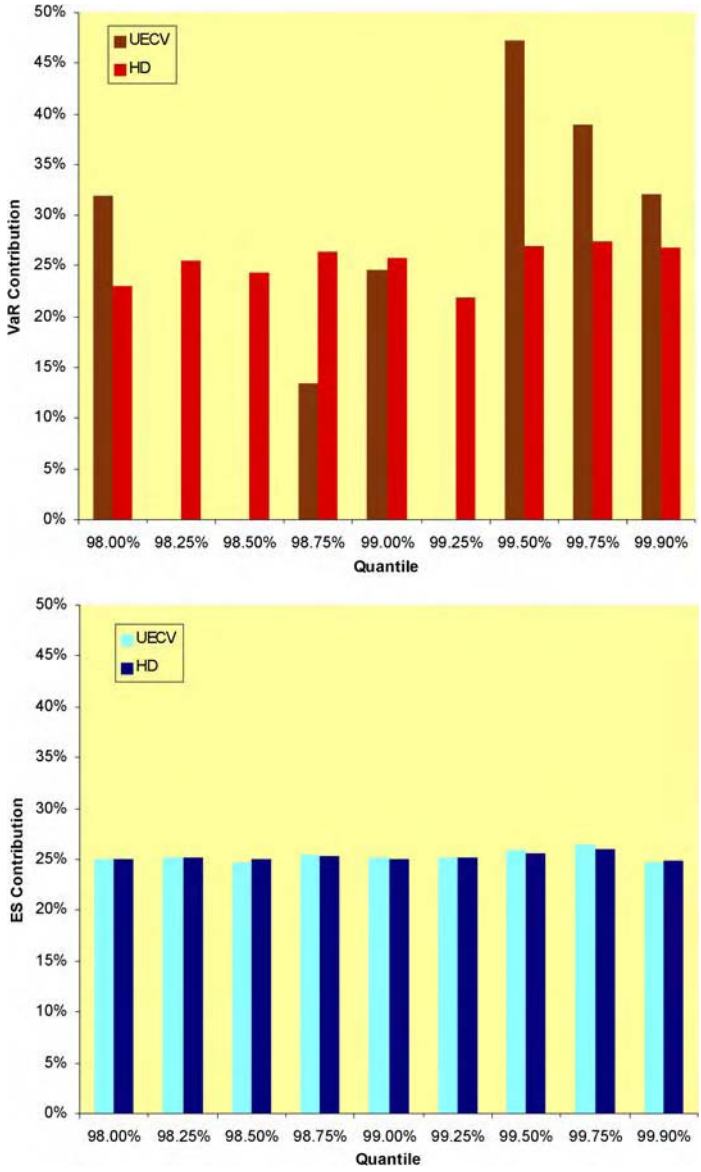


Fig. 7. Risk Contributions for Brazil.

Figure 8 shows the 400 largest portfolio losses in the sample (i.e., the tail of the empirical portfolio loss distribution beyond the 98% quantile) and the component losses due to Brazilian bonds. While the portfolio loss profile is relatively smooth, the Brazilian losses often change drastically from one order statistic to the next. (Note that under the chosen model, an obligor incurs one

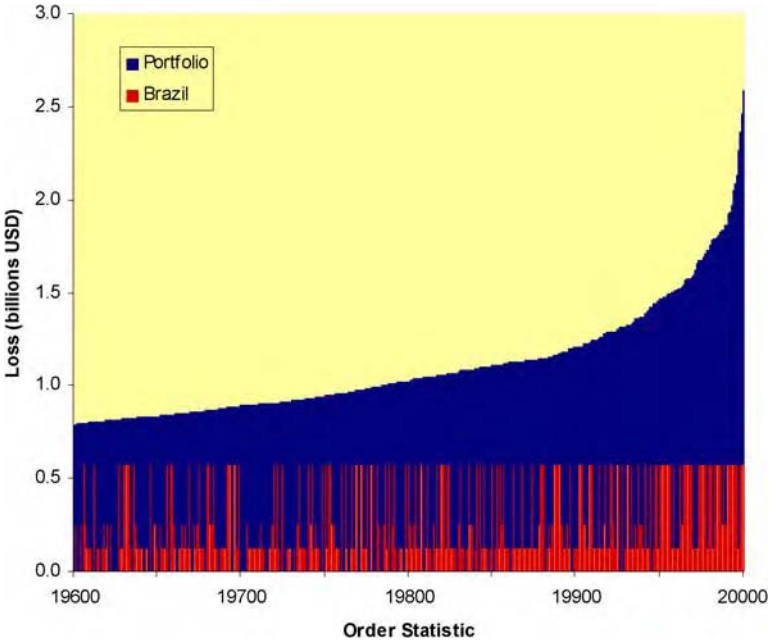


Fig. 8. Brazilian Losses Relative to Portfolio Losses.

of eight possible losses, corresponding to the eight possible credit states, in each scenario.)

In practice, given a sufficiently large sample, a quantile estimator that is based on a single order statistic (e.g., UECV) produces a robust portfolio VaR estimate. Moreover, since the smoothness of the portfolio loss distribution increases with sample size, the accuracy of the VaR estimate improves accordingly. In contrast, UECV estimates of the VaR contributions are unreliable and, since the smoothness of the obligor loss profile is unaffected by sample size, their accuracy cannot necessarily be improved by increasing the number of scenarios.

This property has significant implications for typical credit risk models, in which a loss is triggered by an obligor's default or, more generally, by its transition to a lower credit grade. Since there is a relatively high probability of an individual obligor retaining its original credit rating, many obligors do not incur a loss in a given scenario. As a result, the UECV estimator tends to report an excessive number of zero VaR contributions.

In contrast, ES is defined as an average of the losses spanning a range of order statistics. Since averaging has the effect of smoothing the obligor loss profile, the UECV estimator is generally reliable for ES contributions.

The HD estimator applies a similar averaging approach to VaR estimation. While this improves the stability of the HD VaR contributions, they still show greater variability than the ES contributions (e.g., the dip at the 99.25%

level in Fig. 7). This is due to the different weighting schemes: the ES weights are distributed more or less equally among order statistics beyond the sample quantile, while the VaR weights are focused on a smaller "window" of order statistics around the sample quantile. Since larger samples result in the weights extending over a greater number of order statistics, increasing the sample size produces more reliable estimates of both VaR and ES contributions.

Figure 9 illustrates the estimation of Brazil's contribution to ES and VaR, respectively, at the 99.25% level (the contribution is denoted by the large icon above order statistic 19,850 in each case). The graphs show the relative contribution of Brazil to each of the 400 largest portfolio losses (i.e., the Brazilian loss component divided by the total portfolio loss), shaded to reflect the size of its corresponding estimator weight. The ES contribution is essentially a weighted average of the loss contributions associated with the 200 largest portfolio losses. In contrast, the VaR estimation weights are distributed among order statistics 19,790 through 19,900, with the largest weights surrounding order statistic 19,850.

6.4 Comparing quantile-based and volatility contributions

Assigning capital based on volatility allocation is problematic since an obligor's contribution to volatility may fail to represent the tail of the distribution (e.g., Praschnik et al., 2001; Kurth and Tasche, 2003; Kalkbrener et al., 2004). It is important to understand that risk contributions vary across measures, and also across different confidence levels. This is apparent in Fig. 10, which plots, for the top six obligors, the ranges of the tail-based (VaR and ES) risk contributions, for quantile levels between 98 and 99.9%, against volatility contributions. For example, Brazil contributes between 21.9 and 27.4% of the VaR and between 24.8 and 26.0% of the ES, but accounts for only 20.6% of the volatility.

In this case, the tail-based risk contributions consistently exceed the volatility contribution for the two largest contributors, Brazil and Russia. Moreover, the rankings of Russia and Venezuela are reversed when based on volatility. Also, the range of VaR contributions is typically larger than that for ES. As discussed previously, this might also reflect the relative lack of precision (i.e., a greater sensitivity to random error) on the part of the HD estimator in the former case.

7 Summary and further research

Capital allocation is an important management decision support and business planning tool for financial institutions, which is required for pricing, profitability assessment and limits, building optimal risk-return portfolios and strategies, performance measurement and risk based compensation. This chapter provides a practical overview of the measurement of economic credit capital contributions and their application to capital allocation. We discuss the

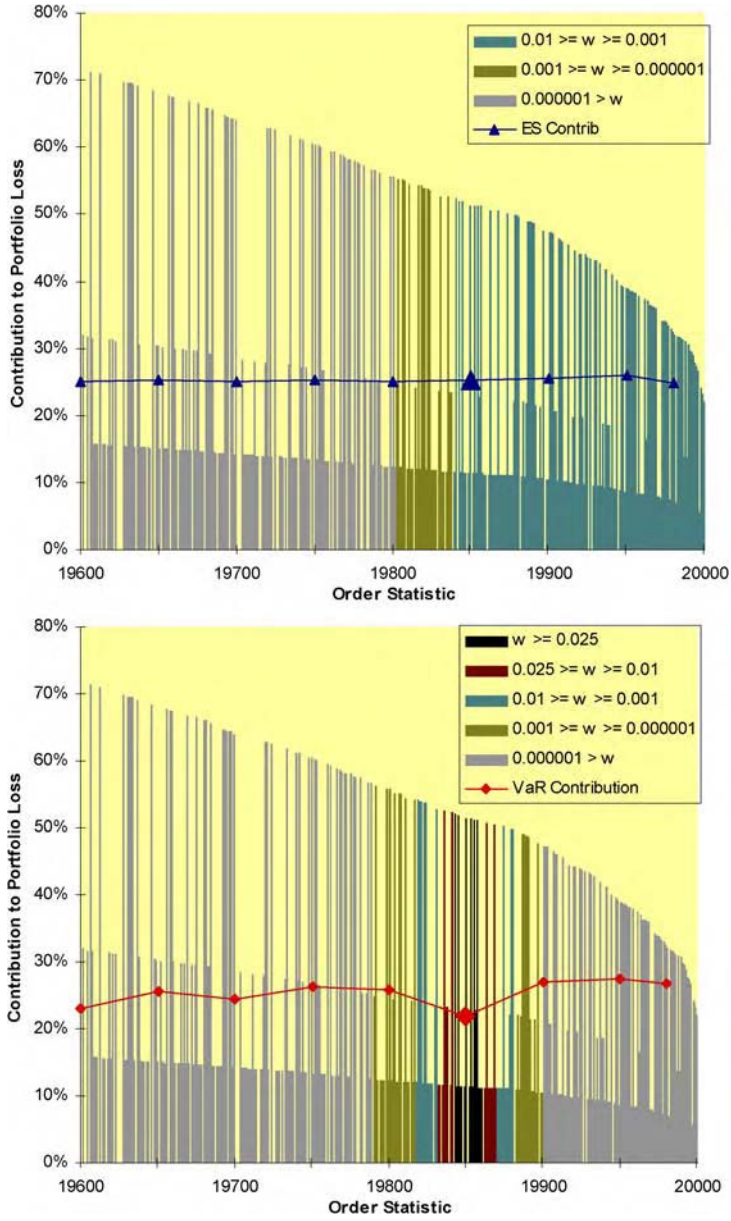


Fig. 9. Estimation of ES and VaR Contributions for Brazil.

advantages and disadvantages of various risk measures and models, the interpretation of various allocation strategies as well as the numerical issues associated with this task.

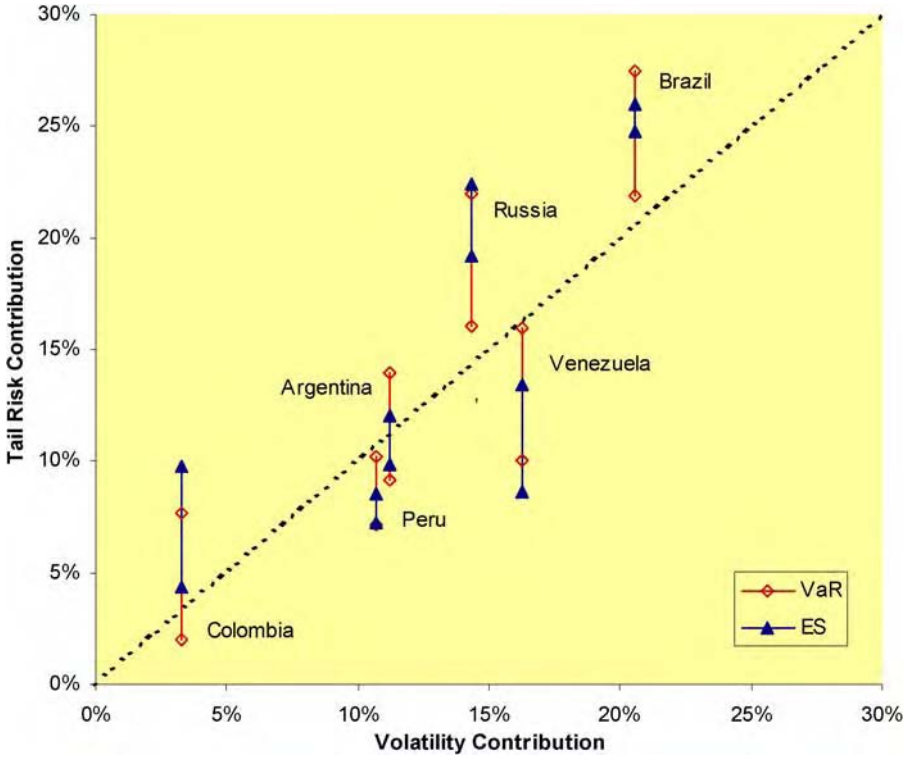


Fig. 10. Most Significant Risk Contributors.

Our key points may be summarized as follows:

- Marginal risk contributions provide a useful basis for allocating EC since they are additive and reflect the benefits of diversification within a portfolio.
- The choice of the risk measure can have a substantial impact on capital allocation. VaR and ES contributions avoid the inconsistencies, and potentially inefficient allocations, associated with the widely-used volatility-based method for EC allocation. The quantile level chosen for measuring risk can also have a significant impact on the relative amount of capital allocated to portfolio components.
- VaR and ES contributions can be calculated analytically under certain simple models (e.g. the Basel II model and several extensions). In addition to providing fast calculations, these models can be used effectively to get a deeper understanding of the behavior of risk contributions. However, these models may present important practical limitations. For example, one-factor, systemic risk models result in linear allocation strategies which are portfolio invariant. Modeling in more detail non-granular portfolios and diversification through multi-factor mod-

els provides a richer picture of diversification and results in more realistic capital allocation strategies.

- Sophisticated credit models that capture the behavior of portfolios in a more realistic manner may entail the use of Monte Carlo simulation for assessing risk. Computing VaR and ES contributions is challenging in this case, specially at the extreme quantiles typically used for credit capital definition. The quality of contribution estimates can be improved in three ways. First, the conditional independence framework can be exploited by using advanced methods to perform the convolution of independent random variables. Second, the use of L-estimators provides more stable contributions, especially for VaR. Finally, the use of variance reduction techniques, such as Importance Sampling can be used effectively to get more accurate contributions at high quantiles in the tails.

Several practical problems relevant to capital allocation are currently the focus of research. These include:

- Statistical estimation of multi-factor models as well as their impact on capital allocation and concentration risk. In addition to requiring accurate PDs, exposures and LGDs, capital allocation methods are very sensitive to the correlations of credit events built into the model. Empirical work is important to understand the relative impact of systematic and idiosyncratic risk, as well as the relationship between economic factors and credit events (e.g. Wendin and McNeil, 2005)
- Consistent capital contributions in large portfolios (and small positions). How can risk contributions be measured accurately when contributions are very small (e.g. for very large portfolios)? In some cases, the size of the risk contribution may be smaller than the error range of the estimator (and the parameters used in model). Examples include retail portfolio with millions of transactions, large corporate portfolios or enterprise portfolios. A practical solution may include the application of a simpler (calibrated) analytical model or the use of a consistent hierarchical methodology to allocate contributions through a large portfolio (e.g. using basic properties of granular, homogeneous portfolios).
- Real-time marginal capital calculations. How can marginal capital be computed consistently for a new loan or transaction (accounting properly for diversification) in "real-time"? In this case, full simulation is typically not an option, although performance might be improved with some semi-analytical models. Some practical solutions may include the application of a simpler analytical model (calibrated to a full economic capital model) as given in Garcia Cespedes et al. (2006). In order to gain acceptance, such a model should be intuitive, based on a small number of parameters and recalibrated frequently over time.

- Contributions of systemic factors. While this chapter has considered the risk contributions of portfolio components, a practitioner may also want to know the contribution to economic capital of the various systemic factors (credit drivers) that are at the heart of a multi-factor economic capital model (such as KMV or CreditMetrics). Such factors explain only the systemic portion of the portfolio's total risk.[24] In addition, the standard theory of marginal capital contributions does not work well since the total capital is not a homogeneous function of these factors. Finally, the most interesting cases, in practice, require simulation of the multi-factor models. For further discussion of systemic risk factor contributions and hedging techniques, see Rosen and Saunders (2006a, 2006b).

Appendix A

A.1 The Harrell–Davis estimator

The Harrell–Davis (HD) estimator derives from the fact that, for $0 < \alpha < 1$, the expected value of order statistic $(M+1)\alpha$ converges to $F^{-1}(\alpha)$ as the sample size increases.[25] Thus, the HD estimator computes VaR_α as $\mathrm{E}[L^{((M+1)\alpha)}]$, regardless of the integrality of $(M+1)\alpha$. The resulting weights are

$$\begin{aligned} w^{VaR}_{\alpha,M,k} &= \frac{1}{\beta[(M+1)\alpha, (M+1)(1-\alpha)]} \\ &\quad \times \int_{(k-1)/M}^{k/M} y^{(M+1)\alpha-1}(1-y)^{(M+1)(1-\alpha)-1}\,\mathrm{d}y \\ &= I_{k/M}\big[(M+1)\alpha, (M+1)(1-\alpha)\big] \\ &\quad - I_{(k-1)/M}\big[(M+1)\alpha, (M+1)(1-\alpha)\big] \end{aligned} \tag{39}$$

where $I_X(a,b)$ is the incomplete beta function. Figure 11 compares the weights of the HD and UECV estimators for computing the 95% VaR from a sample of size 100. A similar comparison for the 95% ES is shown in Fig. 12.

[24] Furthermore, a linear combination of these factors may only explain a portion of the systemic risk (see Rosen and Saunders, 2006a).

[25] In practice, $L^{((M+1)\alpha)}$ is computed as a weighted average of order statistics $L^{(\lfloor (M+1)\alpha \rfloor)}$ and $L^{(\lceil (M+1)\alpha \rceil)}$.

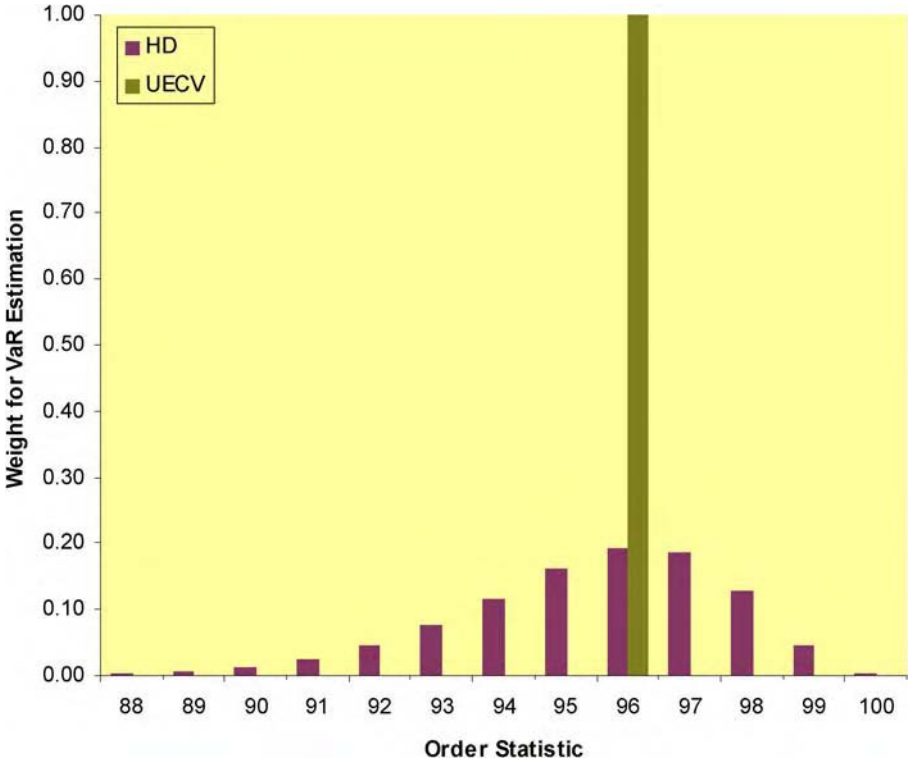


Fig. 11. Weights for Estimating 95% VaR when $S = 100$.

A.2 VaR and ES contributions with CLT in conditional independence models

If conditional losses are Normal, the tail probability of portfolio losses is given by

$$\Pr(L < y) = \frac{1}{M} \sum_{m=1}^{M} \Phi\left(\frac{y - \mu^m}{\sigma^m} \right) \tag{40}$$

where $\Phi(\cdot)$ denotes the cumulative standard Normal probability function,

$$\mu^m = \sum_{j=1}^{N} \mu_j^m \quad \text{and} \quad \left(\sigma^m\right)^2 = \sum_{j=1}^{N} \left(\sigma_j^m\right)^2$$

with the mean and variance of the individual obligor losses L_i^m (see Eq. (3))

$$\mu_j^m = \sum_{r=1}^{R} c_{jr}^m p_{jr}^m \quad \text{and} \quad \left(\sigma_j^m\right)^2 = \sum_{r=1}^{R} p_{jr}^m \left(c_{jr}^m - \mu_j^m\right)^2.$$

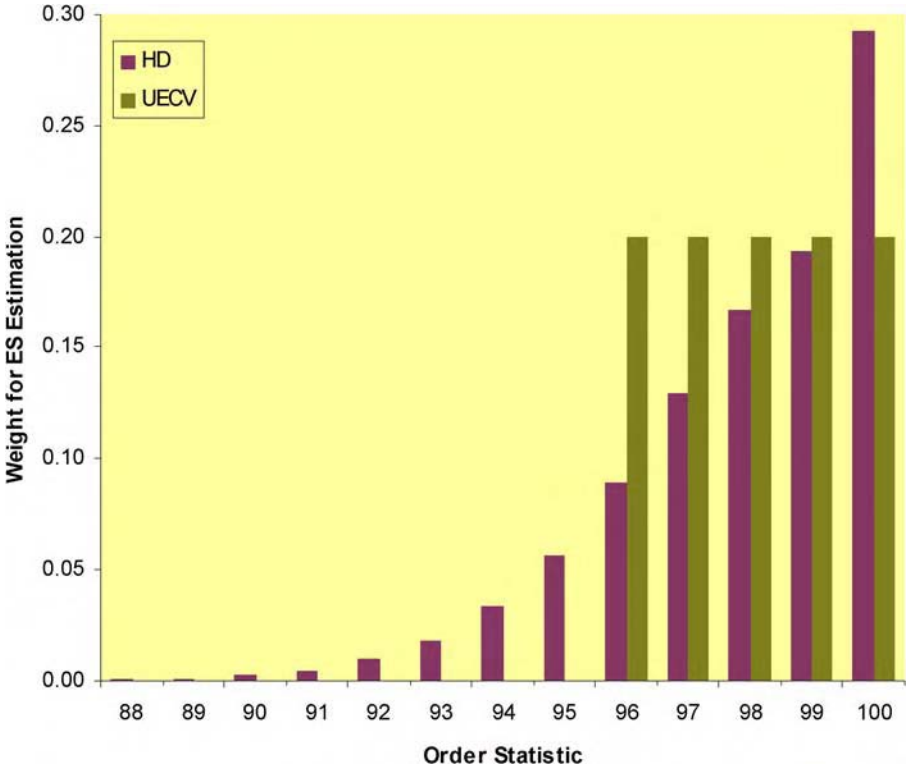


Fig. 12. Weights for Estimating 95% ES when $S = 100$.

Equivalently, if we denote by $\Phi_{\mu,\sigma}$, the $N(\mu, \sigma^2)$ cumulative distribution (with $\varphi_{\mu,\sigma}$, the density functions), we can write Eq. (40) in terms of the estimated $VaR_\alpha(L)$ as

$$\frac{1}{M}\sum_{m=1}^{M} \Phi_{\mu^m,\sigma^m}\big(\overline{VaR}_\alpha(L)\big) = \alpha. \tag{41}$$

Analytical expressions are obtained in this case for the VaR and ES contributions of a given obligor by computing the conditional expectations or taking the derivative of VaR from Eq. (41) (see, for example, Kreinin and Mausser, 2003; Martin, 2004):

$$E\big[L_j \mid L = \overline{VaR}_\alpha\big] = \frac{1}{\varphi_{\mu,\sigma}(\overline{VaR}_\alpha)}\sum_{m=1}^{M}\frac{\varphi_{\mu^m,\sigma^m}(\overline{VaR}_\alpha)}{M}\left(\mu_j^m + \frac{(\sigma_j^m)^2}{\sigma^m}Z_\alpha^m\right) \tag{42}$$

and

$$E\big[L_j \mid L \geqslant \overline{VaR}_\alpha\big] = \frac{1}{1-\alpha}\sum_{m=1}^{M}\frac{1}{M}\left[\mu_j^m\big(1-\Phi_{0,1}\big(Z_\alpha^m\big)\big)+\frac{(\sigma_j^m)^2}{\sigma^m}\varphi_{0,1}\big(Z_\alpha^m\big)\right] \tag{43}$$

where

$$Z_\alpha^m = \frac{\overline{VaR}_\alpha - \mu^m}{\sigma^m}.$$

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance* 26 (7), 1505–1518.

Artzner, P., Delbaen, F., Eber, J.M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9 (3), 203–228.

Aziz, A., Rosen, D. (2004). Capital allocation and RAPM. In: Alexander, C., Sheedy, E. (Eds.), *The Professional Risk Manager's Handbook*. PRMIA Publications, Wilmington, DE, pp. 13–41. www.prmia.org.

Basel Committee on Banking Supervision (BCBS) (1988), *International Convergence of Capital Measurement and Capital Standards*. Available at http://www.bis.org.

Basel Committee on Banking Supervision (BCBS) (2004). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Available at http://www.bis.org.

Bucay, N., Rosen, D. (1999). Credit risk of an international bond portfolio: A case study. *Algo Research Quarterly* 2 (1), 9–29.

Bucay, N., Rosen, D. (2000). Applying portfolio credit risk models to retail portfolios. *Algo Research Quarterly* 3 (1), 45–73.

Credite Suisse Financial Products (1997). *CreditRisk$^+$: A Credit Risk Management Framework*. New York, NY.

Crosbie, P.J. (1999). Modeling default risk. *Manuscript*. KMV Corporation, January 1999.

Denault, M. (2001). Coherent allocation of risk capital. *Journal of Risk* 4 (1), 1–34.

Emmer, S., Tasche, D. (2005). Calculating credit risk capital charges with the one-factor model. *Journal of Risk* 7 (2), 85–101.

Finger, C. (1999). Conditional approaches for CreditMetrics portfolio distributions. *CreditMetrics Monitor* (April), 14–33.

Frey, R., McNeil, A.J. (2003). Dependent defaults in models of portfolio credit risk. *Journal of Risk* 6 (1), 59–92.

Garcia Cespedes, J.C., de Juan Herrero, J.A., Keinin, A., Rosen, D. (2006). A simple multi-factor "factor adjustment" for the treatment of credit capital diversification. *Journal of Credit Risk* 2 (3), 57–85.

Glasserman, P., Li, J. (2005). Importance sampling for portfolio credit risk. *Management Science* 51 (11), 1643–1656.

Glasserman, P. (2005). Measuring marginal risk contributions in credit portfolios. *Journal of Computational Finance* 9, 1–41.

Glasserman, P. (this volume). Calculating portfolio credit risk. In: Linetsky, V., Birge, J. (Eds.), *Handbook of Financial Engineering*.

Gordy, M. (2000). A comparative anatomy of credit risk models. *Journal of Banking and Finance* 24 (1-2), 119–149.

Gordy, M. (2003a). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation* 12 (3), 199–232.
Gordy, M. (2003b). Granularity. In: Szegö, G. (Ed.), *New Risk Measures for Investment and Regulation*. Wiley.
Gouriéroux, C., Laurent, J.P., Scaillet, O. (2000). Sensitivity analysis of values at risk. *Journal of Empirical Finance* 7 (3-4), 225–245.
Gupton, G., Finger, C.A., Bhatia, M. (1997). *CreditMetrics Technical Document*. J.P. Morgan & Co., New York.
Hallerbach, W. (2003). Decomposing portfolio value-at-risk: A general analysis. *Journal of Risk* 5 (2), 1–18.
Harrell, F., Davis, C. (1982). A new distribution-free quantile estimator. *Biometrika* 69 (3), 635–640.
Iscoe, I., Kreinin, A., Rosen, D. (1999). An integrated market and credit risk portfolio model. *Algo Research Quarterly* 1 (2), 21–37.
Kalkbrener, M., Lotter, H., Overbeck, L. (2004). Sensible and efficient capital allocation for credit portfolios. *Risk* (January), S19–S24.
Kupiec, P. (2002). Calibrating your intuition: Capital allocation for market and credit risk. *Working paper WP/02/99*. IMF, available at http://www.imf.org.
Koyluoglu, U., Hickman, A. (1998). Reconcilable differences. *Risk* 11 (10), 56–62 (October).
Koyluoglu, U., Stoker, J. (2002). Honour your contribution. *Risk* (April), 90–94.
Kreinin, A., Mausser, H. (2003). Computation of additive contributions to portfolio risk. *Working paper*. Algorithmics Inc.
Kurth, A., Tasche, D. (2003). Contributions to credit risk. *Risk* (March), 84–88.
Laurent, J.P. (2003). Sensitivity analysis of risk measures for discrete distributions. *Working paper*. http://laurent.jeanpaul.free.fr/var_risk_measure_sensitivity.pdf.
Martin, R. (2004). *Credit Portfolio Modelling Handbook*. Credit Suisse First, Boston.
Martin, R., Thompson, K., Browne, C. (2001). VAR: who contributes and how much? *Risk* (August), 99–102.
Martin, R., Wilde, T. (2002). Unsystematic credit risk. *Risk* 15 (11), 123–128.
Mausser, H. (2003). Calculating quantile-based risk analytics with L-estimators. *Journal of Risk Finance* 4 (3), 61–74.
Mausser, H., Rosen, D. (1998). Beyond VaR: From measuring risk to managing risk. *Algo Research Quarterly* 1 (2), 5–20.
Mausser, H., Rosen, D. (2000). Efficient risk/return frontiers for credit risk. *Journal of Risk Finance* 2 (1), 66–78.
Mausser, H., Rosen, D. (2001). Applying scenario optimization to portfolio credit risk. *Journal of Risk Finance* 2 (2), 36–48.
Mausser, H., Rosen, D. (2005). Scenario-based risk management tools. In: Wallace, S.W., Ziemba, W.T. (Eds.), *Applications of Stochastic Programming*. In: *MPS-SIAM Series in Optimization*, pp. 545–574.
Merino, S., Nyfeler, M.A. (2004). Applying importance sampling for estimating coherent credit risk contributions. *Quantitative Finance* 4, 199–207.
Pykhtin, M. (2004). Multi-factor adjustment. *Risk* (March), 85–90.
Praschnik, J., Hayt, G., Principato, A. (2001). Calculating the contribution. *Risk* (October), S25–S27.
Rosen, D. (2004). Credit risk capital calculation. In: Alexander, C., Sheedy, E. (Eds.), *The Professional Risk Manager's Handbook*. PRMIA Publications, Wilmington, DE, pp. 315–342. http://www.prmia.org.
Rosen, D., Saunders, D. (2006a). Analytical methods for hedging systematic credit risk with linear factor portfolios. *Working paper*. Fields Institute for Mathematical Research and University of Waterloo.
Rosen, D., Saunders, D. (2006b). Measuring capital contributions of systemic factors in credit portfolios. *Working paper*. Fields Institute for Mathematical Research and University of Waterloo.
Scaillet, O. (2004). Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance* 14, 115–129.
Sheather, S.J., Marron, J.S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association* 85, 410–416.

Smithson, C.W. (2003). Economic capital – How much do you really need? *Risk* (November), 60–63.
Tasche, D. (1999). Risk contributions and performance measurement. *Working paper*. Zentrum Mathematik (SCA), TU München.
Tasche, D. (2000). Conditional expectation as quantile derivative. *Working paper*. Technische Universität München.
Tasche, D. (2002). Expected shortfall and beyond. Working paper. Technische Universität München.
Tasche, D. (2006). Measuring sectoral diversification in an asymptotic multi-factor framework. *Journal of Credit Risk* 2 (3), 33–35.
Tchistiakov, V., de Smet, J., Hoogbruin, P.P. (2004). A credit loss control variable. *Risk* (July), 81–85.
Wendin, J., McNeil, A.J. (2005). Dependent credit migrations. *Working paper*. Department of Mathematics, ETH Zurich.
Wilson, T. (1997a). Portfolio credit risk I. *Risk* 10 (9), 111–117.
Wilson, T. (1997b). Portfolio credit risk II. *Risk* 10 (10), 56–61.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15017-9

Chapter 17

Liquidity Risk and Option Pricing Theory

Robert A. Jarrow

Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, USA
E-mail: raj15@cornell.edu

Philip Protter[†]

ORIE-219 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA
E-mail: protter@orie.cornell.edu

Abstract

This paper summarizes the recent advances of Çetin [Çetin, U. (2003). Default and liquidity risk modeling. *Ph.D. thesis*, Cornell University], Çetin et al. [Çetin, U., Jarrow, R., Protter, P. (2004). Liquidity risk and arbitrage pricing theory. *Finance and Stochastics* 8, 311–341], Çetin et al. [Çetin, U., Jarrow, R., Protter, P., Warachka, M. (2006). Pricing options in an extended Black Scholes economy with illiquidity: Theory and empirical evidence. *Review of Financial Studies* 19 (2), 493–529], Blais [Blais, M. (2006). Liquidity and data. *Ph.D. thesis*, Cornell University], and Blais and Protter [Blais, M., Protter, P. (2006). An analysis of the supply curve for liquidity risk through book data, in preparation] on the inclusion of liquidity risk into option pricing theory. This research provides new insights into the relevance of the classical techniques used in continuous time finance for practical risk management.

1 Introduction

Classical asset pricing theory assumes that traders act as price takers, that is, the theory assumes that investors' trades have no impact on the prices paid or received. The relaxation of this price taking assumption and its impact on realized returns in asset pricing models is called *liquidity risk*. Liquidity risk has been extensively studied in the market microstructure literature, but not in the asset pricing literature. In the market microstructure literature, it is well known that a quantity impact on prices can be due to asymmetric information or differential risk tolerances (see Kyle, 1985;

[†] Supported in part by NSF grant DMS-0202958 and NSA grant MDA-904-03-1-0092.

Glosten and Milgrom, 1985; or Grossman and Miller, 1988 in this regard). In an extreme form, liquidity risk has also been studied in the market manipulation literature (see Cvitanic and Ma, 1996; Jarrow, 1992; and Bank and Baum, 2004). And, as argued in Çetin (2003), liquidity risk is related to the transaction costs literature because transaction costs induce a quantity impact on prices paid/received (see also Barles and Soner, 1998; Constantinides and Zariphopoulou, 1999; Cvitanic and Karatzas, 1996; Cvitanic et al., 1999; Jouini, 2000; Jouini and Kallal, 1995; Jouini et al., 2001; Soner et al., 1995 in this regard).

The purpose of this paper is to review the recent research of Çetin (2003), Çetin et al. (2004), and Çetin et al. (2006) on the inclusion of liquidity risk into option pricing theory, and the recent of results of Blais (2006) and Blais and Protter (2006) where these results are interpreted through an analysis of book data. This approach embeds liquidity risk into the classical theory by having investors act as price takers with respect to a C^2 supply curve for the shares. In essence, instead of a single price for all shares traded, investors face a twice continuously differentiable price/quantity schedule. In this framework, it is assumed that the quantity impact on the price transacted is temporary.[1] Given this extension, Çetin et al. (2004) show that appropriate generalizations of the first and second fundamental theorems of asset pricing hold. Briefly stated, in this model, markets are arbitrage free if and only if there exists an equivalent martingale measure. In addition, markets will be approximately complete (in the L^2 sense), if the martingale measure is unique. The converse of this last implication does not hold.

The first and second fundamental theorems extend in this model due to the fact that trading strategies that are both continuous and of finite variation can approximate (in the L^2 sense) arbitrary predictable trading strategies. And, these continuous and finite variation trading strategies can be shown to avoid all liquidity costs. Consequently, the arbitrage-free price of any derivative is shown to be equal to the expected value of its payoff under the risk neutral measure. This is the same price as in the classical economy with no liquidity costs. However, in a world with liquidities, the classical hedge will not be used to replicate the option. Instead, a continuous and finite variation approximation will be used. Both of these observations are consistent with industry usage of the classical arbitrage free pricing methodology. But, they have another set of strong implications for practice.

If one is interested in understanding the quantity impact of trades on prices in options markets as well, then this theory does not readily apply. Indeed, under the C^2 supply curve with continuous trading strategies, all liquidity costs can be avoided when trading in the underlying shares. Although liquidity costs exist, they are nonbinding. Consequently, there can be no quantity impact on

[1] Permanent quantity impacts on prices relates to the previously cited market manipulation literature and it is not studied here.

option prices in such an economy,[2] otherwise arbitrage opportunities exist. To accommodate upward sloping supply curves for options, either the supply curve for the stock must have a discontinuity at 0 (it must violate the C^2 hypothesis) or continuous trading strategies must be excluded. Both extensions are possible. The first extension relates to the transaction cost literature (see Çetin, 2003). The second extension is investigated in Çetin et al. (2006). This second extension is important because continuous trading strategies are not feasible in practice, and only approximating simple trading strategies can be applied. Yet, for simple trading strategies, liquidity costs are binding. This liquidity cost impact implies that the markets are no longer complete, and exact replication is not possible, implying an upward sloping supply curve for options can exist. Çetin et al. (2006) show, in this context, how to super-replication options with minimum liquidity costs. The cost of the super-replication strategy provides an upper bound on the supply curve for the option market.

An outline for this paper is as follows. Section 2 describes the basic economy. Sections 3 and 4 study the first and second fundamental theorems of asset pricing, respectively. Section 5 provides an example – the extended Black–Scholes economy. Section 6 investigates a model with supply curves for options, Section 7 relates the supply curve formulation to transaction costs, Section 8 discusses examples inspired by an analysis of data, and Section 9 concludes the paper.

2 The model

This section presents the model. We are given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\leqslant t\leqslant T}, \mathbb{P})$ satisfying the usual conditions where T is a fixed time. $\mathbb{P}$ represents the statistical or empirical probability measure. We also assume that $\mathcal{F}_0$ is trivial, i.e. $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

We consider a market for a security that we will call a stock with no dividends. Also traded is a money market account that accumulates value at the spot rate of interest. Without loss of generality, we assume that the spot rate of interest is zero, so that the money market account has unit value for all times.[3]

2.1 Supply curve

We consider an arbitrary trader who acts as a price taker with respect to an exogenously given supply curve for shares bought or sold of this stock within the trading interval. More formally, let $S(t, x, \omega)$ represent the stock price, *per share*, at time $t \in [0, T]$ that the trader pays/receives for an order of size $x \in R$

[2] Recall that the previous theory implies that there is a *unique* price for an option (long or short), independent of the quantity of shares in the underlying traded.

[3] A numéraire invariance theorem is proved in Çetin et al. (2004).

given the state $\omega \in \Omega$. A positive order ($x > 0$) represents a buy, a negative order ($x < 0$) represents a sale, and the order zero ($x = 0$) corresponds to the marginal trade. By construction, rather than the trader facing a horizontal supply curve as in the classical theory (the same price for any order size), the trader now faces a supply curve that depends on his order size.[4] Note that the supply curve is otherwise independent of the trader's past actions, endowments, risk aversion, or beliefs. This implies that an investor's trading strategy has no lasting impact on the price process.

We now impose some structure on the supply curve.

Assumption 1 (Supply curve).

1. $S(t, x, \cdot)$ is $\mathcal{F}_t$-measurable and nonnegative.
2. $x \mapsto S(t, x, \omega)$ is a.e. t nondecreasing in x, a.s. [i.e. $x \leqslant y$ implies $S(t, x, \omega) \leqslant S(t, y, \omega)$ a.s. $\mathbb{P}$, a.e. t].
3. S is C^2 in its second argument, $\partial S(t, x)/\partial x$ is continuous in t, and $\partial^2 S(t, x)/\partial x^2$ is continuous in t.
4. $S(\cdot, 0)$ is a semimartingale.
5. $S(\cdot, x)$ has continuous sample paths (including time 0) for all x.

Except for the second condition, these restrictions are self-explanatory. Condition 2 is the situation where the larger the purchase (or sale), the larger the price impact that occurs on the share price. This is the usual situation faced in asset pricing markets, where the quantity impact on the price is due to either information effects or supply/demand imbalances (see Kyle, 1985; Glosten and Milgrom, 1985; Grossman and Miller, 1988). It includes, as a special case, horizontal supply curves.[5]

Example 1 (Supply curve). To present a concrete example of a supply curve, let $S(t, x) \equiv f(t, D_t, x)$ where D_t is an n-dimensional, $\mathcal{F}_t$-measurable semimartingale, and $f : R^{n+2} \to R^+$ is Borel measurable, C^1 in t, and C^2 in all its other arguments. This nonnegative function f can be viewed as a reduced form supply curve generated by a market equilibrium process in a complex and dynamic economy. Under this interpretation, the vector stochastic process D_t represents the state variables generating the uncertainty in the economy, often assumed to be diffusion processes or at least Markov processes (e.g. a solution to a stochastic differential equation driven by a Levy process).

[4] In contrast, the trader is assumed to have no quantity impact due to his trades in the money market account.

[5] This structure can also be viewed as a generalization of the model in Jouini (2000) where the traded securities have distinct selling and buying prices following separate stochastic processes.

2.2 Trading strategies

We start by defining the investor's trading strategy.

Definition 1. A trading strategy is a triplet $((X_t, Y_t\colon\ t \in [0, T]), \tau)$ where X_t represents the trader's aggregate stock holding at time t (units of the stock), Y_t represents the trader's aggregate money market account position at time t (units of the money market account), and τ represents the liquidation time of the stock position, subject to the following restrictions: (a) X_t and Y_t are predictable and optional processes, respectively, with $X_{0-} \equiv Y_{0-} \equiv 0$, and (b) $X_T = 0$ and τ is a predictable $(\mathcal{F}_t\colon\ 0 \leqslant t \leqslant T)$ stopping time with $\tau \leqslant T$ and $X = H1_{[0,\tau)}$ for some predictable process $H(t, \omega)$.

We are interested in a particular type of trading strategy – those that are self-financing. By construction, a self-financing trading strategy generates no cash flows for all times $t \in [0, T)$. That is, purchase/sales of the stock must be obtained via borrowing/investing in the money market account. This implies that Y_t is uniquely determined by (X_t, τ). The goal is to define this self-financing condition for Y_t given an arbitrary stock holding (X_t, τ).

Definition 2. A self-financing trading strategy (s.f.t.s.) is a trading strategy $((X_t, Y_t\colon\ t \in [0, T]), \tau)$ where (a) X_t is càdlàg if $\partial S(t, 0)/\partial x \equiv 0$ for all t, and X_t is càdlàg with finite quadratic variation $([X, X]_T < \infty)$ otherwise, (b) $Y_0 = -X_0 S(0, X_0)$, and (c) for $0 < t \leqslant T$,

$$Y_t = Y_0 + X_0 S(0, X_0) + \int_0^t X_{u-}\,\mathrm{d}S(u, 0) - X_t S(t, 0) - \sum_{0 \leqslant u \leqslant t} \Delta X_u \big[S(u, \Delta X_u) - S(u, 0)\big] - \int_0^t \frac{\partial S}{\partial x}(u, 0)\,\mathrm{d}[X, X]_u^c. \tag{1}$$

Condition (a) imposes restrictions on the class of acceptable trading strategies. Under the hypotheses that X_t is càdlàg and of finite quadratic variation, the right side of expression (1) is always well defined although the last two terms (always being nonpositive) may be negative infinity. The classical theory, under frictionless and competitive markets, does not need these restrictions. An example of a trading strategy that is allowed in the classical theory, but disallowed here, is $X_t = 1_{\{S(t,0) > K\}}$ for some constant $K > 0$ where $S(t, 0)$ follows a Brownian motion. Under the Brownian motion hypothesis this is a discontinuous trading strategy that jumps infinitely often immediately after $S(t, 0) = K$ (the jumps are not square summable), and hence Y_t is undefined.

Condition (b) implies the strategy requires zero initial investment at time 0. When studying complete markets in a subsequent section, condition (b) of the s.f.t.s. is removed so that $Y_0 + X_0 S(0, X_0) \neq 0$.

Condition (c) is the self-financing condition at time t. The money market account equals its value at time 0, plus the accumulated trading gains (evaluated at the marginal trade), less the cost of attaining this position, less the price impact costs of discrete changes in share holdings, and less the price impact costs of continuous changes in the share holdings. This expression is an extension of the classical self-financing condition when the supply curve is horizontal. To see this note that using condition (b) with expression (1) yields the following simplified form of the self-financing condition:

$$\begin{aligned} Y_t + X_t S(t,0) = &\int_0^t X_{u-}\, \mathrm{d}S(u,0) \\ &- \sum_{0 \leqslant u \leqslant t} \Delta X_u \big[S(u, \Delta X_u) - S(u,0) \big] \\ &- \int_0^t \frac{\partial S}{\partial x}(u,0)\, \mathrm{d}[X,X]_u^c \quad \text{for } 0 \leqslant t \leqslant T. \end{aligned} \tag{2}$$

The left side of expression (2) represents the classical "value" of the portfolio at time 0. The right side gives its decomposition into various components. The first term on the right side is the classical "accumulated gains/losses" to the portfolio's value. The last two terms on the right side capture the impact of illiquidity, both entering with a negative sign.

2.3 The marked-to-market value of a s.f.t.s. and its liquidity cost

This section defines the marked-to-market value of a trading strategy and its liquidity cost. At any time prior to liquidation, there is no unique value of a trading strategy or portfolio. Indeed, any price on the supply curve is a plausible price to be used in valuing the portfolio. At least three economically meaningful possibilities can be identified: (i) the immediate liquidation value (assuming that $X_t > 0$ gives $Y_t + X_t S(t, -X_t)$), (ii) the accumulated cost of forming the portfolio (Y_t), and (iii) the portfolio evaluated at the marginal trade ($Y_t + X_t S(t,0)$).[6] This last possibility is defined to be the *marked-to-market value* of the self-financing trading strategy (X, Y, τ). It represents the value of the portfolio under the classical price taking condition.

Motivated by expression (2), we define the liquidity cost to be the difference between the accumulated gains/losses to the portfolio, computed as if all trades

[6] These three valuations are (in general) distinct except at one date, the liquidation date. At the liquidation time τ, the value of the portfolio under each of these three cases are equal because $X_\tau = 0$.

are executed at the marginal trade price $S(t, 0)$, and the marked-to-market value of the portfolio.

Definition 3. The liquidity cost of a s.f.t.s. (X, Y, τ) is

$$L_t \equiv \int_0^t X_{u-}\, \mathrm{d}S(u, 0) - \big[Y_t + X_t S(t, 0)\big].$$

The following lemma follows from the preceding definition.

Lemma 1 *(Equivalent characterization of the liquidity costs).*

$$L_t = \sum_{0 \leqslant u \leqslant t} \Delta X_u \big[S(u, \Delta X_u) - S(u, 0)\big] + \int_0^t \frac{\partial S}{\partial x}(u, 0)\, \mathrm{d}[X, X]_u^c \geqslant 0,$$

where $L_{0-} = 0$, $L_0 = X_0[S(0, X_0) - S(0, 0)]$ *and* L_t *is nondecreasing in* t.

Proof. The first equality follows directly from the definitions. The second inequality and the subsequent observation follow from the fact that $S(u, x)$ is increasing in x.

We see here that the liquidity cost is nonnegative and nondecreasing in t. It consists of two components. The first is due to discontinuous changes in the share holdings. The second is due to the continuous component. This expression is quite intuitive. Note that because $X_{0-} = Y_{0-} = 0$, $\Delta L_0 = L_0 - L_{0-} = L_0 > 0$ is possible. □

3 The extended first fundamental theorem

This section studies the characterization of an arbitrage free market and generalizes the first fundamental theorem of asset pricing to an economy with liquidity risk.

To evaluate a self-financing trading strategy, it is essential to consider its value after liquidation. This is equivalent to studying the portfolio's real wealth, as contrasted with its marked-to-market value or paper wealth, see Jarrow (1992). Using this insight, an arbitrage opportunity can now be defined.

Definition 4. An arbitrage opportunity is a s.f.t.s. (X, Y, τ) such that $\mathbb{P}\{Y_T \geqslant 0\} = 1$ and $\mathbb{P}\{Y_T > 0\} > 0$.

We first need to define some mathematical objects. Let $s_t \equiv S(t, 0)$, $(X_- \cdot s)_t \equiv \int_0^t X_{u-}\, \mathrm{d}S(u, 0)$, and for $\alpha \geqslant 0$, let $\Theta_\alpha \equiv \{\text{s.f.t.s } (X, Y, \tau) \mid (X_- \cdot s)_t \geqslant -\alpha \text{ for all } t \text{ almost surely}\}$.

Definition 5. Given an $\alpha \geqslant 0$, a s.f.t.s. (X, Y, τ) is said to be α-admissible if $(X, Y, \tau) \in \Theta_\alpha$. A s.f.t.s. is admissible if it is α-admissible for some α.

Lemma 2 *($Y_t + X_t S(t, 0)$ is a supermartingale). If there exists a probability measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot, 0)$ is a $\mathbb{Q}$-local martingale, and if $(X, Y, \tau) \in \Theta_\alpha$ for some α, then $Y_t + X_t S(t, 0)$ is a $\mathbb{Q}$-supermartingale.*

Proof. From Definition 3 we have that $Y_t + X_t S(t, 0) = (X_- \cdot s)_t - L_t$. Under the $\mathbb{Q}$ measure, $(X_- \cdot s)_t$ is a local $\mathbb{Q}$-martingale. Since $(X, Y, \tau) \in \Theta_\alpha$ for some α, it is a supermartingale (Duffie, 1996). But, by Lemma 1, L_t is nonnegative and nondecreasing. Therefore, $Y_t + X_t S(t, 0)$ is a supermartingale too. □

Theorem 1 *(A sufficient condition for no arbitrage). If there exists a probability measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot, 0)$ is a $\mathbb{Q}$-local martingale, then there is no arbitrage for $(X, Y, \tau) \in \Theta_\alpha$ for any α.*

Proof. Under this hypothesis, by Lemma 2, $Y_t + X_t S(t, 0)$ is a supermartingale. Note that $Y_\tau + X_\tau S(\tau, 0) = Y_\tau$ by the definition of the liquidation time. Thus, for this s.f.t.s., $\mathbb{E}^{\mathbb{Q}}[Y_\tau] = \mathbb{E}^{\mathbb{Q}}[Y_\tau + X_\tau S(\tau, 0)] \leqslant 0$. But, by the definition of an arbitrage opportunity, $\mathbb{E}^{\mathbb{Q}}[Y_\tau] > 0$. Hence, there exist no arbitrage opportunities in this economy. □

The intuition behind this theorem is straightforward. The marked-to-market portfolio is a hypothetical portfolio that contains zero liquidity costs (see Definition 3). If $S(\cdot, 0)$ has an equivalent martingale measure, then these hypothetical portfolios admit no arbitrage. But, since the actual portfolios differ from these hypothetical portfolios only by the subtraction of nonnegative liquidity costs (Lemma 1), the actual portfolios cannot admit arbitrage either.

In order to get a sufficient condition for the existence of an equivalent local martingale measure, we need to define the notion of a free lunch with vanishing risk as in Delbaen and Schachermayer (1994). This will require a preliminary definition.

Definition 6. A free lunch with vanishing risk (FLVR) is either: (i) an admissible s.f.t.s. that is an arbitrage opportunity or (ii) a sequence of ϵ_n-admissible s.f.t.s. $(X^n, Y^n, \tau^n)_{n \geqslant 1}$ and a nonnegative F_T-measurable random variable, f_0, not identically 0 such that $\epsilon_n \to 0$ and $Y_T^n \to f_0$ in probability.[7]

To state the theorem, we need to introduce a related, but fictitious economy. Consider the economy introduced previously, but suppose instead that

[7] Delbaen and Schachermayer (1994, Proposition 3.6, p. 477) shows that this definition is equivalent to FLVR in the classical economy.

$S(t, x) \equiv S(t, 0)$. When there is no confusion, we denote $S(t, 0)$ by the simpler notation s_t. In this fictitious economy, a s.f.t.s. (X, Y^0, τ) satisfies the classical condition with $X_0 = 0$, the value of the portfolio is given by $Z_t^0 \equiv Y_t^0 + X_t s_t$ with $Y_t^0 = (X \cdot s)_t - X_t s_t$ for all $0 \leqslant t \leqslant T$, and X is allowed to be a general $S(\cdot, 0) \equiv s$ integrable predictable process [see the remark following expression (1)]. So, in this fictitious economy, our definitions of an arbitrage opportunity, an admissible trading strategy, and a NFLVR collapse to those in (Delbaen and Schachermayer, 1994).

Theorem 2 *(First fundamental theorem). Suppose there are no arbitrage opportunities in the fictitious economy. Then, there is no free lunch with vanishing risk (NFLVR) if and only if there exists a probability measure* $\mathbb{Q} \sim \mathbb{P}$ *such that* $S(\cdot, 0) \equiv s$ *is a* $\mathbb{Q}$*-local martingale.*

The proof is in Appendix A.

4 The extended second fundamental theorem

This section studies the meaning and characterization of a complete market and generalizes the second fundamental theorem of asset pricing to an economy with liquidity risk. For this section we assume that there exists an equivalent local martingale measure $\mathbb{Q}$ so that the economy is arbitrage free and there is no free lunch with vanishing risk (NFLVR).

Also for this section, we generalize the definition of a s.f.t.s (X, Y, τ) slightly to allow for nonzero investments at time 0. In particular, a s.f.t.s. (X, Y, τ) in this section will satisfy Definition 2 with the exception that condition (b) is removed. That is, a s.f.t.s. need not have zero initial value $(Y_0 + X_0 S(0, X_0) \neq 0)$.[8]

To proceed, we need to define the space $\mathcal{H}^2_Q$ of semimartingales with respect to the equivalent local martingale measure $\mathbb{Q}$. Let Z be a special semimartingale with canonical decomposition $Z = N + A$, where N is a local martingale under $\mathbb{Q}$ and A is a predictable finite variation process. The $\mathcal{H}^2$ norm of Z is defined to be

$$\|Z\|_{\mathcal{H}^2} = \left\| [N, N]_\infty^{1/2} \right\|_{L^2} + \left\| \int_0^\infty |\mathrm{d}A_s| \right\|_{L^2},$$

[8] In this section we could also relax condition (b) of a trading strategy, Definition 1, to remove the requirement that $X_T = 0$. However, as seen below, it is always possible to approximate any random variable with such a trading strategy. Consequently, this restriction is without loss of generality in the context of our model. This condition was imposed in the previous section to make the definition of an arbitrage opportunity meaningful in a world with liquidity costs.

where the L^2-norms are with respect to the equivalent local martingale measure $\mathbb{Q}$.

Throughout this section we make the assumption that $s(\cdot) = S(\cdot, 0) \in \mathcal{H}^2_Q$. Since we are assuming $s \in \mathcal{H}^2_Q$, it is no longer necessary to require that $X \cdot s$ is uniformly bounded from below.

Definition 7. A contingent claim is any $\mathcal{F}_T$-measurable random variable C with $\mathbb{E}^{\mathbb{Q}}(C^2) < \infty$.

Note that the contingent claim is considered at a time T, prior to which the trader's stock position is liquidated. If the contingent claim's payoff depends on the stock price at time T, then the dependence of the contingent claim's payoff on the shares purchased/sold at time T must be made explicit. Otherwise, the contingent claim's payoff is not well defined. An example helps to clarify this necessity.

Consider a European call option on the stock with a strike price[9] of K and maturity $T_0 \leqslant T$.[10] To write the modified boundary condition for this option incorporating the supply curve for the stock, we must consider two cases: cash delivery and physical delivery.

1. If the option has cash delivery, the long position in the option receives cash at maturity if the option ends up in-the-money. To match the cash settlement, the synthetic option position must be liquidated prior to time T_0. When the synthetic option position is liquidated, the underlying stock position is also liquidated. The position in the stock at time T_0 is, thus, zero.

   If we sell the stock at time T_0 to achieve this position, then the boundary condition is $C \equiv \max[S(T_0, -1) - K, 0]$ where $\Delta X_{T_0} = -1$ since the option is for one share of the stock. However, as we show below, one could also liquidate this stock position just prior to time T_0 using a continuous and finite variation process, so that $\Delta X_{T_0} = 0$. This alternative liquidation strategy might be chosen in an attempt to avoid liquidity costs at time T_0. In this case, the boundary condition is $C \equiv \max[S(T_0, 0) - K, 0]$. Note that using this latter liquidation strategy, the option's payoff is only approximately obtained (to a given level of accuracy) because liquidation occurs just before T_0.
2. If the option has physical delivery, then the synthetic option position should match the physical delivery of the stock in the option contract. With physical delivery, the option contract obligates the short position

[9] To be consistent with the previous construct, one should interpret K as the strike price normalized by the value of the money market account at time T_0.

[10] Recall that interest rates are zero, so that the value of the liquidated position at time T_0 is the same as the position's value at time T.

to deliver the stock shares. To match the physical delivery, the stock position in the synthetic option is not sold. Unfortunately, our model requires the stock position to be liquidated at time T_0. Formally, physical delivery is not possible in our construct. However, to approximate physical delivery in our setting, we can impose the boundary condition $C \equiv \max[S(T_0, 0) - K, 0]$ where $\Delta X_{T_0} = 0$. This boundary condition is consistent with no liquidity costs being incurred at time T_0, which would be the case with physical delivery of the stock.[11]

Definition 8. The market is complete if given any contingent claim C, there exists a s.f.t.s. (X, Y, τ) with $\mathbb{E}^{\mathbb{Q}}(\int_0^T X_u^2 \,\mathrm{d}[s, s]_u) < \infty$ such that $Y_T = C$.

To understand the issues involved in replicating contingent claims, let us momentarily consider a contingent claim C in $L^2(d\mathbb{Q})$ where there exists a s.f.t.s. (X, Y, τ) such that $C = c + \int_0^T X_u \,\mathrm{d}s_u$ where $c \in \mathbb{R}$ and $\mathbb{E}^{\mathbb{Q}}\{\int_0^T X_u^2 \,\mathrm{d}[s, s]_u\} < \infty$. Note that $\mathbb{E}^{\mathbb{Q}}(C) = c$ since $\int_0^0 X_u \,\mathrm{d}s_u = X_0 \Delta s_0 = 0$ by the continuity of s at time 0. This is the situation where, in the absence of liquidity costs, a long position in the contingent claim C is redundant. In this case, Y_0 is chosen so that $Y_0 + X_0 s_0 = c$. But, the liquidity costs in trading this stock position are (by Lemma 1):

$$L_t = \sum_{0 \leqslant u \leqslant t} \Delta X_u \big[S(u, \Delta X_u) - S(u, 0)\big] + \int_0^t \frac{\partial S}{\partial x}(u, 0) \,\mathrm{d}[X, X]_u^c \geqslant 0.$$

We have from Definition 2 that

$$Y_T = Y_0 + X_0 s_0 + \int_0^T X_{u-} \,\mathrm{d}s_u - X_T s_T - L_T + L_0$$

and[12] $\int_0^T X_{u-} \,\mathrm{d}s_u = \int_0^T X_u \,\mathrm{d}s_u$ so that

$$Y_T = C - X_T s_T - L_T + L_0.$$

By assumption, we have liquidated by time T, giving $X_T = 0$. Thus, we have

$$Y_T = C - (L_T - L_0) \leqslant C.$$

[11] We are studying an economy with trading only in the stock and money market account. Expanding this economy to include trading in an option expands the liquidation possibilities prior to time T. Included in this set of expanded possibilities is the delivery of the physical stock to offset the position in an option, thereby avoiding any liquidity costs at time T. Case 2 is the method for capturing no liquidity costs in our restricted setting.

[12] $\int_0^T X_u \,\mathrm{d}s_u = \int_0^T X_{u-} \,\mathrm{d}s_u + \sum_{0 \leqslant u \leqslant T} \Delta X_u \Delta s_u$ and $\Delta X_u \Delta s_u = 0$ for all u since $\Delta s_u = 0$ for all u by the continuity of s.

That is, considering liquidity costs, this trading strategy sub-replicates a long position in this contingent claim's payoffs. Analogously, if we use $-X$ to hedge a short position in this contingent claim, the payoff is generated by

$$\overline{Y}_T = -C - (\overline{L}_T - \overline{L}_0) \leqslant -C,$$

where $\overline{Y}$ is the value in the money market account and $\overline{L}$ is the liquidity cost associated with $-X$. The liquidation value of the trading strategies (long and short) provide a lower and upper bound on attaining the contingent claim's payoffs.

Remark 1.

1. If $\frac{\partial S}{\partial x}(\cdot, 0) \equiv 0$, then $L_\cdot = L_0$ if X is a continuous trading strategy. So, under this hypothesis, all claims C where there exists a s.f.t.s. (X, Y, τ) such that $C = c + \int_0^T X_u \, \mathrm{d}s_u$ with X continuous can be replicated. For example, if $S(\cdot, 0)$ is a geometric Brownian motion (an extended Black–Scholes economy), a call option can be replicated since the Black–Scholes hedge is a continuous s.f.t.s.
2. If $\frac{\partial S}{\partial x}(\cdot, 0) \geqslant 0$ (the general case), then $L = L_0$ if X is a finite variation and continuous trading strategy. So, under this hypothesis, all claims C where there exists a s.f.t.s. (X, Y, τ) such that $C = c + \int_0^T X_u \, \mathrm{d}s_u$ with X of finite variation and continuous can be replicated.

The remark above shows that if we can approximate X using a finite variation and continuous trading strategy, in a limiting sense, we may be able to avoid all the liquidity costs in the replication strategy. In this regard, the following lemma is relevant.

Lemma 3 *(Approximating continuous and finite variation s.f.t.s.). Let* $C \in L^2(d\mathbb{Q})$. *Suppose there exists a predictable* X *with* $\mathbb{E}^{\mathbb{Q}}(\int_0^T X_u^2 \, \mathrm{d}[s, s]_u) < \infty$ *so that* $C = c + \int_0^T X_u \, \mathrm{d}s_u$ *for some* $c \in \mathbb{R}$. *Then, there exists a sequence of s.f.t.s.* $(X^n, Y^n, \tau^n)_{n \geqslant 1}$ *with* X^n *bounded, continuous and of finite variation such that* $\mathbb{E}^{\mathbb{Q}}(\int_0^T (X_u^n)^2 \, \mathrm{d}[s, s]_u) < \infty$, $X_0^n = 0$, $X_T^n = 0$, $Y_0^n = \mathbb{E}^{\mathbb{Q}}(C)$ *for all n and*

$$\begin{aligned} Y_T^n &= Y_0^n + X_0^n S\big(0, X_0^n\big) + \int_0^T X_{u-}^n \, \mathrm{d}s_u - X_T^n S(T, 0) - L_T^n \\ &\to c + \int_0^T X_u \, \mathrm{d}s_u = C \end{aligned} \tag{3}$$

in $L^2(d\mathbb{Q})$.

Proof. Note that for any predictable X that is integrable with respect to s, $\int_0^T X_u \,\mathrm{d}s_u = \int_0^T X_u 1_{(0,T]}(u)\,\mathrm{d}s_u$ since $\int_0^T 1_{(0,T]} X_u \,\mathrm{d}s_u = \int_0^T X_u \,\mathrm{d}s_u - X_0 \Delta s_0$ and $\Delta s_0 = 0$. Therefore, we can without loss of generality assume that $X_0 = 0$.

Given any $H \in \mathbb{L}$ (the set of adapted processes that have left continuous paths with right limits a.s.) with $H_0 = 0$, we define, H^n, by the following:

$$H_t^n(\omega) = n \int_{t-\frac{1}{n}}^{t} H_u(\omega)\,\mathrm{d}u,$$

for all $t \geqslant 0$, letting H_u equal 0 for $u < 0$. Then H is the a.s. pointwise limit of the sequence of adapted processes H^n that are continuous and of finite variation. Note that $H_0^n = 0$ for all n. Theorem 2 in Chapter IV of Protter (2005) remains valid if $\mathbf{b}\mathbb{L}$ is replaced by the set of bounded, continuous processes with paths of finite variation on compact time sets. Let X with $X_0 = 0$ be predictable and $\mathbb{E}^{\mathbb{Q}}(\int_0^T X_u^2 \,\mathrm{d}[s,s]_u) < \infty$. Since $X \cdot s$ is defined to be the $\lim_{k\to\infty} \overline{X}^k \cdot s$, where the convergence is in $L^2(d\mathbb{Q})$ and $\overline{X}^k = X 1_{\{|X| \leqslant k\}}$, and using the above observation, there exists a sequence of continuous and bounded processes of finite variation, $(X^n)_{n\geqslant 1}$, such that $\mathbb{E}^{\mathbb{Q}}(\int_0^T (X_u^n)^2 \,\mathrm{d}[s,s]_u) < \infty$, $X_0^n = 0$ for all n and

$$\int_0^T X_u^n \,\mathrm{d}s_u \to \int_0^T X_u \,\mathrm{d}s_u,$$

in $L^2(d\mathbb{Q})$ (see Protter, 2005, Theorems 2, 4, 5 and 14 in Chapter IV in this respect).

Furthermore, Theorem 12 and Corollary 3 in Appendix A allow us to choose $X_T^n = 0$ for all n. Now, choose $Y^n = \mathbb{E}^{\mathbb{Q}}(C)$ for all n and define Y_t^n for $t > 0$ by (1). Let $\tau^n = T$ for all n. Then, the sequence $(X^n, Y^n, \tau^n)_{n\geqslant 1}$ will satisfy (3). Note that $L^n \equiv 0$ for all n and $\int_0^T X_{u-}^n \,\mathrm{d}s_u = \int_0^T X_u^n \,\mathrm{d}s_u$. □

This lemma motivates the following definition and extension of the second fundamental theorem of asset pricing.

Definition 9. The market is approximately complete if given any contingent claim C, there exists a sequence of s.f.t.s. (X^n, Y^n, τ^n) with $\mathbb{E}^{\mathbb{Q}}(\int_0^T (X_u^n)^2 \,\mathrm{d}[s,s]_u) < \infty$ for all n such that $Y_T^n \to C$ as $n \to \infty$ in $L^2(d\mathbb{Q})$.

Theorem 3 *(Second fundamental theorem). Suppose there exists a* unique *probability measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot, 0) = s$ is a $\mathbb{Q}$-local martingale. Then, the market is approximately complete.*

Proof. The proof proceeds in two steps. Step 1 shows that the hypothesis guarantees that a fictitious economy with no liquidity costs is complete. Step 2

shows that this result implies approximate completeness for an economy with liquidity costs.

Step 1. Consider the economy introduced in this paper, but suppose that $S(\cdot, x) \equiv S(\cdot, 0)$. In this fictitious economy, a s.f.t.s. (X, Y^0, τ) satisfies the classical condition with $Y_t^0 \equiv Y_0 + X_0 S(0,0) + \int_0^t X_{u-}\,\mathrm{d}s_u - X_t s_t$. The classical second fundamental theorem (see Harrison and Pliska, 1981) applies: the fictitious market is complete if and only if $\mathbb{Q}$ is unique.

Step 2. By Step 1, given $\mathbb{Q}$ is unique, the fictitious economy is complete and, moreover, s has the martingale representation property. Hence, there exists a predictable X such that $C = c + \int_0^T X_u\,\mathrm{d}s_u$ with $\mathbb{E}^{\mathbb{Q}}(\int_0^T X_u^2\,\mathrm{d}[s,s]_u) < \infty$ [see Protter (2005, Section 3 of Chapter IV) in this respect]. Then, by applying the lemma above, the market is approximately complete. □

Suppose the martingale measure is unique. Then, by the theorem we know that given any contingent claim C, there exists a sequence of s.f.t.s. $(X^n, Y^n, \tau^n)_{n\geqslant 1}$ with $\mathbb{E}^{\mathbb{Q}}(\int_0^T (X_u^n)^2\,\mathrm{d}[s,s]_u) < \infty$ for all n so that $Y_T^n = Y_0^n + X_0^n S(0, X_0^n) - L_T^n + \int_0^T X_{u-}^n\,\mathrm{d}S(u,0) \to C$ in $L^2(d\mathbb{Q})$. We call any such sequence of s.f.t.s., $(X^n, Y^n, \tau^n)_{n\geqslant 1}$ an *approximating sequence* for C.

Definition 10. Let C be a contingent claim and Ψ^C be the set of approximating sequences for C. The time 0 value of the contingent claim C is given by

$$\inf\Big\{\liminf_{n\to\infty} Y_0^n + X_0^n S\big(0, X_0^n\big)\colon \big(X^n, Y^n, \tau^n\big)_{n\geqslant 1} \in \Psi^C\Big\}.$$

Corollary 1 *(Contingent claim valuation). Suppose there exists a unique probability measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot,0) = s$ is a $\mathbb{Q}$-local martingale. Then, the time* 0 *value of any contingent claim C is equal to $\mathbb{E}^{\mathbb{Q}}(C)$.*

Proof. Let $(X^n, Y^n, \tau^n)_{n\geqslant 1}$ be an approximating sequence for C. Then, $\mathbb{E}^{\mathbb{Q}}(Y_T^n - C)^2 \to 0$, and thus, $\mathbb{E}^{\mathbb{Q}}(Y_T^n - C) \to 0$. However, since $\mathbb{E}^{\mathbb{Q}}(\int_0^T (X_u^n)^2 \mathrm{d}[s,s]_u) < \infty$ for all n, $\int_0^{\cdot} X_{u-}^n\,\mathrm{d}s_u$ is a $\mathbb{Q}$-martingale for each n. This yields $\mathbb{E}^{\mathbb{Q}}(Y_T^n) = Y_0^n + X_0^n S(0, X_0) - \mathbb{E}^{\mathbb{Q}}(L_T^n)$. Combining this with the fact that $L^n \geqslant 0$ for each n and $\mathbb{E}^{\mathbb{Q}}(Y_T^n - C) \to 0$ gives $\liminf_{n\to\infty} Y_0^n + X_0^n S(0, X_0) \geqslant \mathbb{E}^{\mathbb{Q}}(C)$ for all approximating sequences. However, as proven in Lemma 3, there exists some approximating sequence $(\overline{X}^n, \overline{Y}^n, \overline{\tau}^n)_{n\geqslant 1}$ with $\overline{L}^n = 0$ for all n. For this sequence, $\liminf_{n\to\infty} \overline{Y}_0^n + \overline{X}_0^n S(0, X_0) = \mathbb{E}^{\mathbb{Q}}(C)$. □

Remark 2.

1. The above value is consistent with no arbitrage. Indeed, suppose the contingent claim is sold at price $p > \mathbb{E}^{\mathbb{Q}}(C)$. Then, one can short the contingent claim at p and construct a sequence of continuous and of finite variation s.f.t.s., $(X^n, Y^n, \tau^n)_{n\geqslant 1}$, with $Y_0^n = \mathbb{E}^{\mathbb{Q}}(C)$, $X_0^n = 0$ and

$\lim_{n\to\infty} Y_T^n = C$ in L^2, hence, in probability, creating a FLVR. However, this is not allowed since $\mathbb{Q}$ is an equivalent martingale measure for s. Similarly, one can show that the price of the contingent claim cannot be less than $\mathbb{E}^{\mathbb{Q}}(C)$.

2. Given our supply curve formulation, this corollary implies that continuous trading strategies of finite variation can be constructed to both (i) approximately replicate any contingent claim, and (ii) avoid all liquidity costs. This elucidates the special nature of *continuous* trading strategies in a continuous time setting.

5 Example (extended Black–Scholes economy)

To illustrate the previous theory, we consider an extension of the Black–Scholes economy that incorporates liquidity risk. This example along with some empirical evidence regarding the pricing of traded options in the extended Black–Scholes economy can be found in Çetin et al. (2006).

5.1 The economy

Let

$$S(t,x) = e^{\alpha x} S(t,0) \quad \text{with } \alpha > 0, \tag{4}$$

$$S(t,0) \equiv \frac{s_0 e^{\mu t + \sigma W_t}}{e^{rt}}, \tag{5}$$

where μ, σ are constants and W is a standard Brownian motion initialized at zero.

For this section, let the spot rate of interest be constant and equal to r per unit time. The marginal stock price follows a geometric Brownian motion. The normalization by the money market account's value is made explicit in expression (5). Expressions (4) and (5) characterize an extended Black–Scholes economy. It is easy to check that this supply curve satisfies Assumption 1 in Section 2.

Under these assumptions, there exists a unique martingale measure for $S(\cdot, 0) = s$, see Duffie (1996). Hence, we know that the market is arbitrage-free and approximately complete.

5.2 Call option valuation

Consider a European call option with strike price K and maturity date T on this stock with cash delivery. Given cash delivery, in order to avoid liquidity costs at time T, the payoff[13] to the option at time T is selected to be $C_T = \max[S(T,0) - K, 0]$.

[13] The strike price needs to be normalized by the value of the money market account.

Under this structure, by the corollary to the second fundamental theorem of asset pricing, the value of a long position in the option is

$$C_0 = e^{-rT}\mathbb{E}^{\mathbb{Q}}\big(\max\big[S(T,0) - K, 0\big]\big).$$

It is well known that the expectation in this expression gives the Black–Scholes–Merton formula:

$$s_0 N\big(h(0)\big) - Ke^{-rT}N\big(h(0) - \sigma\sqrt{T}\,\big),$$

where $N(\cdot)$ is the standard cumulative normal distribution function and

$$h(t) \equiv \frac{\log s_t - \log K + r(T-t)}{\sigma\sqrt{T-t}} + \frac{\sigma}{2}\sqrt{T-t}.$$

Applying Itô's formula, the classical replicating strategy, $X = (X_t)_{t\in[0,T]}$, implied by the classical Black–Scholes–Merton formula is given by

$$X_t = N\big(h(t)\big). \tag{6}$$

This hedging strategy is continuous, but not of finite variation.

In this economy, we have that $(\frac{\partial S}{\partial x}(t,0) = \alpha e^0 s_t = \alpha s_t)$. Hence, although the call's value is the Black–Scholes formula, the standard hedging strategy will not attain this value. Indeed, using this strategy, it can be shown that the classical Black–Scholes hedge leads to the following nonzero liquidity costs [from expression (1)][14]:

$$L_T = X_0\big(S(0,X_0) - S(0,0)\big) + \int_0^T \frac{\alpha(N'(h(u)))^2 s_u}{T-u}\,\mathrm{d}u. \tag{7}$$

In contrast, an approximate hedging strategy that is continuous and of finite variation having zero liquidity costs is the sequence of s.f.t.s. $(X^n, Y^n, \tau^n)_{n\geqslant 1}$ with

$$X_t^n = 1_{[\frac{1}{n},T-\frac{1}{n})}(t)n \int_{(t-\frac{1}{n})^+}^{t} N\big(h(u)\big)\,\mathrm{d}u, \quad \text{if } 0 \leqslant t \leqslant T - \frac{1}{n},$$

$$X_t^n = \big(nTX^n_{(T-\frac{1}{n})} - nX^n_{(T-\frac{1}{n})}t\big), \quad \text{if } T - \frac{1}{n} \leqslant t \leqslant T, \tag{8}$$

and $Y_0^n = \mathbb{E}^{\mathbb{Q}}(C_T)$. In the limit, this trading strategy attains the call's time T value, i.e. $Y_T^n = Y_0^n + \int_0^T X_{u-}^n\,\mathrm{d}s_u \to C_T = \max[S(T,0) - K, 0]$ in $L^2(d\mathbb{Q})$.

[14] Note that both L_T and Y_T^n are normalized by the value of the money market account.

6 Economies with supply curves for derivatives

Extended first and second fundamental theorems hold in the above economy, with a C^2 supply curve for the stock and allowing continuous trading strategies, consequently, there is a unique price for any option on the stock. This implies that the supply curve for trading an option is horizontal, exhibiting no quantity impact on the price. Otherwise, there would exist arbitrage opportunities (given trading in options and the stock). This is inconsistent with practice. And, it seems that any model analyzing liquidity risk, should imply supply curves for both stocks and options.

The reason they exist for stocks, but not options in the above model, is that continuous trading strategies of finite variation enable the investor to avoid all liquidity costs in the stock. Hence, although liquidity costs exist, they are nonbinding, and the classical theory still applies (albeit in a modified and approximate manner). To make liquidity costs binding (as they are in practice), one must either remove the C^2 condition or disallow continuous trading strategies. The removal of the C^2 condition has been studied in the transaction cost literature (see Çetin, 2003; Barles and Soner, 1998; Constantinides and Zariphopoulou, 1999; Cvitanic and Karatzas, 1996; Cvitanic et al., 1999; Jouini, 2000; Jouini and Kallal, 1995; Jouini et al., 2001; Soner et al., 1995) and will be discussed here directly in Section 7, and in the context of estimating supply curves when summarizing Blais (2006) and Blais and Protter (2006) in Section 8 below. The exclusion of continuous trading strategies, but still retaining the C^2 condition, has been studied by Çetin et al. (2006). This exclusion is consistent with practice because continuous trading strategies are impossible to utilize, except as approximations via simple trading strategies. But, with simple trading strategies, liquidity costs are binding. We discuss this extension next.

We modify the previous theory by considering only the class of the *discrete trading strategies* defined as any simple s.f.t.s. X_t where

$$X_t \in \left\{ x_{\tau_0} 1_{\{\tau_0\}} + \sum_{j=1}^{N} x_{\tau_j} 1_{(\tau_{j-1}, \tau_j]} \;\middle|\; \begin{array}{l} 1.\ \tau_j \text{ are } \mathbb{F} \text{ stopping times for each } j \\ 2.\ x_{\tau_j} \text{ is in } \mathcal{F}_{\tau_{j-1}} \text{ for each } j \text{ (predictable)} \\ 3.\ \tau_0 \equiv 0 \text{ and } \tau_j > \tau_{j-1} + \delta \text{ for a fixed } \delta > 0 \end{array} \right\}.$$

These trading strategies are discontinuous because once a trade is executed, the subsequent trade is separated by a minimum of $\delta > 0$ time units, as in Cheridito (2003). For the remainder of the paper, lower case values x and y denote discrete trading strategies.

By restricting the class of trading strategies in this manner, we retain an arbitrage-free environment (the extended first fundamental theorem still applies), although the minimum distance δ between trades prevents the market from being approximately complete. In an incomplete (not approximately

complete market), the cost of replicating an option depends on the chosen trading strategy. Hence, the extended second fundamental theorem fails. This failure implies that there can be a quantity impact on the price of an option, i.e. the supply curve for an option need not be horizontal.

To investigate no arbitrage constraints on this supply curve, we can study the super-replication of options via the use of discrete trading strategies. For any discrete trading strategy, the liquidity cost equals

$$L_T = \sum_{j=0}^{N} [x_{\tau_{j+1}} - x_{\tau_j}]\big[S(\tau_j, x_{\tau_{j+1}} - x_{\tau_j}) - S(\tau_j, 0)\big]. \tag{9}$$

For a discrete trading strategy with $x_T = 0$, the hedging error is given by

$$C_T - Y_T = C_T - \left[y_0 + x_0 S(0,0) + \sum_{j=0}^{N-1} x_{\tau_{j+1}} \big[S(\tau_{j+1}, 0) - S(\tau_j, 0)\big] \right] + L_T.$$

Thus, there are two components to this hedging error. The first quantity,

$$\left[y_0 + x_0 S(0,0) + \sum_{j=0}^{N-1} x_{\tau_{j+1}} \big[S(\tau_{j+1}, 0) - S(\tau_j, 0)\big] \right] - C_T, \tag{10}$$

is the error in replicating the option's payoff C_T. The second component is the *liquidity cost* L_T defined in Equation (9).

To provide an upper bound on the price a particular quantity of options, one can investigate the minimum cost of super-replication. For a *single* call option on the stock, this cost can be obtained as follows. Define $Z_t = X_t S(t,0) + Y_t$ as the time t marked-to-market value of the replicating portfolio. The optimization problem is

$$\min_{(X,Y)} Z_0 \quad \text{s.t.} \quad Z_T \geqslant C_T = \max\big\{S(T,0) - Ke^{-rT}, 0\big\}, \tag{11}$$

where

$$Z_T = y_0 + x_0 S(0,0) + \sum_{j=0}^{N-1} x_{\tau_{j+1}} \big[S(\tau_{j+1}, 0) - S(\tau_j, 0)\big] - L_T.$$

The solution to this problem requires a numerical approximation. One such numerical procedure involving the binomial approximation is discussed in Çetin et al. (2006).[15]

[15] For multiple call options on the stock, the right side of the equation is premultiplied by the number of option shares.

Since liquidity costs in the underlying stock are quantity dependent, the cost of super-replication will also be quantity dependent. The cost of super-replicating a number of shares of the option then provides an upper bound on the entire supply curve for the option (as a function of the quantity constructed).

In Çetin et al. (2006), for various traded options, some empirical evidence is provided showing that the difference between the classical price and the super-replication cost to an option is economically significant.

7 Transaction costs

As previously stated, transaction costs can be viewed as a special case of our liquidity risk formulation where the $\mathcal{C}^2$ hypothesis on the supply curve is violated. This section provides the justification for this statement. We discuss three kinds of transaction costs, where all costs are *per share* unless otherwise stated. The three kinds are *fixed transaction costs, proportionate transaction costs*, and *mixed fixed and proportionate transaction costs*. We are now ignoring liquidity issues, and we use the mathematics of the supply curve framework to study only transaction costs. To emphasize this distinction, we now call the supply curve the *transaction curve*. Our goal is to see when continuous trading is feasible.[16] This section largely follows Umut Çetin's thesis (Çetin, 2003).

Definition 11. We define three kinds of transaction costs:

1. *Fixed transaction costs* are defined by a transaction curve giving the (per share stock price) by
$$S(t, x) = S(t, 0) + \frac{a}{x}.$$
2. *Proportionate transaction costs* depend proportionately on the dollar value of the trade, and are given by
$$S(t, x) = S(t, 0)\big(1 + \beta \operatorname{sign}(x)\big),$$
where $\beta > 0$ is the proportionate transaction cost per unit value.
3. *Combined fixed and proportionate transaction costs* vary with the specific application. Two examples are[17]:

[16] Obviously in practice continuous trading is not truly feasible, since one cannot physically trade continuously. However, there remains the issue of whether one would desire to approximate a continuous trading strategy arbitrarily closely with discrete trading strategies. If continuous trading strategies have infinite transaction costs, then any approximating sequence would have unboundedly large transaction costs, and be undesirable to utilize. It is the desirability of using approximating sequences that is really being investigated below.

[17] These characterizations were obtained from information provided on both the Fidelty and Vanguard web sites in 2004.

- Fidelity

$$S(t, x) = S(t, 0) + \frac{\beta}{x} + \text{sign}(x)\gamma 1_{\{|x|>\delta\}},$$

where β, γ and δ are positive constants;

- Vanguard

$$S(t, x) = S(t, 0) + \frac{\max\{\alpha, |x|c\}}{x},$$

where α and c are positive constants.

Our first result implies that in the presence of fixed transaction costs, only piecewise constant trading strategies need be considered for modeling purposes.

Theorem 4. *Continuous trading in the presence of fixed transaction costs creates infinite costs in finite time.*

Proof. We assume the structure given in Definition 11, part 1. First assume that our trading strategy X is of the form $X = \sum_{i=0}^{n-1} X_i 1_{[T_i, T_{i+1})}$, for a sequence of trading times $0 = T_0 \leqslant T_1 \leqslant \cdots \leqslant T_n = T$. Then the cumulative trading costs are $\sum_{i=0}^{n-1} a 1_{\{X_i \neq X_{i-1}\}}$, and if we further assume that always $X_i \neq X_{i-1}$, then it is equal simply to na.

Now suppose our trading strategy X has continuous paths. Let $TC(X)$ denote the transaction costs of following the strategy X. We have

$$TC(X) = \limsup_{n\to\infty} \sum_{T_i^n \in \Pi_k} a 1_{\{X_{T_i^n} \neq X_{T_{i+1}^n}\}} = \limsup_{n\to\infty} a N_{\Pi_n}(X),$$

where Π_n is a sequence of random partitions tending to the identity[18] and $N_{\Pi_n}(X)$ is the number of times that $X_{T_i^n} \neq X_{T_{i-1}^n}$ for the random stopping times of Π_n. Note that $\limsup_{n\to\infty} N_{\Pi_n}(X) = \infty$ unless X is a.s. piecewise constant. Thus continuous trading strategies incur infinite transaction costs. Finally, if our trading strategy has both jumps and continuous parts to it, the transaction costs will exceed the costs for each part, hence will also be infinite. □

The situation for proportional transaction costs is more complicated. Here it is possible to trade continuously, provided one follows a trading strategy with paths of finite variation (which is not the case, for example, with the standard Black–Scholes hedge of a European call or put option).

[18] This is terminology from Protter (2005); it means that each Π_n is a finite increasing sequence of stopping times covering the interval $[0, T]$, and the mesh of Π_n tends to 0 as $n \to \infty$.

Theorem 5. *Continuous trading in the presence of proportional transaction costs is infinite if the trading strategy has paths of infinite variation. If the strategy X has paths of finite variation on $[0, T]$ for a subset Λ of Ω, then the cumulative transaction costs are $b\int_0^T S(s,0)|\mathrm{d}X_s|$ a.s. on Λ, where $|\mathrm{d}X_s|$ denotes the total variation Stieltjes path by path integral, and they are infinite on Λ^c. (b is the constant in Definition 11, part 2.)*

Proof. Let Π_n be a sequence of random partitions tending to the identity on $[0, T]$. Let X be a continuous trading strategy. Then the cumulative transaction costs for proportional costs can be written as

$$TC(X) = \limsup_{n\to\infty} \sum_{T_i^n \in \Pi_k} S\big(T_k^n, 0\big)|\Delta X_{n,k}|b,$$

where $\Delta X_{n,k} = X_{T_k^n} - X_{T_{k-1}^n}$. When X has paths of finite variation, this converges to the path by path Stieltjes integral $b\int_0^T S(s,0)|\mathrm{d}X_s|$, and when X paths of infinite variation is diverges to ∞. Since it is a path by path result, we deduce the theorem. □

Theorem 6. *Continuous trading in the presence of combined fixed and proportional transaction costs creates infinite costs in finite time.*

Proof. Suppose our trading strategy X has continuous paths. Let $TC(X)$ denote the transaction costs of following the strategy X. We have

$$TC(X) \geqslant \limsup_{n\to\infty} \sum_{T_i^n \in \Pi_k} \delta 1_{\{X_{T_i^n} \neq X_{T_{i+1}^n}\}} \geqslant \limsup_{n\to\infty} \delta N_{\Pi_n}(X),$$

for some constant δ, and where Π_n is a sequence of random partitions tending to the identity, and $N_{\Pi_n}(X)$ is the number of times that $X_{T_i^n} \neq X_{T_{i-1}^n}$ for the random stopping times of Π_n. This leads to infinite costs as in the proof of Theorem 4. □

8 Examples of supply curves

We now discuss recent results of Blais (2006) and Blais and Protter (2006). These results are inspired by an analysis of a trading book, provided to Blais and Protter by Morgan Stanley, via the good office of Robert Ferstenberg (see Ferstenberg, 2004). See Blais (2006) and Blais and Protter (2006) for a detailed description of the data and its more profound implications. Note that the classical theory, with unlimited liquidity, is embedded in the structure previously discussed, using a standard price process $S_t = S(t, 0)$. In this case the supply curve

$$x \to S(t,x) \quad \text{reduces to} \quad x \to S(t,0),$$

that is, it is a line with slope 0 and vertical axis intercept $S(t, 0)$. If one supposes that the supply curve is linear, that is of the form

$$x \to S(t, x) = M_t x + b_t,$$

then if the classical theory is accurate one must have $M_t = 0$. Taking this as the null hypothesis, Blais (2006) has shown that it can be rejected at the 0.9999 significance level. From this we conclude that the supply curve exists and is nontrivial.

Using linear regression, Blais (2006) has further shown that for *liquid stocks*[19] the supply curve is linear, with time varying slope and intercept; thus for liquid stocks the supply curve can be written

$$x \to S(t, x) = M_t x + b_t,$$

where $b_t = S(t, 0)$. Moreover it is reasonable to assume that $(M_t)_{t \geqslant 0}$ is itself a stochastic process with continuous paths. We have then the following theorem for this case, which is a special case of Theorem 11 of Appendix A.

Theorem 7. *For a liquid stock with linear supply curve of the form*

$$x \to S(t, x) := M_t x + b_t,$$

and a càdlàg trading strategy X with finite quadratic variation, the value in the money market account for a self financing trading strategy is given by

$$Y_t = -X_t S(t, 0) + \int_0^t X_{u-} \, \mathrm{d}S(u, 0) - \int_0^t M_u \, \mathrm{d}[X, X]_u.$$

Note that in this theorem, the quadratic variation differential term can have jumps.

The case for *nonliquid stocks* presents a new problem, and the previously established theory breaks down at one particular point, because the standing hypothesis that the supply curve $x \to S(t, x)$ is $\mathcal{C}^2$ no longer holds. Indeed, in this case the data shows that the supply curve is jump linear, with one jump, which can be thought of as the bid–ask spread. Fortunately the only place Çetin et al. (2004) use the $\mathcal{C}^2$ hypothesis is in the derivation of the self financing strategy, and in the jump linear case the simple structure allows Blais and Protter to eliminate this hypothesis.[20] Since we no longer assume the supply curve is continuous, we can let $S(t, 0^-)$ denote the marginal ask, and $S(t, 0^+)$ will denote the marginal bid, whence we can let $\gamma(t) = S(t, 0^+) - S(t, 0^-)$ denote the

[19] We deliberately leave the definition of a "liquid stock" vague. For a precise definition, see Blais and Protter (2006). Examples of liquid stocks include BP, ATT, IBM.

[20] The $\mathcal{C}^2$ hypothesis of the supply curve in the space variable is of course also not needed for the linear supply curve case, and thus in practice perhaps it is not needed at all.

bid/ask spread. Assume that the supply curve has a jump linear form given by

$$S(t,x) = \begin{cases} \beta(t)x + S(t,0^+) & (x \geqslant 0), \\ \alpha(t)x + S(t,0^-) & (x < 0), \end{cases} \tag{12}$$

where α and β assumed to be continuous stochastic processes. Next define

$$b^+(t) = S(t,0^+) \quad \text{and} \quad b^-(t) = S(t,0^-) \tag{13}$$

and assume that both b^+ and b^- are (random and) continuous functions of t. In the next theorem, we restrict our attention to trading strategies with paths of finite variation, which is reasonable since we are considering the illiquid case. For such a trading strategy X let

$$X_t = X_0 + C_t - A_t \tag{14}$$

which is the path-by path Lebesgue decomposition of X into the difference of two monotone nondecreasing processes with disjoint supports. Note that both A and C can of course contain jumps. Note also that for such an X, $[X,X]_t = \sum_{s \leqslant t} (\Delta X_s)^2$. We have in this case the following result (Blais and Protter, 2006):

Theorem 8. *For an illiquid stock with jump linear supply curve of the form given in Equation (12) and a càdlàg trading strategy X with paths of finite variation (of the form (14)), the value in the money market account for a self financing trading strategy is given by*

$$\begin{aligned} Y_t = Y_0 - X_t S(t,0^+) &+ \int_0^t X_{u-}\, \mathrm{d}S(u,0^+) \\ &- \int_0^t \left\{ \beta(s) 1_{\Lambda^c}(s) + \alpha(s) 1_{\Lambda}(s) \right\} \mathrm{d}[X,X]_s \\ &- \int_0^t \left\{ b^+(s) 1_{\Lambda^c}(s) - b^-(s) 1_{\Lambda}(s) \right\} \mathrm{d}X_s, \end{aligned}$$

where Λ denotes the (random) support of the increasing process A defined in (14) above.

Proof. It is clear that the money market process Y should satisfy

$$\begin{aligned} Y_t &= Y_0 - \lim_{n \to \infty} \sum_{k \geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) \\ &= -X(0) S(0, X_0) \end{aligned}$$

$$- \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) \big[S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big]$$

$$- \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) S\big(T_k^n, 0\big).$$

We know from Example 2 (given in Appendix A) that the last sum converges to $-X_0 S(0,0) + X_t S(t,0) - \int_0^t X_{u-}\, \mathrm{d}S(u,0)$. Let us thus focus on the term

$$- \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) \big[S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big]. \tag{15}$$

Due to our jump linear hypothesis on the structure of the supply curve, we can re-write the sum in expression (15) as:

$$\begin{aligned}
&\sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) \big[S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big] \\
&\quad = \sum_{k\geqslant 1} \Delta X_{n,k} 1_{\{\Delta X_{n,k}\geqslant 0\}} \big[\beta\big(T_k^n\big) \Delta X_{n,k} - b^+\big(T_k^n\big) \big] \\
&\qquad + \sum_{k\geqslant 1} \Delta X_{n,k} 1_{\{\Delta X_{n,k}<0\}} \big[\alpha\big(T_k^n\big) \Delta X_{n,k} - b^-\big(T_k^n\big) \big] \\
&\quad = \sum_{k\geqslant 1} (\Delta X_{n,k})^2 1_{\{\Delta X_{n,k}\geqslant 0\}} \beta\big(T_k^n\big) + \sum_{k\geqslant 1} (\Delta X_{n,k})^2 1_{\{\Delta X_{n,k}<0\}} \alpha\big(T_k^n\big) \\
&\qquad - \sum_{k\geqslant 1} \Delta X_{n,k} 1_{\{\Delta X_{n,k}\geqslant 0\}} b^+\big(T_k^n\big) - \sum_{k\geqslant 1} \Delta X_{n,k} 1_{\{\Delta X_{n,k}<0\}} b^-\big(T_k^n\big),
\end{aligned}$$

where we have written $\Delta X_{n,k}$ as a shorthand for $X_{T_k^n} - X_{T_{k-1}^n}$.

Next we take the limits as indicated in expression (15) and using that b^+ and b^- are both continuing, it follows from standard theorems from stochastic calculus (see, e.g., Protter, 2005) that we get convergence uniformly on compact time sets in probability to the expression

$$- \int_0^t \big\{ \beta(s) 1_{\Lambda^c}(s) + \alpha(s) 1_{\Lambda}(s) \big\}\, \mathrm{d}[X,X]_s$$

$$- \int_0^t \big\{ b^+(s) 1_{\Lambda^c}(s) - b^-(s) 1_{\Lambda}(s) \big\}\, \mathrm{d}X_s$$

and the result follows. □

9 Conclusion

This paper reviews the work of Çetin (2003), Çetin et al. (2004, 2006), Blais (2006), and Blais and Protter (2006) which extends classical arbitrage pricing theory to include liquidity risk. This is accomplished by studying an economy where the security's price depends on the trade size. Extended first and second fundamental theorems of asset pricing are shown to hold. For the first theorem, the economy is shown to be arbitrage free if and only if the stochastic process for the price of a marginal trade has an equivalent martingale probability measure. The second fundamental theory of asset pricing also approximately holds: markets will be approximately complete if the martingale measure is unique. In an approximately complete market, derivative prices are shown to equal the classical arbitrage free price of the derivative security. This implies horizontal supply curves for a derivative on the stock. To obtain upward sloping supply curves for derivatives, continuous trading strategies need to be excluded. This extension implies an incomplete market. Minimal cost super-replicating trading strategies are discussed in this regard. Last, an analysis of the theory and how it applies to data in both the liquid and illiquid cases is reviewed.

Acknowledgement

The authors with to thank Sam Ehrlichman for pointing out a mistake in the original formulation of Theorem 8.

Appendix A

A.1 Proof of the first fundamental theorem

This theorem uses Assumption 1, sample path continuity of $S(t, x)$. The proof proceeds in two steps. Step 1 constructs a fictitious economy where all trades are executed at the marginal stock price. The theorem is true in this fictitious economy by the classical result. Step 2 then shows the theorem in this fictitious economy is sufficient to obtain the result in our economy.

Prior to this proof, we need to make the following observation in order to utilize the classical theory. The classical theory (see Delbaen and Schachermayer, 1994 or alternatively Protter, 2001 for an expository version) has trading strategies starting with $X_0 = 0$, while we have trading strategies with $X_{0-} = 0$ but not $X_0 = 0$. Without loss of generality, in the subsequent proof, we can restrict ourselves to predictable processes with $X_0 = 0$. Here is the argument. Recall $s_u = S(u, 0)$. In our setup, choose Y^0 so that $X_0 S(0,0) + Y_0^0 = 0$ and $X_t S(t,0) + Y_t^0 = X_0 S(0,0) + Y_0^0 + \int_{0+}^{T} X_u \, \mathrm{d}s_u = \int_{0+}^{T} X_u \, \mathrm{d}s_u$. Define

$\widehat{X} = 1_{(0,T]}X$. $\widehat{X}$ is predictable, $\widehat{X}_0 = 0$, and $\int_{0+}^{T} X_u \, \mathrm{d}s_u = \int_0^T \widehat{X}_u \, \mathrm{d}s_u$. The analysis can be done for $\widehat{X}$.

A.1.1 ***Step 1.*** *The fictitious economy*

Consider the fictitious economy introduced in Section 3. Delbaen and Schachermayer prove the following in Section 4 of (Delbaen and Schachermayer, 1994):

Theorem 9. *Given Assumption 1 and no arbitrage, there is NFLVR in the fictitious economy if and only if there exists a measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot, 0)$ is a $\mathbb{Q}$-local martingale.*

Since the stochastic integral of a predictable process can be written as a limit (uniformly on compacts in probability) of stochastic integrals with continuous and finite variation integrands (see Appendix A.3 below), we have the following corollary.[21]

Corollary 2. *Suppose there is no arbitrage opportunity in the fictitious economy. Given Assumption 1, if there is an (FLVR) in the fictitious economy, there exists a sequence of ϵ_n-admissible trading strategies X^n, continuous and of FV, and a nonnegative F_T-measurable random variable f_0, not identically zero, such that $\epsilon_n \to 0$ and $(X^n \cdot S)_T \to f_0$ in probability.*

The proof of this corollary is straightforward, and hence we omit it.

A.1.2 ***Step 2.*** *The illiquid economy*

In the economy with liquidity risk, restricting consideration to s.f.t.s. (X, Y, τ) with X finite variation and continuous processes, by Lemma 1, we have that $Y_t = (X \cdot s)_t - X_t S(t, 0)$. At time T, we have $Y_T = (X \cdot s)_T$. This is the value of the same s.f.t.s. in the fictitious economy. We use this insight below.

Lemma 4. *Given Assumption 1, let X be an α-admissible trading strategy which is continuous and of FV in the fictitious economy. Then there exists a sequence of $(\alpha + \epsilon_n)$-admissible trading strategies, in the illiquid economy, $(H^n, Y^n, \tau^n)_{n \geqslant 1}$ of FV and continuous on $[0, \tau^n)$, such that Y_T^n tends to $(X \cdot S)_T$, in probability, and $\epsilon_n \to 0$.*

[21] In the original paper (Delbaen and Schachermayer, 1994), there is a missing hypothesis in the statement of their theorem related to this corollary. We include here and in other results as needed the missing hypothesis of no arbitrage. We are grateful to Professor Delbaen for providing us with a counterexample that shows one does in fact need this hypothesis (Delbaen, 2003).

Proof. Let $T_n = T - \frac{1}{n}$. Define

$$f_n(t) = 1_{[T_n \leqslant t \leqslant T_{n+1}]} \frac{X_{T_n}}{T_n - T_{n+1}} (t - T_{n+1}) \tag{A.1}$$

so that $f_n(T_n) = X_{T_n}$ and $f_n(T_{n+1}) = 0$. Note that $f_n(t) \to 0$, a.s., $\forall t$. Define

$$X_t^n = X_t 1_{[t<T_n]} + f_n(t). \tag{A.2}$$

By this definition, X^n is continuous and of FV. Note that T is a fixed time and not a stopping time, so X^n is predictable. Moreover,

$$\big(X^n \cdot S\big)_t = (X \cdot S)_{t \wedge T_n} + \int_0^t f_n(s)\, \mathrm{d}S(s, 0). \tag{A.3}$$

Notice that $|f_n(\omega)| \leqslant \sup_t |X_t(\omega)| \equiv K(\omega) \in R$ since X is continuous on $[0, T]$. Thus, f_n is bounded by an $S(\cdot, 0)$-integrable function. Therefore, by dominated convergence theorem for stochastic integrals (see Protter, 2005, p. 145) $\int f_n(s)\, \mathrm{d}S(s, 0)$ tends to 0 in u.c.p. on the compact time interval $[0, T]$, and therefore $X^n \cdot S \to X \cdot S$ in u.c.p. on $[0, T]$.[22]

Now, let $(\epsilon_n)_{n \geqslant 1}$ be a sequence of positive real numbers converging to 0 such that $\sum_n \epsilon_n < \infty$. Define $\tau^n = \inf\{t > 0\colon (X^n \cdot S)_t < -\alpha - \epsilon_n\} \wedge T$. τ^n is a predictable stopping time by the continuity of $S(\cdot, x)$. Due to u.c.p. convergence of $X^n \cdot S$ to $X \cdot S$, passing to a subsequence if necessary, we have the following:

$$\mathbb{P}\Big(\sup_{0 \leqslant t \leqslant T} \big|\big(X^n \cdot S\big)_t - (X \cdot S)_t\big| \geqslant \epsilon_n \Big) \leqslant \epsilon_n. \tag{A.4}$$

Notice that $\mathbb{P}(\tau^n < T) \leqslant \epsilon_n$, i.e. $\tau^n \to T$ in probability. Moreover, $\tau^n \geqslant T_n$ because $X^n = X$ up to time T_n. Choose $H^n = X^n 1_{[0, \tau^n)}$. Consider the sequence of trading strategies $(H^n, \tau^n)_{n \geqslant 1}$. Note that $(H^n \cdot S)_t \geqslant -\alpha - \varepsilon_n$ for all $t \in [0, \tau^n]$ since $H^n_{\tau^n} = 0$ for all n. Therefore, $(H^n, \tau^n)_{n \geqslant 1}$ is a sequence of $(\alpha + \epsilon_n)$-admissible trading strategies. The value of the portfolio at liquidation for each trading strategy is given by

$$Y^n_{\tau^n} = X^n\big(\tau^n\big)\big[S\big(\tau^n, -X^n\big(\tau^n\big)\big) - S\big(\tau^n, 0\big)\big] + \big(X^n \cdot S\big)_{\tau^n} \tag{A.5}$$

since H^n is of FV and jumps only at τ^n for each n by the continuity of X^n. Therefore, it remains to show $X^n(\tau^n) \to 0$ in probability since this, together with $\tau^n \to T$ in probability, will prove the theorem. Indeed, $\sum_n \mathbb{P}(\tau^n < T) \leqslant \sum_n \epsilon_n < \infty$. Therefore, by the first Borel–Cantelli lemma, $\mathbb{P}[\tau^n < T \text{ i.o.}] = 0$, which implies $X^n(\tau^n) = X^n(T) = 0$, with probability 1, for all but at most finitely many n. □

[22] One can also show this using integration by parts.

Lemma 5. *Suppose there is no arbitrage opportunity in the fictitious economy. Given Assumption 1, there is NFLVR in the fictitious economy if and only if there is NFLVR in the illiquid economy.*

Proof. Suppose there is NFLVR in the fictitious economy. Since, given any s.f.t.s. (X, Y, τ) in the illiquid economy, $Y_\tau \leqslant (X \cdot S)_\tau$, it follows there exists NFLVR in the illiquid economy. Conversely, suppose there is FLVR in the fictitious economy. In view of Corollary 2, there is a sequence, $(X^n)_{n\geqslant 1}$, with each X^n continuous, of FV, and ϵ_n-admissible trading strategies such that $(X^n \cdot S)_T \to f_0$ in probability where f_0 is as before and $\epsilon_n \to 0$. However, by the previous lemma, there exists a sequence of α_n-admissible trading strategies, $(H^n, Y^n, \tau^n)_{n\geqslant 1}$, where $\alpha_n \to 0$, in the illiquid economy such that $Y^n_{\tau^n} \to f_0$ in probability, which gives an FLVR in the illiquid economy. □

Theorem 10 *(First fundamental theorem). Suppose there is no arbitrage opportunity in the fictitious economy. Given Assumption 1, there is no free lunch with vanishing risk (NFLVR) in the illiquid economy if and only if there exists a measure $\mathbb{Q} \sim \mathbb{P}$ such that $S(\cdot, 0)$ is a $\mathbb{Q}$-local martingale.*

Proof. By the previous lemma, (NFLVR) in the illiquid economy is equivalent to (NFLVR) in the fictitious economy, which is equivalent to existence of a martingale measure by Theorem 9. □

A.2 Construction of the self-financing condition for a class of trading strategies

The purpose of this section is to provide justification for Definition 2 in the text. This proof uses only the weaker hypotheses of Assumption 2.[23]

Let t be a fixed time and let (σ_n) be a sequence of random partitions of $[0, t]$ tending to identity in the following form:

$$\sigma_n\colon\ 0 = T^n_0 \leqslant T^n_1 \leqslant \cdots \leqslant T^n_{k_n} = t,$$

where T^n_ks are stopping times. For successive trading times, t_1 and t_2, the self-financing condition can be written as

$$Y_{t_2} - Y_{t_1} = -(X_{t_2} - X_{t_1})\big[S(t_2, X_{t_2} - X_{t_1})\big].$$

Note that $Y_t = Y_0 + \sum_{k\geqslant 1}(Y_{T^n_k} - Y_{T^n_{k-1}})$ for all n. Therefore, we will define Y_t to be the following limit whenever it exists:

$$Y_0 - \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T^n_k} - X_{T^n_{k-1}}) S\big(T^n_k, X_{T^n_k} - X_{T^n_{k-1}}\big). \tag{A.6}$$

[23] Note that we have already justified the notion of a self financing strategy in the jump linear illiquid case of Section 8.

Example 2. In the classical case, $S(t,x) = S(t,0)$ for all $x \in \mathbb{R}$. Thus, self-financing condition becomes

$$Y_{t_2} - Y_{t_1} = -[X_{t_2} - X_{t_1}]S(t_2, 0)$$

and initial trades must satisfy $Y(0) = -X(0)S(0,0)$ instead. Therefore,

$$\begin{aligned}
Y_t &= Y_0 - \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) S\big(T_k^n, 0\big) \\
&= Y(0) - \lim_{n\to\infty} \Bigg[\sum_{k\geqslant 1} X_{T_k^n} S\big(T_k^n, 0\big) - \sum_{k\geqslant 1} X_{T_{k-1}^n} S\big(T_k^n, 0\big) \Bigg] \\
&= Y(0) - \lim_{n\to\infty} \Bigg[\sum_{k\geqslant 1} X_{T_k^n} S\big(T_k^n, 0\big) \\
&\quad - \sum_{k\geqslant 1} X_{T_{k-1}^n} \big(S\big(T_k^n, 0\big) - S\big(T_{k-1}^n, 0\big)\big) - \sum_{k\geqslant 1} X_{T_{k-1}^n} S\big(T_{k-1}^n, 0\big) \Bigg] \\
&= Y_0 - X_t S(t,0) + X_0 S(0,0) \\
&\quad + \lim_{n\to\infty} \sum_{k\geqslant 1} X_{T_{k-1}^n} \big(S\big(T_k^n, 0\big) - S\big(T_{k-1}^n, 0\big)\big) \\
&= -X_t S(t,0) + \int_0^t X_{u-}\, \mathrm{d}S(u,0).
\end{aligned}$$

Notice that the limit agrees with the value of $Y(t)$ in classical case. So, we have a framework that contains the existing theory.

Theorem 11. *For X càdlàg and has finite quadratic variation (QV), the value in the money market account is given by*

$$\begin{aligned}
Y_t = -X_t S(t,0) &+ \int_0^t X_{u-}\, \mathrm{d}S(u,0) - \int_0^t S_x^{(1)}(u-,0)\, \mathrm{d}[X,X]_u^c \\
&- \sum_{0\leqslant u\leqslant t} \big[S(u, \Delta X_u) - S(u,0)\big] \Delta X_u,
\end{aligned} \tag{A.7}$$

where $S_x^{(n)}$ is the nth partial derivative of S with respect to x.

Proof. The proof of this theorem is reminiscent of the proof of Theorem 8. Expression (A.6) is

$$Y_t = Y_0 - \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big)$$

$$
\begin{aligned}
&= -X(0)S(0, X_0) \\
&\quad - \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) \big[S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big] \\
&\quad - \lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) S\big(T_k^n, 0\big).
\end{aligned}
$$

We know from Example 2 that the last sum converges to $-X_0 S(0,0) + X_t S(t,0) - \int_0^t X_{u-}\, \mathrm{d}S(u,0)$. Let $A = A(\epsilon, t)$ be a set of jumps of X that has a.s. a finite number of times s, and let $B = B(\epsilon, t)$ be such that $\sum_{s\in B} (\Delta X_s)^2 \leqslant \epsilon^2$, where A and B are disjoint and $A \cup B$ exhaust the jumps of X on $(0, t]$, see proof of Itô's formula in Protter (2005). Thus,

$$
\begin{aligned}
&\lim_{n\to\infty} \sum_{k\geqslant 1} (X_{T_k^n} - X_{T_{k-1}^n}) \big[S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big] \\
&= \lim_{n\to\infty} \sum_{k,A} (X_{T_k^n} - X_{T_{k-1}^n}) \big(S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big) \\
&\quad + \lim_{n\to\infty} \sum_{k,B} (X_{T_k^n} - X_{T_{k-1}^n}) \big(S\big(T_k^n, (X_{T_k^n} - X_{T_{k-1}^n})\big) - S\big(T_k^n, 0\big) \big),
\end{aligned}
$$

where $\sum_{k,A}$ denotes $\sum_{k\geqslant 1} 1_{[A\cap (T_{k-1}^n, T_k^n] \neq \emptyset]}$, and $\sum_{k,B}$ denotes $\sum_{k\geqslant 1} 1_{[B\cap (T_{k-1}^n, T_k^n] = \emptyset]}$ Since A has only finitely many elements, ω by ω, the first limit equals

$$
\sum_{u\in A} \big[S(u, \Delta X_u) - S(u, 0) \big] \Delta X_u. \tag{A.8}
$$

Applying Taylor's formula to each $S(T_k^n, \cdot)$, the second limit becomes

$$
\begin{aligned}
&\lim_{n\to\infty} \sum_{k,B} S_x^{(1)}\big(T_k^n, 0\big) (X_{T_k^n} - X_{T_{k-1}^n})^2 \\
&\quad + \lim_{n\to\infty} \sum_{k,B} (X_{T_k^n} - X_{T_{k-1}^n}) R\big(T_k^n, |X_{T_k^n} - X_{T_{k-1}^n}|\big) \\
&\quad = \lim_{n\to\infty} \sum_{k\geqslant 1} S_x^{(1)}\big(T_k^n, 0\big) (X_{T_k^n} - X_{T_{k-1}^n})^2 \\
&\qquad - \lim_{n\to\infty} \sum_{k,A} S_x^{(1)}\big(T_k^n, 0\big) (X_{T_k^n} - X_{T_{k-1}^n})^2 \\
&\qquad + \lim_{n\to\infty} \sum_{k,B} (X_{T_k^n} - X_{T_{k-1}^n}) R\big(T_k^n, |X_{T_k^n} - X_{T_{k-1}^n}|\big) \\
&\quad = \lim_{n\to\infty} \sum_{k\geqslant 1} S_x^{(1)}\big(T_{k-1}^n, 0\big) (X_{T_k^n} - X_{T_{k-1}^n})^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \lim_{n\to\infty} \sum_{k\geqslant 1} \bigl[S_x^{(1)}\bigl(T_k^n, 0\bigr) - S_x^{(1)}\bigl(T_{k-1}^n, 0\bigr)\bigr](X_{T_k^n} - X_{T_{k-1}^n})^2 \\
&- \lim_{n\to\infty} \sum_{k,A} S_x^{(1)}\bigl(T_k^n, 0\bigr)(X_{T_k^n} - X_{T_{k-1}^n})^2 \\
&+ \lim_{n\to\infty} \sum_{k,B} (X_{T_k^n} - X_{T_{k-1}^n}) R\bigl(T_k^n, |X_{T_k^n} - X_{T_{k-1}^n}|\bigr), \qquad \text{(A.9)}
\end{aligned}
$$

where R is the remainder term in Taylor's formula. The sum of the first three limits converges to[24]

$$
\begin{aligned}
&\int_0^t S_x^{(1)}(u-, 0)\,\mathrm{d}[X, X]_u + \bigl[S_x^{(1)}(\cdot, 0), [X, X]\bigr]_t - \sum_{u\in A} S_x^{(1)}(u, 0)(\Delta X_u)^2 \\
&\quad = \int_0^t S_x^{(1)}(u-, 0)\,\mathrm{d}[X, X]_u + \sum_{0<u\leqslant t} \Delta S_x^{(1)}(u, 0)(\Delta X_u)^2 \\
&\qquad - \sum_{u\in A} S_x^{(1)}(u, 0)(\Delta X_u)^2. \qquad \text{(A.10)}
\end{aligned}
$$

Now we will show as ϵ tends to 0, the last term in (A.9) vanishes. Assuming temporarily that $|S_x^{(2)}| < K < \infty$ uniformly in x and t,

$$
\begin{aligned}
&\bigl|R\bigl(T_k^n, |X_{T_k^n} - X_{T_{k-1}^n}|\bigr)\bigr| \\
&\quad \leqslant \sup_{0\leqslant|x|\leqslant|X_{T_k^n} - X_{T_{k-1}^n}|} \bigl|S_x^{(1)}\bigl(T_k^n, x\bigr) - S_x^{(1)}\bigl(T_k^n, 0\bigr)\bigr| |(X_{T_k^n} - X_{T_{k-1}^n})| \\
&\quad \leqslant \sup_{0\leqslant|y|\leqslant|x|\leqslant|X_{T_k^n} - X_{T_{k-1}^n}|} \bigl|S_x^{(2)}\bigl(T_k^n, y\bigr)x(X_{T_k^n} - X_{T_{k-1}^n})\bigr| \\
&\quad \leqslant K(X_{T_k^n} - X_{T_{k-1}^n})(X_{T_k^n} - X_{T_{k-1}^n}),
\end{aligned}
$$

where the second inequality follows from the Mean Value Theorem. Therefore, the last sum in (A.9) is less than or equal to, in absolute value,

$$
\begin{aligned}
&K \lim_{n\to\infty} \sum_{k,B} \bigl(|X_{T_k^n} - X_{T_{k-1}^n}|\bigr)^3 \\
&\quad < K \lim_{n\to\infty} \sup_{k,B} |X_{T_k^n} - X_{T_{k-1}^n}| \sum_k \bigl(|X_{T_k^n} - X_{T_{k-1}^n}|\bigr)^2 \\
&\quad \leqslant K\epsilon[X, X]_t.
\end{aligned}
$$

[24] Note that the assumption that $S_x^{(1)}(\cdot, 0)$ has a finite QV is not needed when $S_x^{(1)}(\cdot, 0)$ is continuous. In this case, the second limit is zero. This follows from the fact that X has a finite QV and $S_x^{(1)}(\cdot, 0)$ is uniformly continuous, ω by ω, over the compact domain $[0, T]$.

Note that ϵ can be made arbitrarily small and X has a finite QV. Furthermore, since all summands are positive, as $\epsilon \to 0$, (A.8) converges to

$$\sum_{0<u\leqslant t} \big[S(u, \Delta X_u) - S(u, 0)\big]\Delta X_u$$

and (A.10) converges to

$$\begin{aligned}
&\int_0^t S_x^{(1)}(u-, 0)\, \mathrm{d}[X, X]_u + \sum_{0<u\leqslant t} \Delta S_x^{(1)}(u, 0)(\Delta X_u)^2 \\
&\quad - \sum_{0<u\leqslant t} S_x^{(1)}(u, 0)(\Delta X_u)^2 \\
&\quad = \int_0^t S_x^{(1)}(u-, 0)\, \mathrm{d}[X, X]_u - \sum_{0<u\leqslant t} S_x^{(1)}(u-, 0)(\Delta X_u)^2 \\
&\quad = \int_0^t S_x^{(1)}(u-, 0)\, \mathrm{d}[X, X]_u^c.
\end{aligned}$$

For the general case, let $V_k^x = \inf\{t > 0\colon\ S^{(2)}(t, x) > k\}$. Define $\tilde{S}(t, x) := S(t, x)1_{[0,V_k^x)}$. Therefore, (A.7) holds for $\tilde{S}$, for each k. Now, a standard argument using set unions, as in the proof of Itô's formula in Protter (2005), establishes (A.7) for S. □

A.3 Approximating stochastic integrals with continuous and of FV integrands

The next lemma (Lemma 6) is well known and can be found in Protter (2005).

Lemma 6. *Let X be a special semimartingale with the canonical decomposition $X = N + A$, where N is a local martingale and A is predictable. Suppose S has totally inaccessible jumps. Then A is continuous.*

We make the following assumption. (Note that this assumption is satisfied in all classical market models studies, since [for example] a Lévy process has only totally inaccessible jumps, and indeed by a classic theorem of P.A. Meyer, all "reasonable" strong Markov processes have only totally inaccessible jumps.)

Assumption 2. $S(\cdot, 0)$ has only totally inaccessible jumps.

We recall a few definitions.

Definition 12. Let X be a special semimartingale with canonical decomposition $X = \bar{N} + \bar{A}$. The $\mathcal{H}^2$ norm of X is defined to be

$$\|X\|_{\mathcal{H}^2} = \left\|[\bar{N}, \bar{N}]_\infty^{1/2}\right\|_{L^2} + \left\| \int_0^\infty |\mathrm{d}\bar{A}_u| \right\|_{L^2}.$$

The space $\mathcal{H}^2$ of semimartingales consists of all special semimartingales with finite $\mathcal{H}^2$ norm.

Definition 13. The predictable σ-algebra $\mathcal{P}$ on $R_+ \times \Omega$ is the smallest σ-algebra making all processes in $\mathbb{L}$ measurable where $\mathbb{L}$ is the set of processes that have paths that are left continuous with right limits. We let $\mathbf{b}\mathcal{P}$ denote bounded processes that are $\mathcal{P}$-measurable.

Definition 14. Let $X \in \mathcal{H}^2$ with $X = \bar{N} + \bar{A}$ its canonical decomposition, and let $H, J \in \mathbf{b}\mathcal{P}$. We define $d_X(H, J)$ by

$$d_X(H,J) \equiv \left\| \left(\int_0^T (H_u - J_u)^2 \, \mathrm{d}[\bar{N}, \bar{N}]_u \right)^{1/2} \right\|_{L^2} + \left\| \int_0^T |H_u - J_u| |\mathrm{d}\bar{A}_u| \right\|_{L^2}.$$

From here on, we suppose $s \in \mathcal{H}^2$ with the canonical decomposition $s = \bar{N} + \bar{A}$.

Theorem 12. *Let $\epsilon > 0$. For any H bounded, continuous and of FV, there exists H^ϵ, bounded, continuous and of FV, with $H_T^\epsilon = 0$ such that $d_s(H, H^\epsilon) < \epsilon$.*

Proof. Define

$$H_t^m = H_t 1_{[0,T_m]} + H_{T_m} \frac{T - t}{T - T_m} 1_{(T_m, T]},$$

where $T_m = T - \frac{1}{m}$. We will first show $d_s(H, H1_{[0,T_m]}) \to 0$ as $m \to \infty$.

To show $\|(\int_0^T (H_u(\omega) - H_u(\omega)1_{[0,T_m]})^2 \, \mathrm{d}[\bar{N}, \bar{N}]_u(\omega))^{1/2}\|_{L^2} \to 0$, first observe that $[\bar{N}, \bar{N}] = \langle \bar{N}, \bar{N} \rangle + M$, where $\langle \bar{N}, \bar{N} \rangle$ is the compensator, hence predictable, of $[\bar{N}, \bar{N}]$ and M is a local martingale. Since M is a local martingale, there exists a sequence $(T_n)_{n \geqslant 1}$ of stopping times increasing to ∞ such that M^{T_n} is a martingale for each n. Thus, given a bounded G, $G \cdot M^{T_n}$ is a martingale implying $\mathbb{E}[(G \cdot M^{T_n})_t] = 0$ for all t. Moreover,

$$\begin{aligned}\left|\left(G \cdot M^{T_n}\right)_t\right| &\leqslant |G| \cdot [\bar{N}, \bar{N}]_t^{T_n} + |G| \cdot \langle \bar{N}, \bar{N} \rangle_t^{T_n} \\ &\leqslant |G| \cdot [\bar{N}, \bar{N}]_t + |G| \cdot \langle \bar{N}, \bar{N} \rangle_t \end{aligned} \tag{A.11}$$

where the first equality is the triangle inequality and the second follows from $[\bar{N}, \bar{N}]$ and $\langle \bar{N}, \bar{N} \rangle$ being increasing. Furthermore, $G1_{[0,T_n]}$ converges to G hence by Dominated Convergence Theorem for stochastic integrals, $G \cdot M^{T_n}$ converges to $G \cdot M$ in ucp. Moreover, by (A.11), since G is bounded and $[\bar{N}, \bar{N}]$ and $\langle \bar{N}, \bar{N} \rangle$ are integrable, $\mathbb{E}[(G \cdot M^{T_n})_t]$ converges to $\mathbb{E}[(G \cdot M)_t]$ by ordinary Dominated Convergence Theorem. Therefore, $\mathbb{E}[(G \cdot M)_t] = 0$ for all t. Hence, we have

$$\mathbb{E}\big[G \cdot [\bar{N}, \bar{N}]_t\big] = \mathbb{E}\big[G \cdot \langle \bar{N}, \bar{N} \rangle_t\big].$$

Jump times of $[\bar{N}, \bar{N}]$ are those of $\bar{N}$, which are totally inaccessible as a corollary to the previous lemma. Therefore, by the same lemma, $\langle \bar{N}, \bar{N} \rangle$ is continuous. Now,

$$\begin{aligned}&\int_0^T \big(H_u(\omega) - H_u(\omega)1_{[0,T_m]}\big)^2 \, \mathrm{d}\langle \bar{N}, \bar{N} \rangle_u(\omega) \\ &\quad \leqslant \int_0^T \big(H_u(\omega)\big)^2 \, \mathrm{d}\langle \bar{N}, \bar{N} \rangle_u(\omega) < \infty,\end{aligned}$$

for all m, for almost all ω. Thus, by Lebesgue's Dominated Convergence Theorem

$$\int_0^T \big(H_u(\omega) - H_u(\omega)1_{[0,T_m]}\big)^2 \, \mathrm{d}\langle \bar{N}, \bar{N} \rangle_u(\omega) \to 0, \quad \text{a.s.}$$

since $\langle \bar{N}, \bar{N} \rangle$ is continuous. Moreover,

$$\left\| \left((H - H1_{[0,T_m]})^2 \cdot \langle \bar{N}, \bar{N} \rangle\right)^{1/2} \right\|_{L^2} \leqslant \left\| \left(H^2 \cdot \langle \bar{N}, \bar{N} \rangle\right)^{1/2} \right\|_{L^2} < \infty$$

since $H \cdot s \in \mathcal{H}^2$. A second application of Dominated Convergence Theorem yields

$$\left\| \left(\int_0^T \big(H_u(\omega) - H_u(\omega)1_{[0,T_m]}\big)^2 \, \mathrm{d}\langle \bar{N}, \bar{N} \rangle_u(\omega) \right)^{1/2} \right\|_{L^2} \to 0.$$

Since, for any bounded $|G|$, $\mathbb{E}[G \cdot [\bar{N}, \bar{N}]_t] = \mathbb{E}[G \cdot \langle \bar{N}, \bar{N} \rangle_t]$, for all t,

$$\left\| \left(\int_0^T \big(H_u(\omega) - H_u(\omega)1_{[0,T_m]}\big)^2 \, \mathrm{d}[\bar{N}, \bar{N}]_u(\omega) \right)^{1/2} \right\|_{L^2} \to 0,$$

too. By the previous lemma, $\bar{A}$ is continuous as well, so $\| \int_0^T |H_u - H_u 1_{[0,T_m]}||\mathrm{d}\bar{A}_u|\|_{L^2} \to 0$ by a similar argument. Hence, $d_s(H, H1_{[0,T_m]}) \to 0$ as $m \to \infty$.

It remains to show $d_s(H_{T_m}\frac{T-t}{T-T_m}1_{(T_m,T]}, 0) \to 0$, as $m \to \infty$. First note that

$$\int_0^T H_{T_m}^2(\omega)\left(\frac{T-u}{T-T_m}\right)^2 1_{(T_m,T]}\,\mathrm{d}\langle\bar{N},\bar{N}\rangle_u(\omega) \leqslant \int_0^T K\,\mathrm{d}\langle\bar{N},\bar{N}\rangle_u(\omega) < \infty,$$

where $K = \| \max_{0\leqslant t\leqslant T} H_t^2(\omega)\|_\infty < \infty$ since H is bounded. Thus, by the Dominated Convergence Theorem,

$$\int_0^T H_{T_m}^2(\omega)\left(\frac{T-u}{T-T_m}\right)^2 1_{(T_m,T]}\,\mathrm{d}\langle\bar{N},\bar{N}\rangle_u(\omega) \to 0, \quad \text{a.s.}$$

Moreover, another application of the Dominated Convergence Theorem yields

$$\lim_{m\to\infty}\mathbb{E}\left[\int_0^T H_{T_m}^2\left(\frac{T-u}{T-T_m}\right)^2 1_{(T_m,T]}\,\mathrm{d}\langle\bar{N},\bar{N}\rangle_u\right] = 0.$$

A similar argument shows

$$\left\|\int_0^T \left|H_{T_m}\left(\frac{T-u}{T-T_m}\right)\right| 1_{(T_m,T]}|\mathrm{d}A_u|\right\|_{L^2} \to 0$$

which completes the proof. □

Corollary 3. *Let $\epsilon > 0$. For any H, bounded, continuous and of FV, there exists H^ϵ, bounded, continuous and of FV, with $H_T^\epsilon = 0$ such that $\|H\cdot s - H^\epsilon \cdot s\|_{L^2} < \epsilon$.*

Proof. This follows from a combination of Theorem 12 and Theorem 5 of Chapter IV in (Protter, 2005). □

References

Bank, P., Baum, D. (2004). Hedging and portfolio optimization in illiquid financial markets with a large trader. *Mathematical Finance* 14, 1–18.

Barles, G., Soner, H. (1998). Option pricing with transaction costs and a nonlinear Black–Scholes equation. *Finance and Stochastics* 2, 369–397.
Blais, M. (2006). Liquidity and data. *Ph.D. thesis*, Cornell University.
Blais, M., Protter, P. (2006). An analysis of the supply curve for liquidity risk through book data, in preparation.
Çetin, U. (2003). Default and liquidity risk modeling. *Ph.D. thesis*, Cornell University.
Çetin, U., Jarrow, R., Protter, P. (2004). Liquidity risk and arbitrage pricing theory. *Finance and Stochastics* 8, 311–341.
Çetin, U., Jarrow, R., Protter, P., Warachka, M. (2006). Pricing options in an extended Black Scholes economy with illiquidity: Theory and empirical evidence. *Review of Financial Studies* 19 (2), 493–529.
Cheridito, P. (2003). Arbitrage in fractional Brownian motion models. *Finance and Stochastics* 7, 417–554.
Constantinides, G., Zariphopoulou, T. (1999). Bounds on prices of contingent claims in an intertemporal economy with proportional transaction costs and general preferences. *Finance and Stochastics* 3, 345–369.
Cvitanic, J., Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: A martingale approach. *Mathematical Finance* 6, 133–165.
Cvitanic, J., Ma, J. (1996). Hedging options for a large investor and forward–backward SDEs. *Annals of Applied Probability* 6, 370–398.
Cvitanic, J., Pham, H., Touze, N. (1999). A closed-form solution to the problem of super-replication under transaction costs. *Finance and Stochastics* 3, 35–54.
Delbaen, F. (2003). Personal communication.
Delbaen, F., Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* 300, 463–520.
Duffie, D. (1996). *Dynamic Asset Pricing Theory*, second ed. Princeton University Press, New Jersey.
Ferstenberg, R. (2004). Private communication.
Glosten, L., Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14 (March), 71–100.
Grossman, S., Miller, M. (1988). Liquidity and market structure. *Journal of Finance* 43 (3), 617–637.
Harrison, J.M., Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11, 215–260.
Jarrow, R. (1992). Market manipulation, bubbles, corners and short squeezes. *Journal of Financial and Quantitative Analysis,* September, 311–336.
Jouini, E. (2000). Price functionals with bid–ask spreads: An axiomatic approach. *Journal of Mathematical Economics* 34 (4), 547–558.
Jouini, E., Kallal, H. (1995). Martingales and arbitrage in securities markets with transaction costs. *Journal of Economic Theory* 66 (1), 178–197.
Jouini, E., Kallal, H., Napp, C. (2001). Arbitrage in financial markets with fixed costs. *Journal of Mathematical Economics* 35 (2), 197–221.
Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica* 53, 1315–1335.
Protter, P. (2005). *Stochastic Integration and Differential Equations, second ed., Version 2.1*. Springer-Verlag, Heidelberg.
Protter, P. (2001). A partial introduction to financial asset pricing theory. *Stochastic Processes and Their Applications* 91, 169–203.
Soner, H.M., Shreve, S., Cvitanic, J. (1995). There is no nontrivial hedging portfolio for option pricing with transaction costs. *Annals of Applied Probability* 5, 327–355.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15018-0

Chapter 18

Financial Engineering: Applications in Insurance

Phelim Boyle

School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5
E-mail: pboyle@wlu.ca

Mary Hardy

Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada N2L 3G1

Abstract

This chapter provides an introduction to the insurance area and discusses financial engineering applications in insurance. There are some key differences between the nature of the risks found in insurance products and those typically found in banking applications. Historically insurance risks were handled using the actuarial approach and we describe this approach and contrast it with the financial engineering approach. We focus on life insurance applications. Life insurance products often include a variety of investment options which are related to the performance of a stock portfolio or an index portfolio. The pricing and risk management of these contracts has provided a challenge to traditional actuarial techniques. In this chapter we describe how the modern approach to risk management combines concepts from modern financial economics and econometrics with ideas from the traditional actuarial approach. We discuss two specific applications: variable annuity contracts and guaranteed annuity options.

1 Introduction

In the early development of life insurance companies and pension plans it became clear that a scientific approach was required to ensure that these institutions would be able to meet their liabilities. This role was filled initially by mathematicians who performed calculations to ensure that these enterprises charged adequate premiums to maintain their solvency. Over time this role developed into a new profession of insurance mathematicians or actuaries. The first official use of the word actuary was in 1762 to describe Edward Rowe

Mores, the chief officer of a newly formed life insurance company which subsequently became known as[1] Equitable Life. Actuaries were entrusted with maintaining the solvency of insurance and pension enterprises and they have a strong claim to be regarded as the first financial engineers. In many jurisdictions the actuarial profession has a statutory role in certifying the solvency of life insurance companies and pension plans.

The initial focus of the actuarial profession was in safeguarding the solvency of insurance companies and pension plans. Over time the profession developed a more active role in financial engineering and actuaries became key players in the development of risk management techniques. Notable names include Macauley and Redington. However, apart from a few exceptions actuaries have not been heavily involved in the financial engineering revolution of the last thirty years. The application of modern financial engineering to insurance risks has largely evolved in the past ten years. The contingent claims in modern insurance are very similar in some ways to standard capital markets contingent claims. However, there are also some very significant differences.

The insurance market lacks many of the attributes we associate with a perfect frictionless market. Insurance and pension contracts have much longer terms than most other financial contracts. One cannot assume the same degree of rationality with regards to the exercise of the various embedded options in an insurance policy, as in the case of traded options. The market for insurance products is incomplete. Phenomena like adverse selection and moral hazard are much more prevalent in insurance contracts than in most financial contracts. Standard financial engineering methods were developed mostly for shorter term, more marketable contracts. These methods require substantial adaptation when dealing with insurance contracts.

Contemporary life insurance products are more likely to combine traditional death benefit coverage with a product that has more of an investment focus. These contracts developed as the market for traditional products slowed, and insurers recognized that expansion was achievable by offering products that could compete with retail products offered by other financial institutions such as banks and the mutual funds market. In designing these products, it was natural for insurers to introduce guarantees. In part financial guarantees were incorporated because insurance has always guaranteed benefits. In addition, in some jurisdictions, there were tax benefits for insurance contracts which were not available for pure investment contracts. By offering guarantees that were not a conventional part of the mutual fund type contract, the insurer could ensure that the contract qualified for the favorable tax treatment.

In the remainder of this chapter we explain how actuaries have adapted the techniques of capital markets, and combined them with traditional actuarial techniques, to develop a synergetic approach to insurance risks. In the next

[1] See Ogborn (1956). This insurance company – the first in the world – was originally known as The Society for Equitable Assurances on Lives and Survivorships.

section we discuss the evolution of insurance contracts from policies which provided fixed benefits and life insurance protection to contemporary products which often contain direct participation in market performance coupled with financial guarantees. Section 3 summarizes early actuarial approaches to the problem of charging an appropriate premium to cover the risk. These so called premium principles correspond to risk measures. In Section 4 we examine modern approaches to risk management by insurance companies and contrast the actuarial approach of using capital as a cushion to the financial engineering approach of dynamic hedging. Section 5 gives details of an important class of contracts known as variable annuities. We discuss the risk management of one form of these contracts and explain both the actuarial approach and the dynamic hedging approach in this case. Section 6 uses guaranteed annuity options to illustrate the complexity of some of the long term embedded options found in insurance contracts and the risk management challenges they present.

2 Insurance products and markets

The major difference between the traditional type of insurance contract and more modern products is that the liability in the traditional contract is largely diversifiable, while many of the more recent products contain a significant investment component which represents a nondiversifiable risk. If a large number of independent lives buy traditional term insurance, the claims experience is highly predictable. From the strong law of large numbers the experience will tend towards the mean and from the central limit theorem the distribution of claims will tend to the Normal. The more contracts that are sold, the smaller the relative risk. The easiest way to hedge the mortality risk is to sell more contracts.[2]

Traditional life insurance products provided a fixed level of financial protection in the event of the death of the insured life or survival to a given age. A term insurance product provides life insurance coverage for a given period while a whole life contract provides life insurance coverage for the duration of the insured's life. An endowment policy pays the sum assured either if the life insured dies within the term of the policy or if the life insured survives to the end of the contract period. In the beginning, the sum assured was fixed and was funded by level periodic premiums or in some cases by a single lump sum premium.

Let us start with a very simple example. Suppose an insurance company sells a single[3] term life insurance contract. For a premium of \$1050, the insurer will pay a sum assured of \$100,000. This sum assured is payable if and only if the

[2] Another method is to sell longevity risk and mortality risk (life insurance) to the same lives.

[3] Of course insurance companies do not operate in this way. They sell several policies to diversify the risk.

life insured dies within the term of the contract: in this case one year. Suppose, further, that the probability of the life in question dying within one year is 0.01. Then either the life survives or dies within the year. In the first case the insurer pockets the premium and it also survives. In the second case the insurer needs to pay the claim of $100,000. If the insurer does not have access to this amount of capital then it too fails. If this is the case the insurer has a 1% probability of failure.

Now suppose the insurer sells one million identical contracts to independent lives, each with a claim probability of 0.01 and the same sum assured as before. The insurer collects $1050 million in total premiums. This amount will be sufficient to pay all the claims if there are no more than 10,500 deaths. The probability of this happening is less than 10^{-6}. The risk of insurer insolvency has been reduced to a negligible amount by writing more business. The central limit theorem also states that the experience will become very close to the expected loss as more policies are written. This is why the expected loss is one of the key valuation measures used by actuaries.

An important development in the evolution of insurance products was the introduction of participating or with-profits insurance. It was realized that to ensure ongoing solvency of the insurer it was necessary to charge premiums that were greater than the expected value of the benefits. Often this was accomplished by using conservative assumptions in determining the premiums. This device provided a cushion for the insurance company against unfavorable experience and it lead to the accumulation of a profit or surplus. Under these contracts the policyholder shared in the favorable experience of the insurance company, often through a reduction in subsequent premiums.

In markets dominated by UK insurers the profit participation was effected through bonus additions to the sum assured. The periodic increases to the sum assured were called *reversionary bonuses*, and once declared became a guaranteed benefit, provided the policyholder did not surrender the contract early. The introduction of the bonus system allowed a freer investment policy, with policyholders sharing in the resulting profits. As UK insurers increasingly used this freedom to invest in equity investments, they realized that the regular declaration of reversionary bonuses was still too restrictive. In the 1960s, many of them began postponing a large part of the bonus until the policy reached the end of the term, either by death or by reaching the maturity date. This *terminal bonus* was not guaranteed. The terminal bonus sometimes constituted as much as 50% of the final benefit payment. In this way, the traditional life insurance contract took on more of an investment focus, particularly for the endowment insurance contract. In principle these UK with profits policies are designed to smooth out the actual fluctuations in investment returns. In practice the with profits contract is often criticized[4] for being opaque and paternalistic.

[4] In 2004 Callum McCarthy a senior UK regulator commented *More significantly, it requires a lifting of the veils which have traditionally obscured what was actually on offer in a with-profits policy: in future, we*

In the last few decades we have witnessed the development in many insurance markets of new types of retail insurance products where the investment performance is more explicitly linked to the performance of the underlying assets. One of the main factors behind the growth of this business has been the increasing competition from banks and other financial institutions for the savings of retail investors. The increasing globalization of financial services was another factor in this development. In these investment linked contracts the benefits under the insurance policy are tied directly to the performance of some benchmark portfolio or index such as for example the S&P 500. Such contracts are known by different names in different jurisdictions. Examples of such contracts include the variable annuity and equity indexed annuity products in the USA, segregated funds in Canada, unit linked contracts in the UK and other parts of Europe and structured products in parts of Asia. Since all these products carry benefits linked directly to the performance of a portfolio of equities or another equity index (or a mix of equities and other investments) the generic term for such contracts is *equity-linked life insurance*.

We now give an example of an equity-linked policy. In this case the insured pays a single premium for a seven year contract. The maturity benefit under the contract in seven years is the maximum of

- The single premium.
- The single premium rolled up at a rate equal to sixty percent of the return on the S&P index over the seven year period.

To preserve the insurance status of the arrangement there is life cover for the duration of the contract. The policyholder also has the option of lapsing or surrendering the contract and obtaining a cash surrender value. This contract can be viewed as a package of embedded options:

- They are often very long term.
- They are not traded separately in a liquid secondary market but are bundled together in the contract.
- We cannot assume they will be exercised in a rational fashion just as mortgage prepayment option exercise behavior is not fully rational.

An equity-linked insurance policy contains both investment risk and mortality risk, and it is only the mortality risk which can be diversified by pooling a large number of similar contracts. The investment risk is not reduced by pooling. For example, the contract we just described can be viewed as a package consisting of a seven year zero coupon bond, a type of call option on the S&P

expect information to be made available – comprehensibly – on how a firm manages its with-profits funds, its approach to payouts (on maturity and on surrender), its smoothing policy, its investment strategy and any changes to asset share. This is the information needed if a customer is to make an informed choice between competing companies – and, for that matter, between competing investment opportunities (McCarthy, 2004).

and a seven year term insurance benefit. The mortality risk is still diversifiable, but the size of the payout is now also random, depending on the index performance. This type of investment risk is not diversifiable. For this type of contract, insurers are increasingly turning to financial engineering solutions.

These options are more challenging to value than traded financial options for several reasons. The major differences between the embedded option in insurance and the standard options of financial markets are:

- Insurance contracts tend to be very long term. Some contracts have terms of over thirty years.
- The embedded options in insurance contracts are life contingent; the variable annuity guarantee described in the previous paragraph matures on the death of the policyholder. The term is therefore random. In general, the term of these options depends on the survival status of a policyholder.
- The factors that influence the exercise of these options are more complicated than in the case of traded financial options. We cannot assume that they will be exercised in a perfectly rational manner. The exercise of individual consumers is much harder to predict as we have seen in the case of the exercise of mortgage prepayment options.
- Many of the guarantees offered are deeply out-of-the-money at issue.
- When pricing financial options it is often assumed that there is no arbitrage and that the market is complete. In the case of the long term embedded options in insurance contracts, both these assumptions are less likely to hold.

These differences mean that the adaptation and implementation of standard financial engineering techniques will not be straightforward.

3 Premium principles and risk measures

One of the landmark contributions to actuarial science is the work of F. Lundberg, who in 1909 developed the so-called collective risk theory (Lundberg, 1909). Lundberg developed a mathematical model which showed under certain assumptions how the amount of additional loading on the premium was related to the probability of insurer insolvency or ruin. The use of the probability of ruin in this way foreshadowed the concept of Value-at-Risk[5] or VaR. Although from a modern perspective Lundberg's model lacked economic realism, it was a landmark contribution. It provided a precise scientific connection between the premium charged and the probability of the company remaining solvent.

[5] In financial markets Value-at-Risk, or VaR, has become a ubiquitous risk measure. It is now widely used by financial institutions, corporations and regulators as a measure of risk.

Lundberg's work provided an early example of how to construct a risk measure. A risk measure maps the claims distribution to the real numbers, and is used to quantify riskiness according to some criteria. The expected value[6] is an example of a risk measure. The use of risk measures to assess capital requirements has a much longer history in insurance, where risk measures are known as premium principles (Bühlmann, 1970; Gerber, 1979). Premium principles are measures applied to insurance loss distributions, and therefore differ slightly from banking risk measures where profit/loss distributions must be accommodated. For a loss random variable, $X > 0$, the standard premium principles described by Gerber (1979) are defined for some positive parameter $\alpha \geqslant 0$ as follows:

- The *expected value principle* is $(1 + \alpha)\mathrm{E}[X]$.
- The *standard deviation principle* is $\mathrm{E}[X] + \alpha\sqrt{\mathrm{Var}[X]}$.
- The *variance principle* is $\mathrm{E}[X] + \alpha\,\mathrm{Var}[X]$.
- The *zero utility principle*; for some utility function $u(x)$ and surplus w, the premium principle is P such that

  $$u(w) = \mathrm{E}\big[u(w + P - X)\big].$$

  A popular choice for $u()$ is exponential utility, in which case the initial surplus does not affect the calculations, giving the *exponential principle*:

  $$\frac{1}{\alpha}\log\big(\mathrm{E}\big[e^{\alpha X}\big]\big).$$

- The *quantile principle*; let $F_X(x)$ denote the distribution function of X, and α is a parameter such that $0 \leqslant \alpha \leqslant 1$. The quantile principle is $F_X^{-1}(\alpha)$.

Since Buhlmann and Gerber categorized the known premium principles, insurance scholars have been developing new variations. Wang (1995) developed a new approach to risk measures (premium principles) for loss distributions, using distortion of the survival function.

For a nonnegative loss random variable X, with survival function $S(x) = Pr[X > x]$, the mean loss is

$$\int_0^\infty S(x)\,\mathrm{d}x.$$

Wang's contribution was to suggest a risk measure based on a distortion function $g(S(x))$. The distortion function is an increasing function, with $g(0) = 0$ and $g(1) = 1$. The distortion risk measure is $H(X)$, say, where

$$H(X) = \int_0^\infty g\big(S(x)\big)\,\mathrm{d}x.$$

[6] Plus a loading factor.

Wang (1995) suggested the proportional hazard risk measure, where $g(u) = u^{\frac{1}{\rho}}$, for some $\rho \geqslant 1$. The ρ parameter determines the risk loading factor. Another suggestion from Wang (2002) is the normal–normal transform, where $g(u) = \Phi(\Phi^{-1}(u) + k)$ for some parameter $k > 0$. This can be shown to lead to the Black–Scholes–Merton option pricing formula in some cases, with a suitable choice for k. The risk measure might be applied to individual risks or to portfolios. It is interesting to note that the quantile principle corresponds to the widely used Value-at-Risk measure.

While these premium principles can be useful tools they do not capture some important dimensions of the insurance market place. Usually the seller cannot just decide on the price according to some formula and ignore the demand side. This criticism is not new and it dates back to Karl Borch. More economic based approaches have been advocated by Borch and Bühlmann and others.

These measures can be used for either pricing or economic capital calculations, or both. Typically, in life insurance the expected value principle would be used for both, with a larger α value for the economic capital. The risk management would then be fairly passive, with any excess liability over the accumulated premium absorbed by the additional economic capital held. In property and casualty insurance, other premium principles are sometimes used. Risk management in life insurance was originally relatively passive: the insurer set aside enough capital to ensure that the liabilities could be met with a certain probability. Actuaries of course recognized that the premiums should be invested in securities that were appropriate given the nature of the liabilities.

Redington (1952) developed this idea more fully through the concept of immunization, a precursor of dynamic hedging. Redington demonstrated that by selecting assets to equal the liabilities in duration and exceed the liabilities in convexity, it was possible to hedge against small movements in the interest rate. Immunization became an important tool in actuarial risk management, and is still utilized in asset-liability management strategies today. From the start, it was noted that when the interest rate shift occurred, some rebalancing of the asset portfolio would be required. It was also recognized that perfect immunization would not be possible, due to various uncertainties, and because the theory ignores the term structure of interest rates. Nevertheless, immunization was recognized as an important risk management tool for long term insurance liabilities.

4 Risk management for life insurance

Insurance companies use different methods to manage the financial risk associated with embedded options and we can distinguish two main approaches. These are:

- The actuarial reserving method whereby the financial institution sets aside additional capital to ensure that the liabilities under the guaran-

tee will be covered with a high probability. In this case the projection is carried out under the real-world measure often called the P-measure.

- The second approach is the financial engineering approach used by investment banks. The insurer sets up a replicating portfolio of traded securities and dynamically hedges this portfolio over time so that at maturity it matches the liability. In this case the investment proportions in the replicating portfolio are computed using the equivalent martingale measure, or Q-measure.

In practice a combination of these approaches can be used. We describe each method in turn.

4.1 The actuarial approach

Actuaries first began to recognize the challenges of nondiversifiable risk management in the late 1960s and early 1970s when equity-linked contracts first became popular. The Black–Scholes–Merton approach to risk management first became available in 1973 and it was initially viewed with skepticism[7] by many actuaries.

Instead the insurance industry adopted a semi-passive approach. A real-world model was used to project the liability distribution, using Monte Carlo simulation. The liabilities were discounted using a 'conservative' discount rate (which would approximate the risk free rate). Then a risk measure would be applied to the simulated liability present value distribution to determine a capital requirement. The process is described in the report of the Maturity Guarantees Working Party of the Faculty and Institute of Actuaries (MGWP, 1980). The risk measure applied in this report and in much of the subsequent work was the quantile measure, so the capital requirement would be set at the 99% or even the 99.9% quantile of the loss distribution. The approach is essentially passive in principle, but in practice, the requirement to recalculate the capital requirement each year enforced a more dynamic approach.

The liability modeling for embedded options required more sophisticated models than those that were being used by actuaries. Consequently, an important part of the development of risk management for financial guarantees has been concerned with constructing sophisticated integrated models of assets and liabilities that can be used to project the future distribution of the liabilities. These are all real-world models. In the late 1970s, an early version of the Wilkie (1986, 1995) model was first developed for projecting the liabilities for the Maturity Guarantees Working party in the UK. The Wilkie model is an integrated model of inflation, equity prices and dividends and bond prices

[7] This viewpoint is summarized in the following quote from the Maturity Guarantees Working Party (1980). *The Working Party spent time studying the subject and reached varying degrees of confidence that the mathematics was sound. In some cases the confidence was derived from the fact that nobody seems to have seriously challenged the underlying theory.*

that has proven to be a popular basis for asset and liability projections. Other real-world models are discussed in Section 4.5.

We can summarize the traditional P-measure approach as follows. Suppose L_t is the amount payable under an insurance contract at time t. For an embedded put option, the liability might consist of the excess, if any, of the guaranteed amount G_t over the value of the reference fund F_t, say. The instantaneous mortality[8] rate at t is denoted by $\mu_x(t)$ for a life aged x at inception. The survival probability for the life from age x to age $x + t$ is denoted by ${}_tp_x$. We assume the contract lasts for n years and that the risk free rate is r. Then for a guarantee payable on the earlier event of death or expiration the discounted expected present value of the liability at inception ($t = 0$) is

$$A_0 = \int_0^n {}_tp_x \mu_x(t) L_t e^{-rt}\, \mathrm{d}t + {}_np_x L_n e^{-rn}.$$

The price and the initial capital requirement would then be determined from the distribution of A_0, using appropriate risk measures.

In practice, the A_0 distribution is estimated by simulation, and the risk measure applied to the simulated distribution.

4.2 The dynamic hedging approach

The approach described in the previous section does not involve any attempt to use a dynamic hedging strategy to mitigate the risk, even though many of the embedded options in life insurance contracts are relatively straightforward. Originally, the reason was lack of awareness, or lack of credibility in what was a fairly new and very radical approach.

More recently, actuaries have adopted dynamic hedging techniques, but with adaptations. There are challenges with a naïve application of the Black–Scholes methodology arising largely from the issues listed at the end of Section 2 – these are the very long term nature of the options, the dependence on mortality and the fact that often the options are deeply out-of-the-money at issue. Consequently, insurers may now combine financial engineering techniques from the banking world with the models and techniques of the insurance world. We call this the hybrid approach.

Both the long term nature of the liabilities and the moneyness issues mean

- The standard models of Black–Scholes may not be appropriate for insurance guarantees.
- Econometric modeling is critical to successful risk management.
- The practical issues of discrete hedging and transactions costs may have a significant effect on the liability.
- The mortality factor means that the term of the guarantee is random, being dependent on the survival of the policyholder.

[8] This is known as the hazard rate in other applications.

4.3 Mortality dependent options

The mortality issue was first addressed by Brennan and Schwartz (1976) and Boyle and Schwartz (1977). If the value at time 0 of an option which matures at time n with certainty is $H(n)$, then the value of an option which matures at time n dependent on the survival to n of a life who is age x at time $t = 0$, and who has future lifetime random variable T_x at age x, is simply $H(n)Pr[T_x > n]$. That is, the risk neutral measure for the mortality risk is simply the real world measure provided it is fully diversifiable, and independent of the guarantee liability. Similarly, if the guarantee is payable at n conditional on the life dying in the interval $(n - \epsilon, n)$, then the value of the option is

$$H(n)Pr[n - \epsilon < T_x \leqslant n].$$

And in general, the value of a guarantee payoff $H(T_x, n)$, where the term is n-years, and which is dependent on the future lifetime random variable T_x, is

$$E_{T_x}\big[E_Q\big[e^{-rn}H(T_x, n) \mid T_x\big]\big],$$

where the T_x expectation uses the real-world mortality measure, and the Q expectation uses the risk neutral financial measure.

The reason why we can use the P-measure for the future lifetime is explained more fully in Boyle and Schwartz (1977), and also in Lin and Tan (2003). The intuition is that for a fully diversifiable risk one can use the P-measure for pricing.

4.4 The hybrid approach – combining P and Q measure

One approach to the problems of discrete hedging, and the need for more realistic econometric models for longer term options is to model the costs of a hedge strategy under a realistic P-measure. That is, use the Q-measure to determine a hedge strategy, and then use the realistic P-measure to project the hedge, and estimate the unhedged liability, arising from discrete hedging error, transactions costs, and model error.

If we assume that the hedge portfolio follows a pre-determined dynamic strategy, and is re-arranged at unit time intervals, then the Monte Carlo simulation process requires the following steps, at each time unit:

1. Simulate the updated values for the underlying risky assets using the real world measure.
2. Calculate the updated value of the hedge portfolio brought forward from the previous time step.
3. Calculate the revised value of the required hedge based on the updated information.
4. The difference between the value of the hedge carried forward and the value of the hedge brought forward is the hedging or tracking error. Discount all the hedging errors as part of the unhedged liability.

5. Calculate the transactions costs associated with rebalancing the hedge portfolio. This too is discounted, and forms the second part of the unhedged liability.

At the final time step, the hedge required is the simulated option liability.

Combining the present value of the hedging error and the present value of the transaction costs gives the unhedged liability (there may be other items of unhedged cashflow, this is the simplest case).

The insurer can add additional capital to allow for the hedging errors and the transaction costs. We would expect the average hedging error to be close to zero. However, if we apply a risk measure that gives weighting to the more risky outcomes, then the possibility of hedging error leads to a capital requirement. For example, if we select as the capital requirement the 99% quantile of the hedging error distribution, then we would run this fund alongside the hedge portfolio. When the hedging error is negative, the surplus is paid into the fund. When the hedging error is positive, meaning that additional cash is required to make up the hedge, then the cash can be taken from the fund. There is (broadly) only a 99% chance that the whole fund will be used up in meeting the cost of hedging errors. If it is not all needed, the excess would be released back to the company in due course. The total capital requirement at inception would be the cost of the hedge portfolio, plus the capital requirement to cover unhedged costs. This hybrid approach to risk management is permitted for Canadian insurers writing equity-linked contracts with guarantees.

4.5 Realistic models for price projection

In order to apply either the actuarial or the hybrid approaches, a realistic distribution of the reference portfolio is required. For example using a standard lognormal model for the price process involved in an equity-linked contract generally underestimates the risk for an out-of-the-money option. This is due to the fact that the lognormal distribution is too thin-tailed to fit the empirical distribution, and the fact that for some products, stochastic volatility is a significant source of potential liability. The problems arise because the risks are so very long term, and because they tend to start deeply out-of-the-money, so that the tails of the distribution are particularly important. Consequently the identification of models which adequately capture the fat tails and the uncertain volatility of equity prices is now an important component of risk management.

Models which are popular with actuaries include the Wilkie model, mentioned above (Wilkie, 1986, 1995), and various models derived from it, such as in Whitten and Thomas (1999). For the common form of equity-linked life insurance a complex integrated model is not required – the critical risk is from the stock price process. Hardy (2001) shows that the regime switching lognormal model with two regimes provides a good fit to S&P 500 monthly returns over the last forty years. Hardy's model outperforms other competing candidates, including GARCH. Under the regime switching model (which is based

on the framework proposed by Hamilton, 1989), the price process jumps randomly between two regimes. Within each regime the process is lognormal, but the two regimes have different parameters; the more common regime has low volatility and a high mean; the less common regime has high volatility and low mean (capturing the association of high volatility with market crashes). The switching process is a hidden Markov process, with a relatively low chance of moving from the low volatility to the high volatility regimes, and a much higher probability of moving back once the process has switched.

4.6 Risk measures

Once the liability distribution is simulated, we need a risk measure to determine an appropriate price and capital requirement. For capital requirements, quantile measures such as VaR were common, but more recently have been displaced by the Conditional Tail Expectation or CTE risk measure. This measure is also known as the expected shortfall or tail VaR.

The CTE risk measure is defined as the average loss given that the loss falls in the worst $(1 - \alpha)$ part of the distribution. Suppose that $Q_\alpha(X)$ is the α-quantile of the loss distribution, of $X > 0$, and that, further, the quantile does not fall in a probability mass, so that for all $\gamma > \alpha$,

$$Q_\gamma(X) > Q_\alpha(X) \tag{1}$$

then the CTE is defined as

$$CTE_\alpha(X) = E\big[X \mid X > Q_\alpha(X)\big]. \tag{2}$$

Where the constraint in (1) is not met, the fuller definition of the CTE uses $\beta' \geqslant \alpha$ where

$$\beta' = \max\big\{\beta\colon\ Q_\beta(X) = Q_\alpha(X)\big\}$$

then

$$CTE_\alpha(X) = \frac{(1 - \beta')E[X \mid X > Q_\alpha(X)] + (\beta' - \alpha)Q_\alpha(X)}{(1 - \alpha)} \tag{3}$$

which just uses all the distribution above the quantile, plus weighting the quantile enough so that the expectation involves exactly $(1 - \alpha)$ of the distribution.

Using stochastic simulation the estimation of the CTE is achieved very simply by taking the average of the worst $100\alpha\%$ of outcomes. The advantages of the CTE measure over quantile measures are discussed in Artzner et al. (1999). The CTE is the basis of the capital requirements for segregated fund contracts in Canada and is proposed for Variable Annuity business in the USA.

5 Variable annuities

Variable annuities are very popular contracts in the United States. They are investment-insurance vehicles designed to increase retirement income. They

permit participation in the equity markets together with investment guarantees. In this section we first describe the main types of variable annuities. Then we discuss a particular type of contract known as the Guaranteed Minimum Maturity Benefit. Then we describe the risk management of the embedded put option in the Guaranteed Minimum Maturity Benefit.

5.1 Main types of variable annuities

Variable annuities comprise a mutual fund type investment, together with insurance and investment guarantees. As the business has become more competitive, the range and complexity of the guarantees offered has increased. We now briefly describe some of the common types of guarantees.

- *Guaranteed Minimum Death Benefit* (GMDB): If the policyholder dies during the policy term, it is guaranteed that the claim payment will be at least equal to the initial investment, possibly with some interest. This guarantee corresponds to an embedded put option with a stochastic exercise time.
- *Guaranteed Minimum Maturity Benefit* (GMMB): In this contract the proceeds at the end of the policy term are guaranteed to be at least equal to some minimum amount, such as the initial investment, or the initial investment plus some interest. In this case the put option has a fixed maturity.
- *Guaranteed Minimum Withdrawal Benefit* (GMWB): Under a GMWB withdrawals of up to some proportion of the original investment may be made from time to time, and the total withdrawals are guaranteed at least to meet the initial investment. There are a couple of ways that this benefit can be decomposed into more basic option contracts.
- *Guaranteed Minimum Income Benefit* (GMIB): Under this contract, when the initial term of the contract has expired, the policyholder may annuitize the proceeds. The GMIB guarantees the minimum annuity payments.
- *Guaranteed Minimum Accumulation Benefit* (GMAB): This is a form of a GMMB (though it may also be applied to the GMDB). When the initial term of the contract has expired, the policyholder may renew the contract on the original terms. If the value of the policyholder's fund is greater than the original guarantee, the new contract continues at the higher guarantee level. If the market value of the policyholder's fund is less than the original guarantee, the insurer must pay the difference into the fund, and the policy is renewed at the original guarantee level.
- *The Reset Option*: This is not a guarantee, but an option to reset the guarantee level at certain times. The term of the contract is generally then extended, effectively issuing a new contract at the reset date. This is a form of lapse and re-entry option, in insurance terms – or a shout option in finance terminology.

The term of a VA contract is usually at least 10 years, often longer. All VA contracts offer some GMDB, but the GMMB is becoming more popular. The GMMB carries significantly more risk than the GMDB, since in general more policyholders will survive to maturity than will die during the policy term, and because deaths are time diversified. In the case of a GMMB the survivors' policies of a given cohort all mature at the same time and the if the market is depressed then the put options for all the contracts will be in the money.

In the remainder of this section we will work through some examples of the hybrid actuarial/financial engineering approach to the risk management of VA guarantees. We assume that all cashflows and hedge rebalancing occurs at monthly intervals, so we use months (assumed all to be of length 1/12 years) as the time unit. It would also be feasible to use move-based discrete hedging, where re-balancing the hedge portfolio is triggered by the amplitude of the move in the underlying stock price process, rather than time-based hedging where re-balancing is assumed to be carried out at regular intervals. For a comparison between move-based hedging and time-based hedging in an insurance context see Boyle and Hardy (1997).

5.2 Guaranteed Minimum Death Benefit (GMMB) example

We use the following notation/assumptions for this example. These are simplified, but should be adequate for illustration purposes.

- We assume a \$100 initial investment. The contract is a single premium contract, which means that any further investment is effectively a new contract carrying separate guarantees. The term of the contract is n years.
- The premium is invested in a fund which has a market value of F_t at time t. The returns on the fund are driven by an equity index, denoted by S_t. The management charge of $100m\%$, per year is deducted from the policyholder's account. The initial index value is $S_0 = 100$.
- We assume the guaranteed benefit payable immediately on survival to the end of the term is the amount of the initial premium without interest.
- The policyholder is age x at inception. The random time to death is denoted by T_x, and this random variable is assumed to have a known distribution.
- The t-year survival probability for T_x is denoted ${}_tp_x$, in the conventional actuarial notation, and the force of mortality after t years is $\mu_x(t)$. The density function for T_x is then ${}_tp_x\mu_x(t)$. We will also use the fact that, for any $t, u > 0$, ${}_{t+u}p_x = {}_tp_x\,{}_up_{x+t}$.
- The risk free rate of interest is r per year (continuously compounded).
- The maximum term of the contract is n years.

The embedded option represented by the GMMB is a put option with term n and strike price $G = 100$.

5.3 P-measure approach

We use a realistic model of equity prices to project the GMMB liability; the force of mortality μ (hazard rate) is assumed to be deterministic. For each simulated path of index prices, S_t, the simulated present value of the liability is

$$PVAL = {}_np_x(G - F_n)^+ e^{-rn} = {}_np_x\big(100 - S_n e^{-m\,n}\big)^+ e^{-rn}.$$

The initial capital requirement for the contract might be the 95% CTE of the simulated distribution of PVAL. The price of the guarantee might be determined using a lower CTE level, although the price tends to be very small (the option being well out of the money at issue). In practice the price is also influenced by competitive considerations.

5.4 Hybrid approach

Under the hybrid approach we select a Q-measure, and determine the initial price of the hedging portfolio. Then the hedge is projected under the P-measure, allowing the actuary to estimate the distribution of the unhedged liability arising from the realistic P-measure, the discrete hedging error and the transaction costs. We assume in this example that the Q-measure selected is the standard Black–Scholes measure, so that the hedge portfolio price at any stage is a simple Black–Scholes put option price, allowing of course for survival to maturity. We let σ denote the index price volatility.

The price at issue is

$$\mathrm{E}_{T_x}\big[E_Q\big[e^{-rn}(G - F_n)^+ \mid T_x > n\big]\big]$$

and

$$E_Q\big[e^{-rn}(G - F_n)^+\big] = E_Q\big[e^{-rn}\big(G - S_n e^{-mn}\big)^+\big].$$

This expectation is the Black–Scholes price for a put option on the index S_t, with the management charge deduction (which is effectively a negative dividend), and with strike price G.

Let $BSP(K, t, T)$ denote the Black–Scholes price at time, t for a put option on a unit of the index S_t maturing at T, with management charge m per year. We have

$$BSP(K, t, T) = Ke^{-r(T-t)}\Phi\big(-d_2(t)\big) - S_t e^{-mT}\Phi\big(-d_1(t)\big),$$

$$\text{where} \quad d_1(t) = \frac{\log(S_t e^{-mT}/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}$$

$$\text{and} \quad d_2(t) = d_1(t) - \sigma\sqrt{(T - t)}.$$

If we ignore mortality, then the initial price of the GMMB, $H(0)$, say, is

$$H(0) = BSP(100, 0, n).$$

The price can be broken down, as usual, into the delta hedge components. The delta hedge portfolio at issue comprises $HS(0)$ in equities, and $HB(0)$ in risk free bonds, where

$$HB(0) = 100e^{-rn}\Phi\big(-d_2(0)\big), \qquad HS(0) = -S_0 e^{-mn}\Phi\big(-d_1(0)\big).$$

We assume the hedge is re-balanced monthly; after one month, if the policyholder survives, the delta hedge required depends on the prevailing index price. The hedge portfolio after one month is

$$H\left(\frac{1}{12}\right) = HB\left(\frac{1}{12}\right) + HS\left(\frac{1}{12}\right),$$

where

$$HB\left(\frac{1}{12}\right) = 100e^{-r(n-1)}\Phi\left(-d_2\left(\frac{1}{12}\right)\right),$$
$$HS\left(\frac{1}{12}\right) = -S_1 e^{-mn}\Phi\left(-d_1\left(\frac{1}{12}\right)\right).$$

The hedge brought forward from the first month has value

$$HF\left(\frac{1}{12}\right) = HBF\left(\frac{1}{12}\right) + HSF\left(\frac{1}{12}\right)$$

say, where $HBF\big(\frac{1}{12}\big)$ is the value at time $t = \frac{1}{12}$ of the bond hedge established at time 0, and $HSF\big(\frac{1}{12}\big)$ is the value at $t = \frac{1}{12}$ of the equity hedge established at time 0. That is

$$HBF\left(\frac{1}{12}\right) = 100e^{-r(n-\frac{1}{12})}\Phi\big(-d_2(0)\big) \quad \text{and}$$
$$HSF\left(\frac{1}{12}\right) = -S_{\frac{1}{12}}\, e^{-mn}\Phi\big(-d_1(0)\big).$$

We can repeat this at monthly intervals. For each month, say $t = 1/12, 2/12, \ldots, n$, the difference between the hedge required, $H(t)$ and the hedge brought forward $HF(t)$ is the hedging error at t and is part of the unhedged liability. So, assuming $T_x > n$, the hedging error at t is $H(t) - HF(t)$.

Now we incorporate mortality. The hedge required at issue is

$$E_{T_x}\big[H(0) \mid T_x > n\big] = {}_n p_x H(0).$$

If the policyholder survives to t, the hedge required is ${}_{n-t}p_{x+t}H(t)$. The probability of survival is ${}_t p_x$, so the expected cost of the hedge at t, taking expectations at issue, is

$${}_{n-t}p_{x+t}\,{}_t p_x H(t) = {}_n p_x H(t).$$

Similarly, the hedge brought forward at t, given the survival of the policyholder to $t - 1/12$ is ${}_{n-(t-\frac{1}{12})}p_{x+(t-\frac{1}{12})}HF(t)$. At t, the hedging error is

${}_{n-t}p_{x+t}H(t) - {}_{n-(t-\frac{1}{12})}p_{x+(t-\frac{1}{12})}HF(t)$ if the policy holder survives to t, and $0 - {}_{n-(t-\frac{1}{12})}p_{x+(t-\frac{1}{12})}HF(t)$ if the policyholder dies in the month from $t - \frac{1}{12}$ to t. Taking expectations at issue, the expected hedging error at t is

$$he_t = {}_np_x\big(H(t) - HF)\big).$$

This result is very convenient; it means that we can calculate the hedge and hedging error ignoring mortality, and then simply multiply everything by the n-year survival probability.

By using a more realistic P-measure to determine the stock price process S_t the hedging error implicitly captures two sources of unhedged liability; the inadequacy of the lognormal distribution for long term modeling of the stock process, and the error resulting from discrete hedging.

As we simulate the hedge process we can also simulate the transaction costs, which may be substantial for very long contracts. We generally assume that transaction costs are a fixed percentage of the change in value of the stock part of the hedge. That is, if we have stock worth $HSF(t)$ at $t = 1/12, 2/12, \ldots, n$, and we need to rebalance to a stock position of $HS(t)$, and we assume transactions costs of, say, α per \$1 change in value for stocks (and 0% for bonds), then the transactions costs would be

$$\alpha\big|HS(t) - HSF(t)\big|.$$

Allowing for mortality requires simply multiplication by ${}_np_x$, as for the hedging error above, so the transactions costs at t are

$$tc_t = {}_np_x\alpha\big|HS(t) - HSF(t)\big|.$$

The total unhedged liability at t is modeled as $tc_t + he_t$, and this can be discounted at the risk free rate of interest to give the present value at issue, $PVUL$. Using stochastic simulation for the P-measure for the fund value, we can estimate the distribution for $PVUL$ and can apply a risk measure, such as the quantile or CTE measures.

The total capital required at issue would be the sum of the hedge cost and the capital requirement. The price would be the price of the hedge plus some allowance for the cost of carrying the additional capital requirement for the unhedged liability.

In Figure 1 we show the CTE for all possible values of α, $0 \leqslant \alpha < 1$ for the present value of future outgo for both the actuarial and dynamic hedging risk management approaches. The CTE with $\alpha = 0$ is the mean present value of future outgo. On average, the actuarial approach is cheaper. At the other end however, we see that for values of α greater than around 80%, the tail mean values for the actuarial approach are considerably higher than for the dynamic hedge approach. This is all consistent with the principle of risk and return. The dynamic hedging approach mitigates the tail risk, at the expense of a higher mean cost. Actuaries are still somewhat divided on the best approach, though the dynamic hedge approach is gaining in popularity.

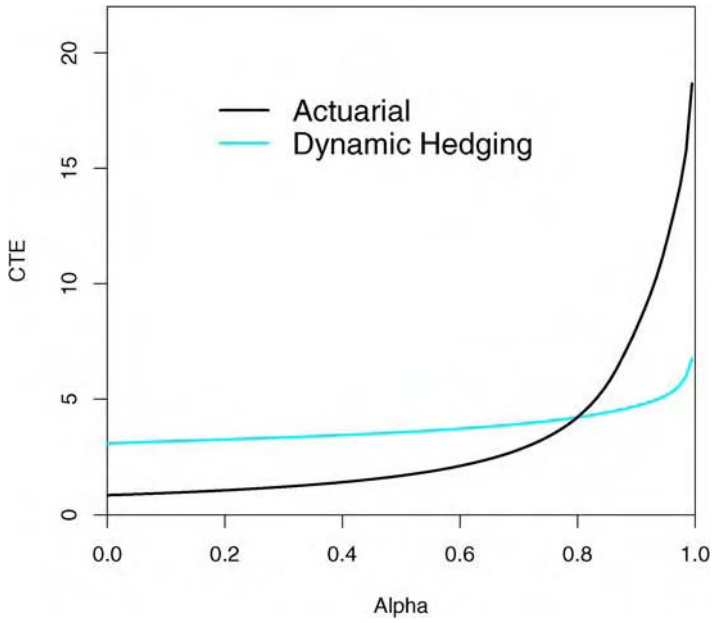


Fig. 1. CTE for GMMB, actuarial and dynamic hedging risk management; Cost per $100 single premium.

6 Guaranteed annuity options

In this section we discuss guaranteed annuity options. These guarantees are sometimes included in contracts designed to produce retirement income. One form of the guarantee provides that the yearly annuity payment will not be less than some minimum amount. The guarantee is more likely to be valuable when interest rates decline. In some respects this guarantee resembles a put option on interest rates but as we will see the actual guarantees offered by insurance companies are often much more complicated and their value depends on other variables as well as the interest rate. The most dramatic example of the importance of these guarantees was in the case of the Equitable Life Insurance company and this is the topic of the current section. These guarantees were responsible for the demise of Equitable Life (UK), the oldest insurance company in the world.

Often the guaranteed annuity options have been viewed by insurers as having negligible value and were not taken into account when the products were priced and were ignored when setting up reserves. These options can be very long dated, lasting 30 to 40 years, and over such long time spans there can be significant fluctuations in economic variables which affect the value of these options. The case of guaranteed annuity options in the UK provides a dramatic illustration of this phenomenon. Guaranteed annuity options have proved to be a significant risk management challenge for several UK insurance compa-

nies. Bolton et al. (1997) describe the origin and nature of these guarantees. The major factors which affected the value of these options included a decline in long term interest rates and improvements in mortality. For many contracts the liability is also related to equity performance.

We now describe these guarantees and explain why they became such a severe problem. Under a guaranteed annuity, the insurance company guarantees to convert the maturing policy proceeds into a life annuity at a fixed rate. Typically, these policies mature when the policyholder reaches a certain age. In the UK the most popular guaranteed rate for males, aged sixty five, was 111 per annum per 1000 and we use this rate in our illustrations. If the prevailing annuity rates at maturity are such that the annual payment per 1000 exceeds 111, a rational policyholder would opt for the prevailing market rate. On the other hand, if the prevailing annuity rates at maturity produce a lower amount than 111, a rational policyholder would take the guaranteed annuity rate. A life annuity is affected by interest rate movements in the same way as a bond. As interest rates rise the annuity amount purchased by a lump sum of 1000 increases and as interest rates fall the annuity amount available per 1000 falls. Hence the guarantee corresponds to a put option on interest rates.

These guarantees began to be included in some UK policies over fifty years ago and they became very popular in the 1970s and 1980s. Long term interest rates throughout the world were quite high in 1970s and 1980s. During these two decades the average UK long term interest rate was around 11%. The break even interest rate implicit in the guaranteed annuity options depends on the mortality assumption, but based on the mortality basis used in the original calculations, this interest rate was in the region of 5–6 percent. We can think of the break even rate as the strike price of the option. At inception these options were far out of the money and the insurance companies apparently assumed that interest rates would never fall to these *low levels* again. This presumption was incorrect and interest rates did fall in the 1990s.

The guaranteed annuity conversion rate is a function of the assumed interest rate and the assumed mortality rate. There was an unprecedented improvement in the mortality of the relevant segment of the UK population during the period 1970–2000. This improvement lead to an increase in the break-even interest rate at which the guarantee applied. We can illustrate this point as follows. An amount of 1000 is equivalent to a thirteen year annuity certain of 111 p.a. at an interest rate of 5.70% per annum. The same lump sum is equivalent to a sixteen year annuity certain of 111 p.a. at a rate of 7.72%. If mortality rates improve the annuity is payable for a longer expected term and the break even interest rate at which the option comes into the money will increase.

There was another factor which also affected the size of the liability under these guarantees. The value of the guarantee at maturity (time T) for the benchmark contract is

$$S(T)\max\left[\left(\frac{a_{65}(T)}{9}-1\right),0\right], \tag{4}$$

where $S(T)$ is the size of the proceeds at time T and $a_{65}(T)$ is market annuity rate[9] at time T for a life aged 65. The market annuity rate depends on long term interest rates and the mortality assumptions used. We see that the option will have a positive value at maturity (*be in the money*) whenever the current annuity factor exceeds the guaranteed factor (9 in this case).

It is clear from Equation (4) that the size of the option if the guarantee is in force is proportional to $S(T)$. The size of $S(T)$ will depend on the nature of the contract and also on the investment returns attributed to the policy. The procedure by which the investment returns are determined depends on the terms of the policy. These guarantees applied to two main types of policies: with profits policies and unit linked policies. For the sake of brevity we will just describe the unit linked contracts.

Under a unit linked policy the investment gains and losses are distributed directly to the policyholder's account. Contracts of this nature have become very popular in many countries in recent years because of their transparency. Under a unit linked contract the size of the option liability, if the guarantee is operative, will depend on the investment performance of the assets in which the funds are invested. In the UK there is a strong tradition of investing in equities and during the period from 1980 until 2000 the rate of growth on the major UK stock market index was 18% per annum.

Hence three principal factors contributed to the growth of the guaranteed annuity option liabilities in the UK over the last few decades. First, there was a large decline in long term interest rates over the period. Second, there was a significant improvement in longevity that was not factored into the initial actuarial calculations. Third, the strong equity performance during the period served to further increase the magnitude of the liabilities. It would appear that these events were not considered when the guarantees were initially granted. It is clear now with the benefit of hindsight that it was imprudent to grant such long term open ended guarantees of this type.

Although these guarantees were neglected until it was too late, a number of papers have discussed ways of better managing the risk using the methods described in Section 4. Yang (2001) and Wilkie et al. (2003) focus mainly on the actuarial approach but they also discuss dynamic hedging as well. They conclude that the actuarial approach, if applied from the outset, would have at least partially solved the problem. At a minimum this approach would have alerted companies much earlier to the costs of these guarantees as the options started to move into the money. Regarding the dynamic hedging approach, they conclude, that it would not have worked because the required traded securities were not available. Boyle and Hardy (2003) discuss the challenges involved in hedging these guarantees and conclude that even with our current knowledge of financial engineering this would be a very difficult task. Pelsser

[9] In other words $a_{65}(T)$ denotes the value at time T of an annuity of one per annum payable during the surviving lifetime of a life aged 65 (at time T).

(2003) analyzes a hedging strategy based on the purchase of long dated receiver swaptions. This approach deals only with the interest rate component of the guarantee. Another possibility is for the insurer to reinsure the liability with another financial institution. Dunbar (1999) discusses some of the details of this approach and describes how Scottish Widows offset its guaranteed annuity liabilities by purchasing a structured product from an investment bank.

7 Conclusions

In this chapter we have given a short account of the application of financial engineering to insurance problems. We saw that the financial options embedded in insurance products can be quite complex and that this leads to challenging risk management problems. We contrasted the traditional actuarial approach with dynamic hedging and discussed the risk management of the Guaranteed Minimum Maturity Benefit and the Guaranteed Annuity Option contract. In a survey chapter of this nature one has to be selective and there are several important topics that we have not discussed.

For example, one current area of research interest to insurance scholars concerns pricing in incomplete markets. As is well known the usual no arbitrage approach does not furnish a unique price in this cases. Under the Föllmer–Schweizer (1991) approach the contract can be priced by minimizing the squared hedging error. El Karoui and Quenez (1995) propose a super hedging procedure while Föllmer and Leukert (1999) describe quantile hedging. These methods have been applied to the pricing of insurance contracts by Moeller (1998). In addition Kolkiewicz and Tan (2004) have implemented a robust hedging approach to deal with the incompleteness in regime switching models.

Another area concerns the question of the optimal contract design. Arrow (1973) and Raviv (1979) have studied the question of optimal contract design in the case of nonlife insurance contracts. Brennan (1993) demonstrated that the UK version of the with profits contract has an inefficient contract design. There is widespread evidence that retail investors like to have downside protection as well as upside participation in the market. It is of interest to explore the characteristics of the optimal design of an equity linked contract. Boyle and Tian (2006) have taken initial steps in this direction. They propose a contract design under which the investor maximizes expected utility subject to a guaranteed minimum and optimal participation in the equity market.

Acknowledgements

Both authors thank the Natural Sciences and Engineering Research Council of Canada for support. They are also grateful to Shannon Kennedy for research assistance.

References

Arrow, K.J. (1973). Optimal insurance and generalized deductibles. Rand Corporation.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9 (3), 203–228.

Bolton, M.J., Carr, D.H., Collins, P.A., George, C.M., Knowles, V.P., Whitehouse, A.J. (1997). Reserving for annuity guarantees. *Report of the Annuity Guarantees Working Party*, Faculty and Institute of Actuaries Technical Report.

Boyle, P.P., Hardy, M.R. (1997). Reserving for maturity guarantees: Two approaches. *Insurance: Mathematics and Economics* 21 (2), 113–127.

Boyle, P.P., Hardy, M. (2003). Guaranteed annuity options. *Astin Bulletin* 33 (2), 125–152.

Boyle, P.P., Schwartz, E.S. (1977). Equilibrium prices of guarantees under equity-linked contracts. *Journal of Risk and Insurance* 44 (4), 639–660.

Boyle, P.P., Tian, W. (2006). Optimal equity indexed annuity design. *Working paper*, University of Waterloo.

Brennan, M.J. (1993). Aspects of insurance, intermediation and finance. *Geneva Papers on Risk and Insurance* 18, 7–30.

Brennan, M.J., Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee. *Journal of Financial Economics* 3, 195–213.

Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*. Springer-Verlag, New York.

Dunbar, N. (1999). Sterling swaptions under new scrutiny. *Risk* December, 33–35.

El Karoui, Quenez (1995). Dynamic programming and pricing of contingent claims in an incomplete market. *SIAM Journal of Control and Optimization* 33, 29–66.

Föllmer, H., Leukert, P. (1999). Quantile hedging. *Finance and Stochastics* 3, 251–273.

Föllmer, Schweizer (1991). Hedging of contingent claims under incomplete information. In: Davis, M.H., Elliot, R.J. (Eds.), *Applied Stochastic Analysis*. In: *Stochastic Monographs*, vol. 5. Gordon and Breach, New York/London, pp. 387–414.

Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*. *Huebner Foundation Monograph*, vol. 8. Wharton School, University of Pennsylvania.

Hamilton, J.D. (1989). A new approach to the economic analysis of non-stationary time series. *Econometrica* 57, 357–384.

Hardy, M.R. (2001). A regime switching model of long term stock returns. *North American Actuarial Journal* 5 (2), 41–53.

Kolkiewicz, W.A., Tan, K.S. (2004). Volatility risk for regime switching models. *North American Actuarial Journal* 8, 127–145.

Lin, X., Tan, K.S. (2003). Valuation of equity-indexed annuities under stochastic interest rate. *North American Actuarial Journal* 7 (4), 72–91.

Lundberg, F. (1909). Über die Thoerie der Rückversichung. *Transactions VI International Congress of Actuaries*, 877–895.

Maturity Guarantees Working Party (MGWP) (1980). Report of the Maturity Guarantees Working Party. *Journal of the Institute of Actuaries* 107, 103–209.

McCarthy, C. (2004). *Talk at the Insurance Institute of London Luncheon*. The Mansion House, London.

Moeller, T. (1998). Risk minimizing hedgeing strategies for unit linked life insurance contracts. *ASTIN Bulletin* 28, 17–47.

Ogborn, M.E. (1956). The professional name of actuary. *Journal of the Institute of Actuaries* 82, 833–846.

Pelsser, A. (2003). Pricing and hedging guaranteed annuity options via static option replication. *Insurance: Mathematics and Economics* 33 (2), 283–296.

Raviv, A. (1979). The design of an optimal insurance policy. *American Economic Review* 69, 84–96.

Redington, F.M. (1952). A review of the principles of life office valuation. *Journal of the Institute of Actuaries* 78, 286–315.

Wang, S.X. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transform. *Insurance: Mathematics and Economics* 17, 43–54.

Wang, S.X. (2002). A universal framework for pricing financial and insurance risks. *ASTIN Bulletin* 32 (2), 213–234.

Whitten, S.P., Thomas, R.G. (1999). A non-linear stochastic model for actuarial use. *British Actuarial Journal* 5, 919–953.

Wilkie, A.D. (1986). A stochastic investment model for actuarial use. *Transactions of the Faculty of Actuaries* 39, 341–381.

Wilkie, A.D. (1995). More on a stochastic asset model for actuarial use. *British Actuarial Journal* 1 (V), 777–964.

Wilkie, A.D., Waters, H.R., Yang, S. (2003). Reserving, pricing and hedging for policies with guaranteed annuity options. *British Actuarial Journal* 9 (2), 263–291.

Yang, S. (2001). Reserving, pricing and hedging for guaranteed annuity options. *Ph.D. thesis*, Department of Actuarial Mathematics and Statistics, Heriot Watt University, Edinburgh.

PART VI

Portfolio Optimization

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15019-2

Chapter 19

Dynamic Portfolio Choice and Risk Aversion

Costis Skiadas

Kellogg School of Management, Department of Finance, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA
E-mail: c-skiadas@kellogg.northwestern.edu

Abstract

This chapter presents a theory of optimal lifetime consumption-portfolio choice in a continuous information setting, with emphasis on the modeling of risk aversion through generalized recursive utility. A novel contribution is a decision theoretic development of the notions of source-dependent first- or second-order risk aversion. Backward stochastic differential equations (BSDEs) are explained heuristically as continuous-information versions of backward recursions on an information tree, and are used to formulate utility functions as well as optimality conditions. The role of scale invariance and quadratic BSDEs in obtaining tractable solutions is explained. A final section outlines extensions, including optimality conditions under trading constraints, and tractable formulations with nontradeable income.

1 Introduction

This chapter analyzes the optimal consumption-portfolio choice of a risk-averse agent, with emphasis on the modeling of risk aversion given a stochastic investment opportunity set. The main part of the analysis is based on Schroder and Skiadas (2003). A novel contribution is a decision theoretic development of the notions of source-dependent first- or second-order risk aversion that are implicit in the utility representations of Schroder and Skiadas (2003). These ideas unify, at least in the context of continuous information, standard notions of risk aversion with some models of ambiguity aversion or robustness that have recently received considerable attention in the literature. The dynamic portfolio methodology presented should, however, also be of interest to readers only concerned with conventional source-independent risk aversion in a dynamic setting.

Following Merton's (1969, 1971) seminal work, most papers on dynamic portfolio choice assume that the investor maximizes time-additive expected

utility, that we refer to as "additive utility" for the purposes of this discussion. The limitations of additive utility in modeling risk aversion are, however, well recognized by now (see, for example, Epstein, 1992). We will argue that any two additive utilities that imply identical preferences over deterministic plans must be ordinally equivalent, and therefore equally risk-averse. In this chapter, we consider utility functions for which risk aversion can be adjusted without changing the utility value of deterministic plans. The stochastic setting is one in which information is revealed continuously by a finite set of Brownian motions. Utility is defined over consumption plans in terms of a single nondurable good, with an exogenous planning horizon and endowment. Markets can be incomplete, but they are sufficiently complete so that the investor's endowed income stream is tradeable, and there are no other trading constraints or transaction costs. The last section outlines extensions dealing with trading constraints and nontradeable income, and points to further extensions in the literature relaxing various combinations of the above assumptions (typically at the cost of other restrictions).

In its simplest form, the utility function we will adopt is the Duffie and Epstein (1992) continuous-information limit of the recursive utility of Kreps and Porteus (1978), which includes the widely used homothetic recursive utility specification of Epstein and Zin (1989) (a special case of which is expected discounted power or logarithmic utility[1]). In the Kreps–Porteus formulation, current utility is computed as a function of current consumption and a von Neumann–Morgenstern (1944) certainty equivalent of the continuation utility. Given sufficient smoothness, the classic analysis of small risks of Arrow (1965, 1970) and Pratt (1964) implies that the certainty equivalent can be replaced by a quadratic criterion in the continuous-information limit, which is the reason why some elements of the original single-period mean–variance portfolio analysis of Markowitz (1952) survive in a continuous-information setting with Duffie–Epstein utility. Assuming constant relative risk aversion, the optimal portfolio is a weighted sum of an instantaneously mean–variance efficient portfolio and a hedging portfolio that accounts for the stochastic nature of the investment opportunity set (and vanishes in the case of i.i.d. instantaneous returns).

An extension of Duffie–Epstein utility we will consider allows risk aversion to depend on the source of risk. For example, investors have been documented to show a preference toward investing in the familiar: domestic stocks, firms whose products are familiar, local firms, one's employer's stock.[2] The well-known experimental findings of Ellsberg (1961), and a large literature following it, show that subjects prefer to bet on risk sources to which probabilities can

[1] Under some regularity, a homothetic additive utility is necessarily the additive special case of Epstein–Zin utility. On the other hand, Epstein–Zin utility is only a parametric special case of the much broader class of homothetic Duffie–Epstein utilities.

[2] Daniel et al. (2002) survey such psychological biases in asset markets.

be more unambiguously assigned, a phenomenon known as ambiguity aversion.[3] One can think of risk as reflecting not only the risk that is conditional on the assumed model of the risk source, but also uncertainty about the model's validity, which is itself too difficult to model. Since model risk can vary with the source of risk, it is useful to consider source-dependent risk aversion. With this motivation, we will extend the Kreps–Porteus recursion by letting the certainty equivalent be a function of the entire vector of continuation utilities attributable to each Brownian motion separately. The locally-quadratic analysis under Duffie–Epstein utility extends to this case, but with a different coefficient of risk aversion assigned to each source of risk.

As shown in Skiadas (2003), Duffie–Epstein utility includes the "robust" specifications of Anderson et al. (2000), Hansen et al. (2001) and Maenhout (1999). Similarly, the criterion of Uppal and Wang (2003) is equivalent to a special form of recursive utility with source-dependent risk aversion (included in the "quasi-quadratic proportional aggregator" specification of Schroder and Skiadas, 2003). The multiple-prior expressions of these authors suggest a robustness interpretation of risk aversion. Conversely, their robustness interpretation of multiple-prior formulations can be thought of as risk aversion in the context of recursive utility. To avoid this semantic redundancy, in this chapter we define formally only risk aversion, and we think of robustness or ambiguity aversion as an informal consideration in selecting the degree of risk aversion toward a given source of risk.

Another way in which the Duffie–Epstein representation will be extended relates to the distinction between first- and second-order risk aversion made in a static setting by Segal and Spivak (1990). The Arrow–Pratt analysis, and by extension the Duffie–Epstein limit of Kreps–Porteus utility, relies on the smoothness of the von Neumann–Morgenstern certainty equivalent, an assumption for which there is no compelling decision-theoretic justification. Smooth expected utility implies local risk-neutrality, meaning that an investor should be willing to undertake any actuarially favorable investment in sufficiently small scale, and should be unwilling to perfectly insure a sufficiently small risk at actuarially unfavorable terms. We will consider a source-dependent extension of Kreps–Porteus utility with nonsmooth certainty equivalent for which these conclusions are invalidated, and we will derive corresponding optimal trading strategy expressions that highlight the relationship between first-order risk aversion and portfolio holdings.

Motivated by the notion of ambiguity aversion, Epstein and Schneider (2003) formulated a multiple prior utility, whose continuous-information limit was studied by Chen and Epstein (2002). Consistent with the view of ambiguity aversion as being a form of risk aversion, the Chen–Epstein "κ-ignorance" formulation is mathematically equivalent to a case of the above mentioned extension of Duffie–Epstein utility with source-dependent first-order risk aversion.

[3] The view of ambiguity aversion as a form of risk aversion is further supported by the arguments of Klibanoff et al. (2002).

Lazrak and Quenez (2003) analyzed the properties of a utility that is defined as a solution to a general backward stochastic differential equation (BSDE), and includes the Chen–Epstein formulation. Lazrak and Quenez made the important observation that comparative risk aversion can depend on the "direction" of risk. Complementing the Lazrak–Quenez analysis, this chapter provides a heuristic decision–theoretic foundation of their proposed utility form, that we will refer to simply as "recursive utility." The more specific models of risk aversion discussed above correspond to special functional forms of recursive utility.

Following the development of Schroder and Skiadas (2003), optimality conditions will first be derived for general concave recursive utilities, as a system of forward–backward stochastic differential equations (FBSDEs). The forward component of the system is the wealth process, which follows the investor's budget equation, and the backward components are the utility and shadow-price-of-wealth processes. The FBSDE system uncouples if the problem is scale-invariant (with respect to wealth). Combining scale-invariance with the various types of risk aversion discussed above, we will be able to formulate some interesting optimal trading strategy expressions, in terms of the solution to a single BSDE. Moreover, we will give some examples of preferences and stochastic investment opportunity sets for which the BSDE of the optimality conditions takes a quadratic form whose solution can be reduced to a tractable ODE system. A parallel theory based on translation-invariant recursive utility (which generalizes expected discounted exponential utility) can be found in Schroder and Skiadas (2005a), and is briefly discussed in the final section.

Merton approached the dynamic optimal portfolio selection problem using the Hamilton–Jacobi–Bellman equation of optimal control theory, modern expositions of which are given by Fleming and Soner (1993) and Yong and Zhou (1999). Examples of solutions with Epstein–Zin utility using this method are Giovannini and Weil (1989), Svensson (1989), Obstfeld (1994), Zariphopoulou and Tiu (2002), and Chacko and Viceira (2005). Cox and Huang (1989) and Karatzas et al. (1987) rederived the Merton solution by using the state-price density property of marginal utilities at the optimum, in a way that relied on utility additivity. This "utility gradient approach" was generalized to include recursive utilities in Duffie and Skiadas (1994), Schroder and Skiadas (1999), El Karoui et al. (2001), and Schroder and Skiadas (2003, 2005a, 2005b), and is the method adopted in this chapter. [An alternative dynamic programming derivation of the scale-invariant solutions is outlined in Schroder and Skiadas (2003).] While some further leads to the literature will be given in the final section, this chapter is not intended as a literature survey, and no attempt has been made to be comprehensive. Monographs or collected papers on dynamic portfolio choice include Merton (1990), Korn (1997), Sethi (1997), Karatzas and Shreve (1998), Gollier (2001), and Campbell and Viceira (2002). An overview of the econometrics of portfolio choice is given by Brandt (forthcoming).

The mathematical background for this chapter is outlined in the appendices of Duffie (2001), and is covered in detail by Karatzas and Shreve (1988). Less widely known are the more recent mathematical tools of BSDEs and FBSDEs,

a general perspective on which can be found in the expositions of El Karoui et al. (1997) and Ma and Yong (1999).

The remainder of this chapter is organized in five sections. Section 2 sets up the problem and characterizes the optimum with minimal restrictions on preferences over consumption plans. Section 3 develops recursive utility, and the associated optimality conditions. Section 4 motivates some more specialized recursive utility forms, representing the various types of risk aversion introduced above. Section 5 formulates optimality conditions for these special recursive utility forms, assuming utility homotheticity. Section 6 concludes with an outline of several extensions.

2 Optimality and state pricing

This section introduces the stochastic setting, the investor problem, and the basic optimality verification argument in terms of the state price density property of a utility supergradient density. The essential tool of linear BSDEs is introduced in the context of state pricing. The section imposes only minimal preferences restrictions, and concludes with a discussion of the inadequacy of additive utility as a representation of risk aversion. The discussion of recursive utility begins with Section 3.

2.1 Dynamic investment opportunity set

Uncertainty is represented by the probability space $(\Omega, \mathcal{F}, P)$, on which is defined a d-dimensional Brownian motion $B = (B^1, \ldots, B^d)'$ over a finite time-horizon $[0, T]$. As with every vector in this chapter, B is a column vector, and the prime denotes transposition. Information is represented by the (augmented) filtration $\{\mathcal{F}_t\colon\ t \in [0, T]\}$ generated by the Brownian motion B. Intuitively, we think of an information tree whose time-t nodes or *spots* correspond to the possible paths of B up to time t. A *time-t spot* is therefore a continuous function of the form $\omega^t : [0, t] \to \mathbb{R}^d$. Conditional expectation given time-t information, $\mathcal{F}_t$, is denoted E_t. Similarly, covariance (variance) given $\mathcal{F}_t$ is denoted cov_t (var_t). We assume that $\mathcal{F} = \mathcal{F}_T$, and therefore $E_T[x] = x$ for every random variable x.

A *process* in this chapter is by definition a stochastic process that is progressively measurable with respect to $\{\mathcal{F}_t\}$. For any process x, we think of the time-t value x_t (alternatively denoted $x(t)$) as a function of the realized spot ω^t. In heuristic explanations (that ignore issues regarding null sets) we will write $x[\omega^t]$ to express this dependence. Given any subset S of some Euclidean space, we let $\mathcal{L}(S)$ denote the set of processes of the form $x : \Omega \times [0, T] \to S$. For any integer p, typically $p = 1$ or 2, we define the set $\mathcal{L}_p(S)$ of all $x \in \mathcal{L}(S)$ such that $\int_0^T |x_t|^p \, \mathrm{d}t < \infty$ with probability one (where $|\cdot|$ denotes Euclidean norm).

We consider a financial market allowing instantaneous default-free borrowing and lending at a continuously-compounded rate given by the process r.

A dollar can be invested from time t to time $t + \mathrm{d}t$ earning interest $r_t\,\mathrm{d}t$, which is risk-free in the sense that $\mathrm{Var}_t[r_t\,\mathrm{d}t] = 0$, but whose value depends on time-t information. For expositional simplicity, r is assumed bounded (although this assumption is violated in some later applications). The rest of the market consists of trading in m risky assets, whose cumulative *excess* returns are represented by the m-dimensional process $R = (R^1, \ldots, R^m)'$. A dollar invested at time t in risky asset i is worth $1 + r_t\,\mathrm{d}t + \mathrm{d}R_t^i$ at time $t + \mathrm{d}t$.

We assume that R is an Itô process with dynamics

$$\mathrm{d}R_t = \mu_t^R\,\mathrm{d}t + \sigma_t^{R\prime}\,\mathrm{d}B_t, \quad \mu^R \in \mathcal{L}_1\big(\mathbb{R}^m\big),\ \sigma^R \in \mathcal{L}_2\big(\mathbb{R}^{d\times m}\big). \tag{1}$$

There is, therefore, one column of σ^R for every risky asset, and one row for every component of the Brownian motion B. The investment opportunity set is defined by the triple (r, μ^R, σ^R), whose value can vary from spot to spot. We think of (1) as an instantaneous linear factor model, where

$$\mu_j^R(t)\,\mathrm{d}t = E_t\big[\mathrm{d}R_t^j\big] \quad \text{and} \quad \sigma_{ij}^R(t)\,\mathrm{d}t = \mathrm{cov}_t\big[\mathrm{d}B_t^i, \mathrm{d}R_t^j\big],$$
$$i = 1, \ldots, d,\ j = 1, \ldots, m.$$

Since $E_t[\mathrm{d}B_t] = 0$ and $E_t[\mathrm{d}B_t\,\mathrm{d}B_t'] = I\,\mathrm{d}t$ (where I is an identity matrix) the conditional variance–covariance matrix of $\mathrm{d}R_t$ is

$$E_t\big[\big(\mathrm{d}R_t - E_t[\mathrm{d}R_t]\big)\big(\mathrm{d}R_t - E_t[\mathrm{d}R_t]\big)'\big] = \sigma_t^{R\prime}\sigma_t^R\,\mathrm{d}t.$$

A time-t *allocation* is an $\mathcal{F}_t$-measurable random vector $\psi_t = (\psi_t^1, \ldots, \psi_t^m)'$, where ψ_t^i represents the proportion of wealth invested at time t in risky asset i, with the remaining nonconsumed wealth invested risk-free. Negative proportions indicate short positions. The choice of a time-t allocation can depend on time-t information, and therefore we think of ψ_t as a function of the realized time-t spot. A dollar invested at time t according to allocation ψ_t is worth

$$1 + r_t\,\mathrm{d}t + \psi_t'\,\mathrm{d}R_t = 1 + \big(r_t + \psi_t'\mu_t^R\big)\,\mathrm{d}t + \big(\sigma_t^R\psi_t\big)'\,\mathrm{d}B_t$$

at time $t + \mathrm{d}t$. The vector $\sigma_t^R\psi_t$ represents the *risk profile* of the allocation ψ_t, since it specifies the loadings of the instantaneous excess return $\psi_t'\,\mathrm{d}R_t$ on the instantaneous factors $\mathrm{d}B_t$.

If the column span of σ^R is $\mathbb{R}^d$ at all times then the market is complete, in the sense that every risk profile is feasible through some allocation at all times. We do *not* assume that the market is complete, allowing the rank of σ^R to be less than d. While the market can be effectively complete even if σ^R drops rank, we will consider applications in which market incompleteness is a binding constraint. We will not allow, however, the rank of σ^R to vary from spot to spot, and we assume that at no spot of the information tree are any of the assets redundant over an infinitesimal time interval. This is the economic content of the following condition, assumed throughout:

Asset nonredundancy. The columns of σ^R, corresponding to the m risky assets, are everywhere linearly independent, and therefore $m \leqslant d$.

As a consequence of this assumption, the $m \times m$ instantaneous variance–covariance rate matrix $\sigma^{R\prime}\sigma^R$ is everywhere invertible. If σ_t is a risk profile attainable through the allocation ψ_t, meaning that $\sigma_t^R \psi_t = \sigma_t$, then ψ_t is the unique allocation with this property, and is given by

$$\psi_t = \left(\sigma_t^{R\prime}\sigma_t^R\right)^{-1}\sigma_t^{R\prime}\sigma_t. \tag{2}$$

The traditional portfolio analysis of Markowitz (1952) can be applied conditionally spot-by-spot on the information tree. Selecting an allocation ψ_t results in an instantaneous excess return with conditional mean and variance

$$E_t[\psi_t' \,\mathrm{d}R_t] = \psi_t'\mu_t^R \,\mathrm{d}t \quad \text{and} \quad \mathrm{var}_t[\psi_t' \,\mathrm{d}R_t] = \psi_t'\sigma_t^{R\prime}\sigma_t^R\psi_t \,\mathrm{d}t.$$

Let μ_t be any $\mathcal{F}_t$-measurable random variable. Minimizing $\mathrm{var}_t[\psi_t' \,\mathrm{d}R_t]$ subject to the constraint $E_t[\psi_t' \,\mathrm{d}R_t] = \mu_t \,\mathrm{d}t$ results in an allocation of the form

$$\psi_t = k_t\left(\sigma_t^{R\prime}\sigma_t^R\right)^{-1}\mu_t^R,$$

for some $\mathcal{F}_t$-measurable random variable k_t that depends on the choice of μ_t. We call an allocation of this form *instantaneously minimum-variance efficient*. The corresponding squared conditional instantaneous Sharpe ratio is maximized, and is given by

$$\frac{E_t[\psi_t' \,\mathrm{d}R_t]^2}{\mathrm{var}_t[\psi_t' \,\mathrm{d}R_t]} = \mu_t^{R\prime}\left(\sigma_t^{R\prime}\sigma_t^R\right)^{-1}\mu_t^R \,\mathrm{d}t. \tag{3}$$

2.2 *Strategies, utility, and optimality*

An optimal investment strategy is one that finances a consumption plan for which there exists no other consumption plan that is both more desirable and feasible. In this subsection we formalize this notion, while placing minimal restrictions on investor preferences.

We let $\mathcal{H}$ denote the Hilbert space of every $x \in \mathcal{L}(\mathbb{R})$ such that $E[\int_0^T x_t^2 \,\mathrm{d}t + x_T^2] < \infty$, with the inner product

$$(x \mid y) = E\left[\int_0^T x_t y_t \,\mathrm{d}t + x_T y_T\right], \quad x, y \in \mathcal{H}.$$

The set of *strictly positive*[4] elements of $\mathcal{H}$ is $\mathcal{H}_{++} = \mathcal{H} \cap \mathcal{L}(\mathbb{R}_{++})$. The element of $\mathcal{H}$ that is identically equal to one is denoted $\mathbf{1}$.

We postulate a convex cone $\mathcal{C} \subseteq \mathcal{H}_{++}$ of *consumption plans* such that $\mathbf{1} \in \mathcal{C}$, and for any $x \in \mathcal{H}$ and $y, z \in \mathcal{C}$, $y \leqslant x \leqslant z$ implies $x \in \mathcal{C}$. For any $c \in \mathcal{C}$ and

[4] More precisely, any two processes x, y such that $(x - y \mid x - y) = 0$ are identified as elements of $\mathcal{H}$. A *strictly positive* element of $\mathcal{H}$ is one that can be identified in this way with a process in $\mathcal{L}(\mathbb{R}_{++})$.

time $t < T$, we interpret c_t as the time-t consumption rate, while c_T represents a terminal lump-sum consumption or bequest. In a typical application, $\mathcal{C}$ is specified by some integrability restriction required for a utility function to be well defined. The strict positivity of consumption plans reflects our implicit assumption that a consumption nonnegativity constraint is nonbinding. In later sections, the positivity of optimal consumption will be enforced by assuming infinite marginal utility at zero.

We consider an investor with initial wealth $w_0 > 0$ and no subsequent income. (This includes the case of an endowed income stream as long as it can be traded.) A *consumption strategy* is any process $\rho \in \mathcal{L}_1(\mathbb{R}_{++})$ such that $\rho_T = 1$. For $t < T$, we interpret ρ_t as the investor's consumption rate as a proportion of time-t wealth, while the convention $\rho_T = 1$ is used below to express the assumption that final wealth equals terminal consumption. A *trading strategy* is any process $\psi \in \mathcal{L}(\mathbb{R}^m)$ such that $\psi'\mu^R \in \mathcal{L}_1(\mathbb{R})$ and $\sigma^R\psi \in \mathcal{L}_2(\mathbb{R}^d)$, with ψ_t representing a time-t allocation. A *strategy* is a pair (ρ, ψ) of a consumption strategy and a trading strategy.

The *wealth process* W *generated* by a strategy (ρ, ψ) is defined through the *budget equation*

$$\frac{\mathrm{d}W_t}{W_t} = (r_t - \rho_t)\,\mathrm{d}t + \psi_t'\,\mathrm{d}R_t, \quad W_0 = w_0. \tag{4}$$

The consumption plan c is *financed* by the strategy (ρ, ψ) if $c = \rho W$, meaning that $c_t = \rho_t W_t$ for every time t (and therefore $c_T = W_T$). A consumption plan is *feasible* if it can be financed by some strategy.

The investor's problem is to select a feasible consumption plan that is optimal. To define optimality, we introduce utility functions. We say that the investor *prefers* plan b to plan a at spot ω^t if, conditionally on the realization of ω^t, an agent with plan a as the status quo would switch to plan b if presented with the opportunity to do so at no cost. The investor is *indifferent* between two plans if neither plan is preferred to the other.

We are going to measure utility concretely by taking as a unit the consumption plan $\mathbf{1}$. We assume throughout that the investor prefers more consumption to less, and therefore, given any scalars α, β such that $\beta > \alpha > 0$, the agent prefers $\beta\mathbf{1}$ to $\alpha\mathbf{1}$ at every spot. We further assume that, given any consumption plan c and spot ω^t, there exists a (necessarily unique) scalar α such that, conditionally on the realization of spot ω^t, the agent is indifferent between plans c and $\alpha\mathbf{1}$. We call this value of α the *spot-*ω^t *cardinal utility* of c, and denote it $U(c)[\omega^t]$. Specifying a value at every spot of the information tree defines the *cardinal utility process* $U(c)$ of plan c. We note that, by definition, $U_T(c) = c_T$.

Another preference assumption we adopt is that if the investor is indifferent between a and a' and between b and b', then the investor prefers b to a if and only if the investor prefers b' to a'. Applying this condition with $a' = U(a)[\omega^t]\mathbf{1}$ and $b' = U(b)[\omega^t]\mathbf{1}$, we conclude that, conditionally on the realization of spot ω^t, the investor prefers plan b to plan a if and only if $U(b)[\omega^t] > U(a)[\omega^t]$.

The investor's objective at spot ω^t is therefore to select the feasible consumption plan c of maximum spot-ω^t utility $U(c)[\omega^t]$.

Utility maximization at every spot can be an inconsistent objective, since the investor may have an incentive to deviate at some spot from a strategy selected at an earlier spot. We exclude this possibility by assuming the following key condition throughout.

Dynamic consistency. Suppose two consumption plans a and b are equal up to a stopping time τ, and $P[U_\tau(b) \geqslant U_\tau(a)] = 1$. Then $U_0(b) \geqslant U_0(a)$, with the inequality being strict if $P[U_\tau(b) > U_\tau(a)] > 0$.

Suppose time-zero utility is maximized by the strategy (ρ, ψ), which finances the consumption plan c, and generates the wealth process W. Then there cannot exist a stopping time τ and trading strategy $(\tilde{\rho}, \tilde{\psi})$, that finances consumption plan $\tilde{c}$ and generates a wealth process $\tilde{W}$, such that $W_\tau = \tilde{W}_\tau$, $P[U_\tau(\tilde{c}) \geqslant U_\tau(c)] = 1$, and $P[U_\tau(\tilde{c}) > U_\tau(c)] > 0$. Otherwise, by dynamic consistency, the strategy that starts as (ρ, ψ) and switches to $(\tilde{\rho}, \tilde{\psi})$ at time τ would result in higher time-zero utility than $U_0(c)$, contradicting the time-zero optimality of strategy (ρ, ψ).

Dynamic consistency justifies the following definition of optimality in terms of the single time-zero utility function $U_0 : \mathcal{C} \to \mathbb{R}$.

Definition 1. The consumption plan c is *optimal* if it is feasible and there exists no feasible consumption plan $\tilde{c}$ such that $U_0(\tilde{c}) > U_0(c)$. A strategy (ρ, ψ) is *optimal* if it finances an optimal consumption plan. Finally, a consumption or trading strategy is *optimal* if it is part of an optimal strategy.

A function $\tilde{U}_0 : \mathcal{C} \to \mathbb{R}$ is *ordinally equivalent* to $U_0 : \mathcal{C} \to \mathbb{R}$ if $\tilde{U}_0 = f \circ U_0$ for some strictly increasing function $f : \mathbb{R} \to \mathbb{R}$. We call such a function $\tilde{U}_0$ an *ordinal utility* representation of the investor's time-zero preferences. A property of U_0 is *ordinal* if it is also true of any utility that is ordinally equivalent to U_0. Optimality of a given consumption plan relative to U_0 is an example of an ordinal property of U_0. (Extending the notion of ordinal equivalence to utility at every spot, we note that dynamic consistency is an ordinal property of the entire utility process U.)

We henceforth take as given the time-zero utility function $U_0 : \mathcal{C} \to \mathbb{R}$, which can be either ordinal or cardinal, the distinction made only where relevant. The following two properties are assumed throughout the chapter:

Monotonicity. For any c, $c + x \in \mathcal{C}$, $0 \neq x \geqslant 0$ implies $U_0(c + x) > U_0(c)$.

Concavity. For all $c^0, c^1 \in \mathcal{C}$, $\alpha \in (0, 1)$ implies $U_0(\alpha c^1 + (1 - \alpha)c^0) \geqslant \alpha U_0(c^1) + (1 - \alpha)U_0(c^0)$.

Monotonicity is an ordinal property, while concavity is not. For cardinal utility, concavity can be thought of as an expression of a preference for consumption smoothing. Later we will introduce the important class of scale-invariant problems in which U_0 is assumed to have the additional ordinal property of homotheticity. A cardinal utility is homothetic if and only if it is homogeneous of degree one, in which case concavity is equivalent to the ordinal property of quasiconcavity.

Let (ρ, ψ) be a candidate optimal strategy that generates the wealth process W and finances the consumption plan $c = \rho W$. We will verify the optimality of c by constructing a utility supergradient density at c that is also a state price density at c. These notions are defined below.

Definition 2. (a) A process $\pi \in \mathcal{H}$ is a *state price density at* c if $(\pi \mid x) \leqslant 0$ for any $x \in \mathcal{H}$ such that $c + x$ is a feasible consumption plan.

(b) A process $\pi \in \mathcal{H}$ is a *supergradient density* of U_0 at c if $U_0(c + x) \leqslant U_0(c) + (\pi \mid x)$ for every $x \in \mathcal{H}$ such that $c + x \in \mathcal{C}$.

Interpreting $(\pi \mid x)$ as a present value of x, the state-price density property states that there is no feasible incremental consumption plan relative to c that has positive present value. A supergradient density can be thought of as a generalized notion of marginal utility. Since U_0 is assumed (strictly) increasing and concave, any supergradient density of U_0 is necessarily strictly positive. Given a reference plan, the state-price density property depends on the market opportunities and not on preferences, while the supergradient density property depends on preferences and not on the market opportunities.

The following observation is the basis for optimality verification in this chapter. [While we will not discuss the necessity of optimality conditions in this chapter, a simple partial converse is given in Schroder and Skiadas (2003).]

Proposition 3. *Suppose c is a feasible consumption plan, and $\pi \in \mathcal{H}_{++}$ is a supergradient density of U_0 at c that is also a state price density at c. Then the plan c is optimal.*

Proof. If $c + x \in \mathcal{C}$ is feasible, then $U_0(c + x) \leqslant U_0(c) + (\pi \mid x) \leqslant U_0(c)$. □

2.3 State price dynamics and linear BSDEs

In order to apply the optimality verification argument of Proposition 3, we study below the dynamics of a state price density. In the process we introduce the mathematical tool of a linear backward stochastic differential equation (BSDE), which plays a basic role in this chapter and asset pricing theory in general.

The key to understanding the state price density dynamics is the following notion of risk pricing:

Definition 4. A *market-price-of-risk process* is any process $\eta \in \mathcal{L}_2(\mathbb{R}^d)$ such that

$$\mu^R = \sigma^{R\prime}\eta. \tag{5}$$

Recalling the linear-factor-model interpretation (1), the above equation can be thought of as (exact) factor pricing, with η_t^i representing the time-t price of instantaneous linear factor $\mathrm{d}B_t^i$. Since σ^R is assumed everywhere full-rank, a market-price-of-risk process is unique if and only if $m = d$.

The existence of a market-price-of-risk process is implied by the absence of arbitrage opportunities. While a rigorous statement and proof of this claim can be found in Karatzas and Shreve (1998), it is worth recalling the essential idea. In an arbitrage-free market there cannot be an instantaneously riskless allocation with positive instantaneous excess returns; that is,

$$\sigma_t^R \psi_t = 0 \quad \text{implies} \quad \psi_t' \mu_t^R = 0. \tag{6}$$

The existence of a market price of risk process is the dual equivalent to (6). Clearly, (5) implies (6). Conversely, we define the orthogonal decomposition $\mu_t^R = \sigma_t^{R\prime}\eta_t + \varepsilon_t$, where $\sigma_t^R \varepsilon_t = 0$. If (6) holds, then $\varepsilon_t' \mu_t^R = 0$, and therefore $\varepsilon_t' \varepsilon_t = \varepsilon_t' \mu_t^R = 0$, proving that $\mu_t^R = \sigma_t^{R\prime}\eta_t$.

Suppose that the process $\pi \in \mathcal{H}_{++}$ follows the dynamics

$$\frac{\mathrm{d}\pi_t}{\pi_t} = -r_t\,\mathrm{d}t - \eta_t'\,\mathrm{d}B_t, \quad t \in [0, T], \tag{7}$$

for some market-price-of-risk process η. We will argue that π is a state-price density at any given consumption plan satisfying an integrability condition.

Consider any strategy (ρ, ψ), generating the wealth process W, and financing the consumption plan $c = \rho W$. Letting $\Sigma = W\sigma^R\psi$ in the budget equation (4) and using the assumption $\mu^R = \sigma^{R\prime}\eta$ results in

$$\mathrm{d}W_t = -(c_t - r_t W_t - \eta_t'\Sigma_t)\,\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t, \quad W_T = c_T. \tag{8}$$

This is a *linear BSDE*. The Itô process W *solves* the BSDE if (8) is satisfied for some $\Sigma \in \mathcal{L}_2(\mathbb{R}^d)$. Given the solution W, the corresponding $\Sigma \in \mathcal{L}_2(\mathbb{R}^d)$ is uniquely determined (by the uniqueness of Itô representations) and therefore we can also think of a solution as being the pair $(W, \Sigma) \in \mathcal{L}_1(\mathbb{R}) \times \mathcal{L}_2(\mathbb{R}^d)$. Nonlinear BSDEs are introduced in the following section, where it is explained that a BSDE is essentially a backward recursion on the information tree. For the linear case, the backward recursion interpretation is suggested by a present value formula given in the lemma below. Even though the symbols have specific meanings in this context, the lemma is stated in a way that applies to a general linear BSDE.

Lemma 5. *Suppose that W solves BSDE (8) for some $c \in \mathcal{H}$, $r \in \mathcal{L}_1(\mathbb{R})$, and $\eta \in \mathcal{L}_2(\mathbb{R})$, and that $\pi \in \mathcal{H}_{++}$ follows the dynamics (7).*

(a) *If* $W \in \mathcal{L}(\mathbb{R}_+)$*, then*

$$W_t \geqslant \frac{1}{\pi_t} E_t \left[\int_t^T \pi_s c_s \, \mathrm{d}s + \pi_T c_T \right], \quad t \in [0, T]. \tag{9}$$

(b) *If* $E[\sup_t \pi_t |W_t|] < \infty$*, then*

$$W_t = \frac{1}{\pi_t} E_t \left[\int_t^T \pi_s c_s \, \mathrm{d}s + \pi_T c_T \right], \quad t \in [0, T]. \tag{10}$$

Proof. Suppose (W, Σ) satisfies (8). Integration by parts gives $\mathrm{d}(\pi W) = -\pi c \, \mathrm{d}t + \ldots \mathrm{d}B$. Let $\{\tau_n\}$ be an increasing sequence of stopping times converging to T almost surely, and such that the $\ldots \mathrm{d}B$ term stopped at τ_n is a martingale. Integrating the last equation from t to T, and applying the operator E_t on both sides, we find

$$\pi_t W_t = E_t \left[\int_t^{\tau_n} \pi_s c_s \, \mathrm{d}s + \pi_{\tau_n} W_{\tau_n} \right].$$

If $W \geqslant 0$, we can take the limit as $n \to \infty$ and apply Fatou's lemma to conclude (9). If $E[\sup_t \pi_t |W_t|] < \infty$, then we can apply dominated convergence to conclude (10). □

Remark 6. Conversely, if W is given by (10), then W solves BSDE (8). This can be shown by rearranging (10), and using integration by parts and a martingale representation theorem.

In our context, where W is the wealth process generated by a strategy financing the consumption plan c, the above lemma implies the state-price-density property of π:

Proposition 7. *Suppose* $\pi \in \mathcal{H}_{++}$ *follows the dynamics* (7) *for a market-price-of-risk process* η*. If* $E[\sup_t \pi_t W_t] < \infty$*, then* π *is a state price density at* c*.*

Proof. Suppose $c + x$ is a feasible consumption plan. By Lemma 5, $\pi_0 w_0 \geqslant (\pi \mid c + x)$ and $\pi_0 w_0 = (\pi \mid c)$. Therefore, $(\pi \mid x) \leqslant 0$. □

Remark 8. The necessity of condition (7) for an Itô process $\pi \in \mathcal{H}_{++}$ to be a state price density at c is shown, under some regularity assumptions, in Schroder and Skiadas (2003), where the characterization is also extended to allow for trading constraints. For example, necessity follows if $\mathcal{C} = \mathcal{H}_{++}$ and $c \in \mathcal{C}$ is continuous.

In Lemma 5, we saw that the linear term $rW + \eta'\Sigma$ in BSDE (8) corresponds to stochastic discounting in the present value formula (10). Alternatively, the two terms can be interpreted separately, with rW corresponding to temporal discounting and $\eta'\Sigma$ corresponding to a change of measure. To see how, we define, given any $\eta \in \mathcal{L}_2(\mathbb{R}^d)$, the processes ξ^η and B^η by

$$\frac{\mathrm{d}\xi_t^\eta}{\xi_t^\eta} = -\eta_t' \,\mathrm{d}B_t, \quad \xi_0^\eta = 1, \quad \text{and} \quad \mathrm{d}B_t^\eta = \mathrm{d}B_t + \eta_t \,\mathrm{d}t, \quad B_0^\eta = 0. \tag{11}$$

We recall that ξ^η is a positive supermartingale, and is a martingale if and only if $E\xi_T^\eta = 1$. Suppose $\eta \in \mathcal{L}_2(\mathbb{R}^d)$ is such that ξ^η is a martingale. In this case an equivalent-to-P probability measure P^η, with expectation operator E^η, is well defined through the change-of-measure formula $E^\eta[x] = E[\xi_T x]$ (or $\mathrm{d}P^\eta/\mathrm{d}P = \xi_T$). By Girsanov's theorem, B^η is standard Brownian motion under P^η. The linear BSDE (8) can equivalently be stated as

$$\mathrm{d}W_t = -(c_t - r_t W_t)\,\mathrm{d}t + \Sigma_t' \,\mathrm{d}B_t^\eta, \quad W_T = c_T.$$

Applying Lemma 5 and Remark 6 to this BSDE with underlying probability P^η, we conclude that, if $E^\eta[\sup_t \exp(-\int_0^t r_\tau \,\mathrm{d}\tau)|W_t|] < \infty$, then W solves BSDE (8) if and only if

$$W_t = E_t^\eta\left[\int_t^T e^{-\int_t^s r_\tau \,\mathrm{d}\tau} c_s \,\mathrm{d}s + e^{-\int_t^T r_\tau \,\mathrm{d}\tau} c_T\right]. \tag{12}$$

Equation (12) is the familiar risk-neutral-pricing version of the present-value formula (10), stating that financial wealth is equal to the present value of the future cash flow that this wealth finances. In a Markovian setting, such a present value can be computed (under some regularity) in terms of a corresponding PDE solution, sometimes referred to as the Feynman–Kac solution (see Duffie, 2001). The PDE form can be derived by writing W as a function of time and the underlying Markov state, applying Itô's lemma, and matching terms with the linear BSDE (8). This type of construction applies more generally to BSDEs, and can be used to characterize optimal portfolios, as will be outlined for a class of scale-invariant solutions in Section 5.

2.4 Expected time-additive utility and what is wrong with it

Having understood the structure of state price dynamics, which is unrelated to preferences, we turn our attention to the utility side. Our objective is to specify some utility functional structure that properly captures a notion of risk aversion, and then compute the supergradient density dynamics. Combining the latter with the state price dynamics will result in optimality conditions.

A widely used functional form of the time-zero utility function $U_0 : \mathcal{C} \to \mathbb{R}$ is

$$U_0(c) = E\left[\int_0^T e^{-\beta t} u(c_t)\,\mathrm{d}t + e^{-\beta T} v(c_T)\right], \tag{13}$$

for some $\beta \in \mathbb{R}$ and concave increasing functions $u, v : \mathbb{R}_{++} \to \mathbb{R}$. The more concave u is, the more risk-averse the utility. An advantage of this specification is that a supergradient density can be computed separately at each spot, simplifying the investor problem, at least under complete markets. For example, suppose that (13) holds with $u = v$, the derivative u' exists and maps $\mathbb{R}_{++}$ onto $\mathbb{R}_{++}$, and the optimal consumption plan c satisfies $u'(c) \in \mathcal{H}_2$. It is straightforward to check that the process $e^{-\beta t} u'(c_t)$ is a supergradient density of U_0 at c. If the market is complete ($m = d$), then there exists a unique state price density π with $\pi_0 = 1$, given by the dynamics (7) with $\eta = \sigma^{R\prime -1} \mu^R$. The optimal consumption is $c_t = u'^{-1}(e^{\beta t} k \pi_t)$, where the scalar $k > 0$ is selected so that $(\pi \mid c) = w_0$. The corresponding wealth process W is given by the present value formula (10). If $\mathrm{d}W/W = \ldots \mathrm{d}t + \sigma' \mathrm{d}B$, then we have seen that the corresponding optimal trading strategy ψ is given by (2). This is essentially the analysis of the Merton problem by Cox and Huang (1989) and Karatzas et al. (1987). Later in this chapter, we will see that much of the simplicity of the above argument is lost if markets are incomplete.

We argue that, despite its popularity, the time-additive utility specification (13) is fundamentally flawed as a representation of risk aversion, which is a good reason for investing some effort in studying recursive utility. Emphasizing the temporal aspect of consumption, we focus in the remainder of this section on preferences over consumption plans with fixed terminal consumption or bequest, and we therefore assume that (13) holds with $v = 0$. We show below that in this case the investor's preferences over deterministic choices determine, up to ordinal equivalence, the investor's entire utility function, and in particular the investor's risk aversion. On the other hand, we will see that with recursive utility two investors can have identical preference in a deterministic environment, and yet one investor can be more risk-averse than the other.

We use the following standard uniqueness result from additive representation theory. A proof can be found in Narens (1985) or Wakker (1989).

Lemma 9. *For any integer $N > 1$, and each $i \in \{1, 2\}$, suppose the function $F^i : \mathbb{R}^N_{++} \to \mathbb{R}$ has the additive structure $F^i(x_1, \ldots, x_N) = \sum_{n=1}^N f_n^i(x_n)$, $x \in \mathbb{R}^N_{++}$, where $f_n^i : \mathbb{R}_{++} \to \mathbb{R}$, $n = 1, \ldots, N$. Suppose also that F^1 and F^2 are ordinally equivalent, meaning that $F^1(x) \geqslant F^1(y)$ if and only if $F^2(x) \geqslant F^2(y)$. Then there exists $\alpha \in \mathbb{R}_{++}$ and $\beta \in \mathbb{R}^N$ such that $f_n^1 = \alpha f_n^2 + \beta_n$, $n = 1, \ldots, N$.*

Proposition 10. *For each $i \in \{1, 2\}$, suppose the utility function $U_0^i : \mathcal{C} \to \mathbb{R}$ takes the form $U_0^i(c) = E \int_0^T v^i(t, c_t)\,\mathrm{d}t$, where $v^i : [0, T] \times \mathbb{R}_{++} \to \mathbb{R}$ is continuous.*

Suppose also that, for any deterministic[5] *consumption plans* $a, b \in \mathcal{C}$, $U_0^1(a) \geqslant U_0^1(b)$ *if and only if* $U_0^2(a) \geqslant U_0^2(b)$. *Then the utility functions* U_0^1 *and* U_0^2 *are ordinally equivalent on the entire space* $\mathcal{C}$.

Proof. After replacing U_0^i with $U_0^i - U_0^i(\mathbf{1})$, we can and do assume that $v^i(t, 1) = 0$ for all t. Given any integer $N > 1$, we define the time intervals $J^n = [(n-1)T/N, nT/N)$, $n = 1, \ldots, N$, partitioning $[0, T)$. Let D_N be the set of deterministic plans of the form $\sum_{n=1}^N x_n 1_{J^n}$. Since U_0^1 and U_0^2 order the elements of D_N the same, we can apply the above lemma with $f_n^i(x) = \int_{J^n} v^i(t, x)\,\mathrm{d}t$ to conclude that, for some $\alpha \in \mathbb{R}_{++}$ and all n,

$$\int_{J^n} v^1(t, x)\,\mathrm{d}t = \alpha \int_{J^n} v^2(t, x)\,\mathrm{d}t. \tag{14}$$

Repeating the above argument with $2N$ in place of N results in the same relationship, with the same constant α, since D_N can be embedded into D_{2N}. For any $x > 0$, we can therefore take a sequence of intervals $\{J^n\colon\ n = 1, 2, \ldots\}$ containing x, whose length converges to zero and (14) holds for all n. Dividing both sides of (14) by the length of J_n and taking the limit as $n \to \infty$, we conclude that $v^1(t, x) = \alpha v^2(t, x)$. □

The limitation of additive utility in capturing risk aversion is illustrated in the following variant of what seems to be a folklore example (which I learned from Duffie and Epstein, 1992).[6]

Example 11. Suppose that $T = 100$ and $U_0(c) = E[\int_0^{100} v(t, c_t)\,\mathrm{d}t]$ for some continuous function $v : [0, 100] \times \mathbb{R}_{++} \to \mathbb{R}$. The plans a and b are defined by

$$a_t = 1 + 1,000 \times 1_{\{t>1,\ B_1>0\}} \quad \text{and}$$

$$b_t = 1 + 1,000 \times \sum_{n=1}^{99} 1_{\{1+n\geqslant t>n,\ B_n - B_{n-1}>0\}}.$$

While $Ea_t = Eb_t$ for all t, one could reasonably argue that plan b is less risky than plan a. Yet, it is straightforward to check that $U_0(a) = U_0(b)$.

[5] We call a process x *deterministic* if x_t is $\mathcal{F}_0$-measurable for every time t.

[6] In their introduction, Duffie and Epstein (1992) give another example of the limitation of additive utility that is based on the notion of preferences for the timing of resolution of uncertainty of Kreps and Porteus (1978). The notion was extended in Skiadas (1998) in terms of preferences over pairs of consumption plans and information streams (filtrations). Additivity relates to the nondependence of utility on the filtration argument.

3 Recursive utility

In this section, we motivate and define recursive utility, and we derive its basic properties. By combining a computation of the utility supergradient dynamics with last section's state price dynamics, we obtain optimality conditions under recursive utility as a FBSDE system. Finally, we introduce homothetic recursive utility and its role in uncoupling the FBSDE system of the optimality conditions. Continuous-time recursive utility was first defined and analyzed by Duffie and Epstein (1992), who imposed some special structure that is useful in modeling risk aversion. Following Lazrak and Quenez (2003), we adopt a broader definition of recursive utility as a solution to a general BSDE. In the following section, we will see that the broader definition of recursive utility allows for interesting models of risk aversion that go beyond the Duffie–Epstein specification.

3.1 Recursive utility and BSDEs

We begin with a heuristic derivation from general principles of recursive utility. The argument should also help clarify the interpretation of a BSDE as a continuous-time representation of a backward recursion on an information tree.

We consider the cardinal dynamic utility function $U : \mathcal{C} \to \mathbb{R}$ that was informally constructed in terms of preferences in Section 2.2, and we assume that $U(c)$ is an Itô process for every $c \in \mathcal{C}$. In addition to our earlier assumptions of dynamic consistency, monotonicity, and concavity, we impose the following simplifying restriction.

Irrelevance of past or unrealized consumption. For any consumption plans a and b, any time $t \leqslant T$, and any event $A \in \mathcal{F}_t$, if $a = b$ on[7] $A \times [t, T]$, then $U(a) = U(b)$ on $A \times [t, T]$.

This assumption is not an essential aspect of a recursive utility structure, but serves as a natural benchmark in an analysis whose main focus is risk aversion. Together with dynamic consistency, it implies that, for any consumption plan c and times $t < u$, the restriction of $U_t(c)$ on a time-t event A can be expressed as a function of the restriction of c on $A \times [t, u]$ and the restriction of $U_u(c)$ on A. More formally, we can show[8]:

[7] This means that the indicator of $\{(\omega, u)\colon\ \omega \in A,\ u \in [t, T],\ a(\omega, u) \neq b(\omega, u)\}$ is zero as an element of $\mathcal{H}$.

[8] **Proof.** Let $D = A \cap \{U_t(a) > U_t(b)\}$, and define the stopping times $\sigma = t1_D + T1_{\Omega\setminus D}$ and $\tau = u1_D + T1_{\Omega\setminus D}$. We define the plan a' (respectively b') to be equal to a (respectively b) on $D \times [t, T]$, and some arbitrary plan c outside $D \times [t, T]$. Since $a = a'$ on $[\sigma, T]$, we have $U(a) = U(a')$ on $[\sigma, T]$, and therefore $U_\tau(a') = U_\tau(a)$ a.s. Analogously, $U_\tau(b') = U_\tau(b)$ a.s., and therefore $U_\tau(a') = U_\tau(b')$ a.s. Since a' and b' are equal up to the stopping time τ, dynamic consistency implies $U_0(a') = U_0(b')$.

Lemma 12. *Given any times $t < u \leqslant T$ and event $A \in \mathcal{F}_t$, suppose that the consumption plans a and b are equal on $A \times [t, u]$ and $U_u(a) = U_u(b)$ on A. Then $U_t(a) = U_t(b)$ on A.*

Proceeding heuristically, we apply the above functional relationship with the time-event (t, A) corresponding to a single spot ω^t and $u = t + \mathrm{d}t$, where $\mathrm{d}t$ is an infinitesimal time-interval. Fixing any $c \in \mathcal{C}$, we let $U = U(c)$. Given the instantaneous factor decomposition

$$U_{t+\mathrm{d}t} = m_t + \Sigma_t' \,\mathrm{d}B_t, \quad \text{where}$$
$$m_t = E_t[U_{t+\mathrm{d}t}] \quad \text{and} \quad \Sigma_t^i = \mathrm{cov}_t\big[U_{t+\mathrm{d}t}, \mathrm{d}B_t^i\big], \quad i = 1, \ldots, d, \tag{15}$$

we obtain the functional restriction

$$U_t = \Phi(t, c_t, m_t, \Sigma_t), \tag{16}$$

for some (possibly state-dependent) function $\Phi : \Omega \times [0, T] \times \mathbb{R}_{++} \times \mathbb{R}^{1+d} \to (0, \infty)$ that is adapted to the underlying information structure. Utility monotonicity and concavity heuristically imply[9] that $\Phi(\omega, t, c, m, \Sigma)$ is increasing in (c, m) and concave in (c, m, Σ). Given U_{t+dt}, Equation (16), with m_t and Σ_t defined in (15), is used to compute U_t. Equation (16) is therefore a heuristic backward recursion on the information tree, which determines the entire utility process U given the terminal value U_T.

To formulate a rigorous version of the utility recursion, we assume that the function F, called an (infinitesimal) aggregator, is implicitly defined, at any state ω and time $t < T$, by

$$\mu = -F(\omega, t, c, U, \Sigma) \quad \Longleftrightarrow \quad U = \Phi(\omega, t, c, U + \mu\,\mathrm{d}t, \Sigma). \tag{17}$$

By monotonicity of Φ in the conditional mean argument, there is at most one value μ satisfying the right-hand side equation in (17), and therefore F is uniquely determined given Φ. Moreover, the monotonicity and concavity properties of Φ imply that $F(\omega, t, c, U, \Sigma)$ is increasing in c and concave in (c, U, Σ). (If Φ is strictly concave in m it also follows[10] that F is decreasing in U. We will not need to assume this condition, although it is helpful in verifying technical regularity conditions.) We use the notation $U_T = F(T, c_T)$ to

On the other hand, a' and b' are equal up to σ, $U_\sigma(a') > U_\sigma(b')$ on D, and $U_\sigma(a') = U_\sigma(b')$ on $\Omega \setminus D$. If $P(D) > 0$, then dynamic consistency would imply $U_0(a') > U_0(b')$, a contradiction. Therefore $P(D) = 0$. This shows $U_t(a) \leqslant U_t(b)$ on A. The reverse inequality is true by symmetry. □

[9] The idea is that the dependence of $\Phi(\omega, t, c, m, \Sigma)$ on (c, m, Σ) is through the pair (c, U), where $U = m + \Sigma' \,\mathrm{d}B$. One can heuristically identify (c, U) with a plan that is equal to c at spot ω^t (corresponding to (ω, t)), takes the value U on $[t+\mathrm{d}t, T]$ conditionally on spot ω^t having occurred, and it takes, say, the value one at all remaining spots. Utility monotonicity and concavity over the set of such plans translates to the corresponding properties for $\Phi(\omega, t, \cdot)$.

[10] To see that, make a plot of $\Phi(\omega, t, c, U + \mu\,\mathrm{d}t, \Sigma)$ as a function of U. The concave graph intersects the 45° line at U. As μ increases, the graph moves up and the intersection with the 45° line moves to the right.

express the dependence of terminal utility on terminal consumption (which is the identity for cardinal utility).

Assuming the Itô decomposition $\mathrm{d}U = \mu\,\mathrm{d}t + \Sigma'\,\mathrm{d}B$, and therefore $m = U + \mu\,\mathrm{d}t$, recursion (16) is equivalent to the drift restriction $\mu_t = -F(t, c_t, U_t, \Sigma_t)$, resulting in the utility dynamics

$$\mathrm{d}U_t = -F(t, c_t, U_t, \Sigma_t)\,\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t, \quad U_T = F(T, c_T). \tag{18}$$

Equation (18) is a BSDE to be solved jointly in the (adapted) process pair (U, Σ). The function $f(\omega, t, y, z) = F(\omega, t, c(\omega, t), y, z)$, is known as the BSDE *driver*. We say that the Itô process U *solves* BSDE (18) if there exists a (necessarily unique) $\Sigma \in \mathcal{L}_2(\mathbb{R}^d)$ such that (18) is satisfied. While we have motivated BSDE (18) in terms of cardinal utility, it can also be used to characterize other ordinally equivalent utility versions, as in the following example.

Example 13 (Expected discounted utility). In the above heuristic argument, suppose

$$\Phi(\omega, t, c, m, \Sigma) = u(\omega, t, c)\,\mathrm{d}t + m\exp\big(-\beta(\omega, t)\,\mathrm{d}t\big),$$
$$F(\omega, t, c, U, \Sigma) = u(\omega, t, c) - \beta(\omega, t)U, \quad t < T.$$

By Lemma 5, under a regularity assumption, the solution to BSDE (18) is

$$U_t = E_t\left[\int_t^T \exp\left(-\int_t^s \beta_\tau\,\mathrm{d}\tau\right)u(s, c_s)\,\mathrm{d}s + \exp\left(-\int_t^T \beta_\tau\,\mathrm{d}\tau\right)F(T, c_T)\right].$$

Initial BSDE existence and uniqueness results, based on the type of Lipschitz-growth assumptions on the driver familiar from SDE theory, were first obtained by Pardoux and Peng (1990) and Duffie and Epstein (1992). An improved version of the Pardoux–Peng argument can be found in El Karoui et al. (1997). These conditions are violated in our main homothetic application to follow, which includes the widely used Epstein–Zin utility [a continuous-time version of the recursive utility parametrization used in Epstein and Zin (1989)]. Existence, uniqueness and basic properties for continuous-time Epstein–Zin utility were shown in Appendix A of Schroder and Skiadas (1999). BSDE theory has been further developed by Hamadene (1996), Lepeltier and Martin (1997, 1998, 2002), Kobylanski (2000), and others (see also El Karoui and Mazliak, 1997). Moreover, the numerical solution of BSDEs has received increasing attention, with contributions by Douglas Jr. et al. (1996), Chevance (1997), Bally and Pages (2002), Ma et al. (2002), Zhang (2004), Bouchard and Touzi (2004), Bouchard and Elie (2005), Gobet et al. (2005), Lemor et al. (2006), and others. Issues of existence, uniqueness, or numerical computation will not be further addressed in this chapter.

Given the above motivation, we now formally define the utility class used in the remainder of this chapter. We assume that utility takes values in an open

interval $I_U \subseteq \mathbb{R}$, which is equal to $\mathbb{R}_{++}$ for cardinal utility. Utility processes will be assumed to be members of a linear subspace $\mathcal{U} \subseteq \mathcal{L}(I_U)$, taken as a primitive. We assume throughout that every $U \in \mathcal{U}$ is an Itô process and satisfies $E[\sup_t U_t^2] < \infty$. Below we define a *dynamic utility*, meaning that an entire utility process $U(c)$ is assigned to a plan c. Later we will verify that dynamic consistency is satisfied, and is therefore sufficient to maximize time-zero utility.

Definition 14. An (increasing in consumption and concave) *aggregator* is a progressively measurable function of the form $F : \Omega \times [0, T] \times \mathbb{R}_{++} \times I_U \times \mathbb{R}^d \to \mathbb{R}$ satisfying:

1. $F(\omega, t, c, U, \Sigma)$ is strictly increasing in c and concave in (c, U, Σ).
2. $F(\omega, T, c, U, \Sigma)$ does not depend on (U, Σ), and is therefore denoted $F(\omega, T, c)$.

The function $U : \mathcal{C} \to I_U$ is *recursive utility* with aggregator function F if, for any $c \in \mathcal{C}$, $U(c)$ solves BSDE (18) uniquely in $\mathcal{U}$. The aggregator F is *deterministic* if it does not depend on the state variable. The recursive utility U is *state-independent* if the corresponding aggregator F is deterministic, and for any deterministic plan c, $U(c)$ is the unique deterministic element of $\mathcal{U}$ solving the ODE $\mathrm{d}U_t = -F(t, c_t, U_t, 0)\,\mathrm{d}t$, $U_T = F(T, c_T)$.

Remark 15 (Aggregator and beliefs). Suppose that U is recursive utility with aggregator F, and the process $b \in \mathcal{L}(\mathbb{R}^d)$ is (for simplicity) bounded. Consider the modified aggregator

$$F^b(\omega, t, c, U, \Sigma) = F(\omega, t, c, U, \Sigma) + b(\omega, t)'\Sigma.$$

Recalling the notation in (11), we note that

$$\mathrm{d}U_t = -F^b(t, c_t, U_t, \Sigma_t)\,\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t^b \quad \text{and}$$
$$\mathrm{d}R_t = \left(\mu_t^R - \sigma_t^{R\prime} b_t\right)\mathrm{d}t + \sigma_t^{R\prime}\,\mathrm{d}B_t^b.$$

Since B^b is Brownian motion under the probability P^b (where $\mathrm{d}P^b/\mathrm{d}P = \xi_T^b$), an investor with prior P^b still assesses the same risk profile σ^R, but believes that the instantaneous expected returns are $\mu^R - \sigma^{R\prime} b$. A solution method to the investor's problem for $b = 0$ extends to any value of b after the formal substitution $(P, B, \mu^R) \to (P^b, B^b, \mu^R - \sigma^{R\prime} b)$.

3.2 Some basic properties of recursive utility

In this subsection we derive, under regularity assumptions, some basic properties of recursive utility. We first verify dynamic consistency, monotonicity, concavity, and the irrelevance of past or unrealized alternatives. We then discuss comparative risk aversion, and finally we compute the dynamics of a utility supergradient density.

The following notation will be useful. For any function of the form $f:\Omega \times [0,T] \times S \to \mathbb{R}$, where S is a convex subset of some Euclidean space X, we define the superdifferential notation:

$$\partial f(\omega,t,s) = \big\{\delta \in X\colon\ f(\omega,t,s+h) \leqslant f(\omega,t,s) + \delta' h \quad \text{for all } h \in X \text{ such that } s+h \in S\big\}.$$

Given any processes $d \in \mathcal{L}(X)$ and $x \in \mathcal{L}(S)$, the notation $d \in \partial f(x)$ means that the indicator function of the set of all (ω,t) such that $d(\omega,t) \notin \partial f(\omega,t,x(\omega,t))$ is the zero element of $\mathcal{H}$. Given any $d=(a,b) \in \mathcal{L}_1(\mathbb{R}) \times \mathcal{L}_2(\mathbb{R}^d)$, we let $\mathcal{E}(d)$ or $\mathcal{E}(a,b)$ denote the stochastic exponential with dynamics

$$\frac{\mathrm{d}\mathcal{E}_t(a,b)}{\mathcal{E}_t(a,b)} = a_t\,\mathrm{d}t + b_t'\,\mathrm{d}B_t, \quad \mathcal{E}_0(a,b) = 1.$$

The key to deriving properties of recursive utility is the so-called comparison principle, stated below in terms of the (progressively measurable) driver functions $f^i:\Omega \times [0,T] \times I_U \times \mathbb{R}^d \to \mathbb{R}$, $i=0,1$.

Condition 16 (Comparison principle). For each $i \in \{0,1\}$, suppose $(U^i,\Sigma^i) \in \mathcal{U} \times \mathcal{L}(\mathbb{R}^d)$ solves the BSDE

$$\mathrm{d}U_t^i = -f^i\big(t,U_t^i,\Sigma_t^i\big)\,\mathrm{d}t + \Sigma_t^{i\prime}\,\mathrm{d}B, \quad t \in [0,T],\ U_T^i \text{ given}.$$

Given stopping times σ,τ such that $\sigma \leqslant \tau$ a.s., suppose also that[11]

$$f^0\big(t,U^1,\Sigma^1\big) \leqslant f^1\big(t,U^1,\Sigma^1\big) \quad \text{on } [\sigma,\tau] \quad \text{and} \quad U_\tau^0 \leqslant U_\tau^1 \quad \text{a.s.}$$

Then $U_\sigma^0 \leqslant U_\sigma^1$ a.s. Assuming further that $P[U_\tau^0 < U_\tau^1] > 0$, then $P[U_\sigma^0 < U_\sigma^1] > 0$.

A comparison lemma (or stochastic Gronwall–Bellman inequality in the language of Duffie and Epstein) imposes sufficient regularity restrictions for the comparison principle to hold. Various comparison lemmas are given in the BSDE literature referenced earlier. We show below an apparently new version whose applicability relies on our concavity assumption.

Lemma 17 *(Comparison lemma). The comparison principle* (*Condition* 16) *holds if there exists some* $d \in \mathcal{L}_1(\mathbb{R}) \times \mathcal{L}_2(\mathbb{R}^d)$ *such that* $d \in \partial f^0(U^1,\Sigma^1)$ *a.e. and* $E[\sup_t \mathcal{E}_t(d)^2] < \infty$.

[11] For processes x,y, we say that $x \leqslant y$ on $[\sigma,\tau]$ if the indicator function of the set of all (ω,t) such that $x(\omega,t) > y(\omega,t)$ and $\sigma(\omega) \leqslant t \leqslant \tau(\omega)$ is zero as an element of $\mathcal{H}$.

Proof. Notationally suppressing the arguments (ω, t), we define the processes $y = U^1 - U^0 \in \mathcal{U}$, $z = \Sigma^1 - \Sigma^0$, and $p = f^1(U^1, \Sigma^1) - f^0(U^1, \Sigma^1)$, and we note that

$$\mathrm{d}y = -\big(f^0\big(U^1, \Sigma^1\big) - f^0\big(U^0, \Sigma^0\big) + p\big)\,\mathrm{d}t + z'\,\mathrm{d}B.$$

Let $d = (\mathrm{d}_U, \mathrm{d}_\Sigma)$ be as in the lemma's statement, and define the process $q = f^0(U^1, \Sigma^1) - f^0(U^0, \Sigma^0) - (\mathrm{d}_U y + \mathrm{d}'_\Sigma z)$. Then the above dynamics for y can be restated as

$$\mathrm{d}y = -(\delta + \mathrm{d}_U y + \mathrm{d}'_\Sigma z)\,\mathrm{d}t + z'\,\mathrm{d}B,$$

where $\delta = p + q$. Our assumptions imply that $\delta \geqslant 0$ on $[\sigma, \tau]$ and $y_\tau \geqslant 0$ a.s. Arguing as in the proof of Lemma 5, we can select a sequence of stopping times $\{\tau_n\}$ such that $\tau_n \uparrow \tau$ a.s. and

$$\mathcal{E}_\sigma(d) y_\sigma = E_\sigma\left[\int_\sigma^{\tau_n} \mathcal{E}_t(d)\delta_t\,\mathrm{d}t + \mathcal{E}_{\tau_n}(d) y_{\tau_n}\right] \geqslant E_\sigma[\mathcal{E}_{\tau_n} y_{\tau_n}] \quad \text{a.s.}$$

We recall that, by assumption, $y \in \mathcal{U}$ implies $E[\sup_t y_t^2] < \infty$. Letting $n \to \infty$, and using dominated convergence, it follows that $y_\sigma \geqslant 0$ a.s. □

Next we introduce a regularity condition that will allow us to apply the comparison lemma to derive the utility properties we are interested in. The reader who wants to skip technicalities can read "regular" as meaning "we can apply the comparison principle where we have to."

Given any aggregator F and $c \in \mathcal{C}$, we use the notation $F^c(\omega, t, y, z) = F(\omega, t, c(\omega, t), y, z)$. We call an aggregator F *regular* if, given any $(c, U) \in \mathcal{C} \times \mathcal{U}$ with $\mathrm{d}U = \ldots \mathrm{d}t + \Sigma'\,\mathrm{d}B$, there exists $d \in \mathcal{L}_1(\mathbb{R}) \times \mathcal{L}_2(\mathbb{R}^d)$ such that $d \in \partial F^c(U, \Sigma)$ a.e. and $E[\sup_t \mathcal{E}_t(d)^2] < \infty$. For example, suppose F is differentiable and $F_U \leqslant 0$ [which follows from the strict concavity of Φ in Equation (16)]. In this case, regularity of F becomes an integrability restriction on F_Σ, which is satisfied if F_Σ is bounded. Boundedness of F_Σ is usually too strong an assumption, however, and confirming regularity is more challenging.

Proposition 18. *A recursive utility with a regular aggregator is dynamically consistent, monotonically increasing, concave, and satisfies the irrelevance of past or unrealized alternatives condition.*

Proof. Suppose U is recursive utility with aggregator F, and $c^0, c^1 \in \mathcal{C}$. We use the notation $U^i = U(c^i)$ and $\mathrm{d}U^i = \ldots \mathrm{d}t + \Sigma^{i\prime}\,\mathrm{d}B$. To show monotonicity, suppose $c^1 \geqslant c^0$. The comparison lemma with $f^i = F^{c^i}$ implies $U^1 \geqslant U^0$. To show concavity, we fix any $\alpha \in (0, 1)$ and define the notation $x^\alpha = (1-\alpha)x^0 + \alpha x^1$. Notationally suppressing the arguments (ω, t), we define the process $p = F(c^\alpha, U^\alpha, \Sigma^\alpha) - (1-\alpha)F(c^0, U^0, \Sigma^0) - \alpha F(c^1, U^1, \Sigma^1)$, and note that

$$\mathrm{d}U^\alpha = -\big(F\big(c^\alpha, U^\alpha, \Sigma^\alpha\big) - p\big)\,\mathrm{d}t + \Sigma^\alpha\,\mathrm{d}B, \quad U_T^\alpha = F\big(T, c_T^\alpha\big) - p_T.$$

The concavity assumption on F implies that $p \geqslant 0$. Applying Lemma 17 with $f^1 = F^{c^\alpha}$ and $f^0 = F^{c^\alpha} - p$, we conclude that $U(c^\alpha) \geqslant U^\alpha$, confirming concavity. The remaining claims are left as an exercises in the application of Lemma 17. □

A state-independent recursive utility with aggregator F ranks deterministic plans in a way determined by the function $(t, c, U) \mapsto F(t, c, U, 0)$, while the dependence of F on Σ can be used to adjust risk aversion without affecting the utility of deterministic plans. The formal statement of this property is based on the following partial order of utility functions.

Comparative risk aversion. A utility function $U_0^1 : \mathcal{C} \to \mathbb{R}$ is *more risk-averse* than a utility function $U_0^2 : \mathcal{C} \to \mathbb{R}$ if

- For any deterministic plans $a, b \in \mathcal{C}$,
$$U_0^1(a) \geqslant U_0^1(b) \iff U_0^2(a) \geqslant U_0^2(b).$$
- For any $c \in \mathcal{C}$ and deterministic $\bar{c} \in \mathcal{C}$,
$$U_0^2(\bar{c}) \geqslant U_0^2(c) \implies U_0^1(\bar{c}) \geqslant U_0^1(c).$$

Remark 19. If U_0^1 and U_0^2 are cardinal utilities, then U_0^1 is more risk-averse than U_0^2 if and only if $U_0^1(c) = U_0^2(c)$ for every deterministic plan c, and $U_0^1(c) \leqslant U_0^2(c)$ for every plan c.

Proposition 20. *Suppose that, for $i \in \{1, 2\}$, U^i is state-independent recursive utility with aggregator F^i, and F^1 is regular. If $F^1(t, c, U, 0) = F^2(t, c, U, 0)$ and $F^1(t, c, U, \Sigma) \leqslant F^2(t, c, U, \Sigma)$ for all (t, c, U, Σ), then U_0^1 is more risk-averse than U_0^2.*

Proof. By definition, $F^1(t, c, U, 0) = F^2(t, c, U, 0)$ implies that $U^1(c) = U^2(c)$ for every deterministic plan c. The proof is completed using Lemma 17. □

Finally, we derive a utility supergradient density expression for recursive utility, which will be key in establishing optimality conditions.

Proposition 21. *Suppose U is recursive utility with aggregator F such that $F_c, F_U \in \mathcal{L}_1(\mathbb{R})$ and $F_\Sigma \in \mathcal{L}_2(\mathbb{R}^d)$ satisfy*

$$(F_c, F_U, F_\Sigma) \in \partial F(c, U, \Sigma) \tag{19}$$

and $E[\sup_t \mathcal{E}_t(F_U, F_\Sigma)^2] < \infty$. Let the process π be defined by

$$\pi = \mathcal{E}(F_U, F_\Sigma) F_c.$$

Provided it belongs to $\mathcal{H}$, the process π is a supergradient density of U_0 at c.

Proof. Assuming $c + x \in \mathcal{C}$, we define $\delta = U(c + x) - U(c)$, $\Delta = \Sigma(c + x) - \Sigma(c)$, and $p = F(c, U, \Sigma) + F_c x + F_U \delta + F'_{\Sigma} \Delta - F(c + x, U + \delta, \Sigma + \Delta) \geqslant 0$, where the last inequality follows from the assumed condition (19). The BSDEs for $U(c + h)$ and $U(c)$ imply the linear BSDE

$$\mathrm{d}\delta = -(F_c x + F_U \delta + F'_{\Sigma} \Delta - p)\,\mathrm{d}t + \Delta'\,\mathrm{d}B, \quad \delta_T = F_c(T) x_T - p_T.$$

An exercise, using Lemma 17 and Lemma 5, shows that $\delta_0 \leqslant (\pi \mid x)$. □

3.3 *Optimality under recursive utility*

Proposition 3 verifies the optimality of a feasible consumption plan c based on the existence of a process that is both a utility supergradient density at c and a state price density at c. Specializing this argument to recursive utility, in this subsection, we apply Itô's lemma to the supergradient density expression of Proposition 21, and we use the state price dynamics of Proposition 7 to derive sufficient optimality conditions for recursive utility as a FBSDE system.

We fix a reference recursive utility $U : \mathcal{C} \to I_U$ with aggregator F, relative to which optimality is defined. By definition, $F(\omega, t, \cdot)$ is concave but not necessarily differentiable. In the following section, we will see that nonsmoothness of $F(\omega, t, c, U, \Sigma)$ in (U, Σ) is useful in modeling first-order risk aversion. On the other hand, we will have no use for nonsmoothness of F in the consumption argument, and we therefore assume the existence of the corresponding partial derivative F_c. In addition, we finesse the issue of a consumption nonnegativity constraint by the usual trick of making marginal utility go to infinity near zero. Finally, we assume that marginal utility converges to zero as consumption goes to infinity. These assumptions and some associated notation are summarized below, and are adopted for the remainder of this chapter's main part.

Regularity assumptions and notation. The partial derivative F_c exists everywhere, and the function $F_c(\omega, t, \cdot, U, \Sigma)$ is strictly decreasing and maps $(0, \infty)$ onto $(0, \infty)$, for any (ω, t, U, Σ). The function $\mathcal{I} : \Omega \times [0, T] \times (0, \infty) \times I_U \times \mathbb{R}^d \to (0, \infty)$ is therefore well-defined implicitly by

$$F_c\big(\omega, t, \mathcal{I}(\omega, t, \lambda, U, \Sigma), U, \Sigma\big) = \lambda, \quad \lambda \in (0, \infty).$$

The superdifferential of F with respect to (U, Σ) is defined by

$$\partial_{U,\Sigma} F(\omega, t, c, U, \Sigma) = \big\{(a, b) \in \mathbb{R} \times \mathbb{R}^d \colon \big(F_c(\omega, t, c, U, \Sigma), a, b\big) \in \partial F(\omega, t, c, U, \Sigma)\big\}.$$

We fix a reference strategy (ρ, ψ), generating the wealth process W, and financing the consumption plan $c = \rho W$. To formulate sufficient conditions for the optimality of c, we define the strictly positive process

$$\lambda_t = F_c(t, c_t, U_t, \Sigma_t). \tag{20}$$

The last equation is equivalent to

$$c_t = \mathcal{I}(t, \lambda_t, U_t, \Sigma_t). \tag{21}$$

By the usual envelope-type argument of microeconomics, if c is optimal then λ_t represents the shadow price of the time-t wealth constraint. Although we will not need to formalize this interpretation, it will be helpful to keep in mind that λ is a shadow-price-of-wealth process. The dynamics of λ are denoted

$$\frac{\mathrm{d}\lambda_t}{\lambda_t} = \mu_t^\lambda \,\mathrm{d}t + \sigma_t^{\lambda\prime} \,\mathrm{d}B_t. \tag{22}$$

We know that (under regularity assumptions) $\pi = \mathcal{E}\lambda$ is a supergradient density at c, where $\mathcal{E} = \mathcal{E}(F_U, F_\Sigma)$ is computed as in Proposition 21. Integration by parts gives

$$\frac{\mathrm{d}\pi}{\pi} = \big(F_U + \mu^\lambda + \sigma^{\lambda\prime} F_\Sigma\big)\,\mathrm{d}t + \big(F_\Sigma + \sigma^\lambda\big)'\,\mathrm{d}B.$$

To ensure that π is also a state-price density at c, we match terms with the dynamics of Proposition 7, resulting in the restrictions:

$$r = -\big(F_U + \mu^\lambda + \sigma^{\lambda\prime} F_\Sigma\big), \quad \eta = -\big(F_\Sigma + \sigma^\lambda\big), \quad \mu^R = \sigma^{R\prime}\eta. \tag{23}$$

We round up the optimality conditions by combining the above restrictions with the utility and wealth dynamics, as well as Equations (21) and (22):

Condition 22 (Optimality conditions for recursive utility). The trading strategy ψ and the processes $(U, \Sigma, \lambda, \sigma^\lambda, W) \in \mathcal{U} \times \mathcal{L}_2(\mathbb{R}^d) \times \mathcal{L}(\mathbb{R}_{++}) \times \mathcal{L}_2(\mathbb{R}^d) \times \mathcal{L}(\mathbb{R}_{++})$ solve

$$\begin{aligned}
&\mathrm{d}U = -F\big(\mathcal{I}(\lambda, U, \Sigma), U, \Sigma\big)\,\mathrm{d}t + \Sigma'\,\mathrm{d}B, \quad U_T = F(T, W_T),\\
&\frac{\mathrm{d}\lambda}{\lambda} = -\big(r + F_U + \sigma^{\lambda\prime} F_\Sigma\big)\,\mathrm{d}t + \sigma^{\lambda\prime}\,\mathrm{d}B, \quad \lambda_T = F_c(T, W_T),\\
&\mathrm{d}W = \big(W\big(r + \psi'\mu^R\big) - \mathcal{I}(\lambda, U, \Sigma)\big)\,\mathrm{d}t + W\psi'\sigma^{R\prime}\,\mathrm{d}B, \quad W_0 = w_0,\\
&\mu^R + \sigma^{R\prime}\big(F_\Sigma + \sigma^\lambda\big) = 0, \quad (F_U, F_\Sigma) \in (\partial_{U,\Sigma} F)\big(\mathcal{I}(t, \lambda, U, \Sigma), U, \Sigma\big).
\end{aligned}$$

Proposition 23. *Suppose Condition 22 holds, and let $c_t = \mathcal{I}(t, \lambda_t, U_t, \Sigma_t)$ and $\rho_t = c_t/W_t$. If $c \in \mathcal{C}$, $\pi = \mathcal{E}(F_U, F_\Sigma)\lambda \in \mathcal{H}$, and $E[\sup_t \pi_t W_t] < \infty$, then the strategy (ρ, ψ) is optimal, it generates the wealth process W, and it finances the consumption plan c, whose utility process is U.*

Proof. The dynamics of W can be used to verify that (ρ, ψ) finances c with wealth process W. By Proposition 21, π is a utility supergradient density at c. By Proposition 7, π is also a state price density at c. By Proposition 3, c is optimal. The dynamics of U show that $U = U(c)$. □

Remark 24. In Schroder and Skiadas (2003) the above conditions are extended to include convex trading constraints, and a necessity argument is given for a smooth aggregator under some regularity assumptions. The case of no intermediate or no terminal consumption is essentially the same as above, omitting the appropriate consumption arguments in the formulation.

Condition 22 is a FBSDE system. The wealth dynamics are computed recursively forward in time, starting with $W_0 = w_0$, and the dynamics of (U, λ) are computed recursively backward on the information tree, starting with their terminal values. The forward and backward components are coupled. In the following subsection we will introduce scale-invariance as a way of uncoupling this FBSDE system. In a Markovian setting, a PDE version of the FBSDE system can be obtained as in Ma et al. (1994). The construction is outlined in Schroder and Skiadas (2003), as well as later in this chapter for a more special class of homothetic recursive utilities.

3.4 Homothetic recursive utility

The utility function $U_0 : \mathcal{C} \to \mathbb{R}$ is *homothetic* (or *scale-invariant*) if for any $c^1, c^2 \in \mathcal{C}$,

$$U_0\big(c^1\big) = U_0\big(c^2\big) \quad \text{implies} \quad U_0\big(kc^1\big) = U_0\big(kc^2\big) \quad \text{for all } k \in (0, \infty).$$

If U_0 is homothetic and cardinal, then[12] it is homogeneous (of degree one). For recursive utility, with $I_U = (0, \infty)$, homogeneity of U_0 is implied by (and is essentially equivalent to) homogeneity of the aggregator with respect to the utility argument; that is, an aggregator of the form

$$F(\omega, t, c, U, \Sigma) = UG\left(\omega, t, \frac{c}{U}, \frac{\Sigma}{U}\right), \quad F(\omega, T, c) = c, \tag{24}$$

for some function $G : \Omega \times [0, T] \times (0, \infty) \times \mathbb{R}^d \to \mathbb{R}$ that we call a *proportional aggregator*.

Assuming the above aggregator form, suppose that (ρ, ψ) is an optimal strategy with corresponding wealth process W and utility process U. Recalling the interpretation of the process λ as the sensitivity of the optimal time-t utility value on time-t wealth, the homogeneity of the utility function implies that

$$U = \lambda W. \tag{25}$$

The intuition behind this relationship is straightforward. Suppose that at some spot ω^t the investor with unit wealth finds a consumption plan $\bar{c}$ optimal, resulting in the spot-ω^t optimal utility value $\lambda[\omega^t]$. If the same investor's wealth

[12] **Proof.** For any $c \in C$ and $k \in (0, \infty)$, $U_0(c) = U_0(U_0(c)\mathbf{1})$ implies $U_0(kc) = U_0(kU_0(c)\mathbf{1}) = kU_0(c)$. □

at ω^t were instead $W[\omega^t]$, then, by the homogeneity of the utility function and the budget equation, the investor would find the consumption plan $W[\omega^t]\bar{c}$ optimal at ω^t, resulting in the optimal utility value $U[\omega^t] = \lambda[\omega^t]W[\omega^t]$. In other words, the optimization problem at every spot is a scaled version of the unit-wealth version of the problem.

Equation (25) allows us to reduce the optimality conditions to a single BSDE for λ, whose general form can be found in Schroder and Skiadas (2003). Rather than dealing with the general case here, we will instead consider, in Section 5, optimality under special proportional aggregator functional forms that are motivated by the models of risk aversion of the following section.

We close this section with an example of a proportional aggregator specification, under which the optimal consumption strategy is a given process, for any investment opportunity set.

Example 25 (A robustly optimal consumption strategy). Let the aggregator F be given by Equation (24) for a proportional aggregator of the form

$$G(t, x, \sigma) = \beta(t)\log(x) + H(t, \sigma), \quad t < T, \tag{26}$$

where β is any strictly positive and (for simplicity) bounded process. While the optimal trading strategy depends on the specification of H and the investment opportunity set, the optimal consumption strategy is independent of both, and is simply equal to β. Ignoring technical details, we assume the sufficiency and necessity of the optimality conditions and the existence of an optimum. To see the essential part of the argument, suppose (ρ, ψ) is an optimal strategy, with corresponding wealth process W, consumption plan $c = \rho W$, utility process $U = \ldots \mathrm{d}t + \Sigma' \mathrm{d}B$, and shadow-price-of-wealth process λ. Using Equation (25), we observe that

$$\lambda_t = F_c(t, c, U, \Sigma) = \beta_t \frac{U_t}{c_t} = \beta_t \frac{W_t}{c_t}\frac{U_t}{W_t} = \frac{\beta_t}{\rho_t}\lambda_t.$$

Therefore, $\rho = \beta$, independently of the investment opportunity set. [In Schroder and Skiadas (2003) it is shown that this conclusion is valid even under trading constraints.]

4 Modeling risk aversion

This section formulates some concrete representations of possibly source-dependent second- or first-order risk aversion in the context of recursive utility. These representations will be used in the following section to derive optimal trading strategy formulas that help clarify the relationship between risk aversion and optimal portfolio allocations.

4.1 Conditional certainty equivalents

Let us recall the essential intuition of a recursive utility formulation, captured by Lemma 12. We fix a consumption plan c with cardinal utility process $U = U(c)$. For any time-t spot ω^t, the corresponding utility value $U_t[\omega^t]$ can be computed as a function of ω^t, the immediate consumption $c[\omega^t]\,\mathrm{d}t$, and the restriction of the random variable $U_{t+\mathrm{d}t}$ to the spot ω^t, which we denote $U_{t+\mathrm{d}t}[\omega^t]$.

In this section, we assume that the functional dependence of U_t on $U_{t+\mathrm{d}t}$ enters through the conditional certainty equivalent $\nu_t(U_{t+\mathrm{d}t})$, an $\mathcal{F}_t$-measurable random variable such that $\nu_t(U_{t+\mathrm{d}t})[\omega^t]$ depends on $U_{t+\mathrm{d}t}$ only through its restriction $U_{t+\mathrm{d}t}[\omega^t]$, and is the identity if $U_{t+\mathrm{d}t}[\omega^t]$ is constant. The value of $\nu_t(U_{t+\mathrm{d}t})$ is a conditional certainty equivalent in the sense that, conditionally on the spot ω^t and immediate consumption $c[\omega^t]\,\mathrm{d}t$, the investor is indifferent between the continuation of the plan c and a constant consumption rate of $\nu_t(U_{t+\mathrm{d}t})[\omega^t]$ over the entire remaining period $[t + \mathrm{d}t, T]$. Under this assumption, we can write the heuristic recursion for the utility process $U = U(c)$ as

$$U_t = \phi\big(t, \mathrm{d}t, c_t, \nu_t(U_{t+\mathrm{d}t})\big), \tag{27}$$

where ϕ can be spot-dependent. The dependence of ϕ on the recursion interval $\mathrm{d}t$ is important in the approximation argument that follows. We further assume that ϕ has continuous partial derivatives $\phi_{\mathrm{d}t}$, ϕ_c, and ϕ_U. Since preferences are increasing, ϕ_c and ϕ_U are strictly positive.

In the following three subsections we derive the functional form of the aggregator F for various specifications of the certainty equivalent ν. In each case, the conditional certainty equivalent has a local representation in terms of the utility dynamics $\mathrm{d}U_t = \mu_t\,\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t$ that takes the form

$$\nu_t(U_{t+\mathrm{d}t}) = U_t + \mu_t\,\mathrm{d}t - \mathcal{A}(t, U_t, \Sigma_t)\,\mathrm{d}t, \tag{28}$$

where $\mathcal{A}(t, U, 0) = 0$. The function $\mathcal{A}$ represents the risk aversion implicit in ν. Recalling Equations (16) and (27), and using a first-order Taylor expansion of ϕ, we obtain

$$\begin{aligned} U_t &= \Phi(t, c_t, U_t + \mu_t\,\mathrm{d}t, \Sigma_t) \\ &= \phi\big(t, \mathrm{d}t, c_t, U_t + \big(\mu_t - \mathcal{A}(t, U_t, \Sigma_t)\big)\,\mathrm{d}t\big) \\ &= U_t + \big[\phi_{\mathrm{d}t}(t, 0, c_t, U_t) + \phi_U(t, 0, c_t, U_t)\big(\mu_t - \mathcal{A}(t, U_t, \Sigma_t)\big)\big]\,\mathrm{d}t. \end{aligned}$$

Using the definition of F in terms of Φ in (17) and the last equation, we obtain the aggregator functional form

$$F(\omega, t, c, U, \Sigma) = f(\omega, t, c, U) - \mathcal{A}(\omega, t, U, \Sigma), \tag{29}$$

where $f(\omega, t, c, U) = \phi_{\mathrm{d}t}(\omega, t, 0, c, U)/\phi_U(\omega, t, 0, c, U)$.

Suppose now that ϕ and $\mathcal{A}$ are state independent, and therefore f and F are also state independent. If the plan c is deterministic, then $\Sigma = 0$ and [since

$\mathcal{A}(t, U, 0) = 0]$ utility can be computed in terms of the aggregator section $F(t, c, U, 0) = f(t, c, U)$. The function f (or ϕ) therefore determines the investor's preferences over deterministic choices. By Proposition 20, given f, the larger $\mathcal{A}$ is the more risk-averse the investor. This hierarchical separation of preferences toward deterministic choices and risk aversion can also be seen directly in the recursive form (27). If c is deterministic, then so is U, and therefore $\nu_t(U_{t+\mathrm{d}t}) = U_{t+\mathrm{d}t}$. This shows that utility over deterministic plans is determined by ϕ. Given ϕ, increasing $\mathcal{A}$ decreases the conditional certainty equivalent value and therefore U_t, resulting in more risk-averse utility.

The key behavioral restriction introduced by the assumption of the recursive form (27) is that, given the agent's preferences over deterministic choices, the agent's risk aversion at a spot ω^t, represented by $\mathcal{A}(U, \Sigma)[\omega^t]$, does not depend on the amount $c[\omega^t]\,\mathrm{d}t$ consumed at time t. This separation of current consumption and risk aversion is reflected in the separable representation (29).

Homothetic utility with an aggregator of the form (29) is obtained by further imposing the functional restriction (24). In this case, the proportional aggregator G takes the functional form

$$G(\omega, t, x, \sigma) = g(\omega, t, x) - \mathcal{R}(\omega, t, \sigma), \tag{30}$$

where $g(\omega, t, x) = f(\omega, t, x, 1)$ and $\mathcal{R}(\omega, t, \sigma) = \mathcal{A}(\omega, t, 1, \sigma)$.

4.2 The Duffie–Epstein limit of Kreps–Porteus utility

The first specialization of the aggregator form (29) we consider results from the continuous-time formulation of Kreps and Porteus (1978) utility due to[13] Duffie and Epstein (1992). In this formulation, the conditional certainty equivalent ν_t is defined by

$$u\big(\nu_t(U_{t+\mathrm{d}t})\big) = E_t\big[u(U_{t+\mathrm{d}t})\big], \tag{31}$$

for some strictly increasing, concave, twice continuously differentiable function $u : I_U \to \mathbb{R}$. We denote the corresponding coefficient of absolute risk aversion by

$$a(U) = -\frac{u''(U)}{u'(U)}.$$

In the current context, the classic Arrow (1965, 1970) and Pratt (1964) approximation of expected utility for small risks can be expressed through Itô's

[13] In fact, Duffie–Epstein utilities are obtained as the continuous-time limit of a broader class of discrete-time utilities than the Kreps–Porteus class, since the investor's certainty equivalent over continuation utility need only be von Neumann–Morgenstern in an approximate local sense. It is sufficient for our purposes, however, to think of Duffie–Epstein utility as (sufficiently smooth) continuous-time Kreps–Porteus utility.

lemma as

$$u(U_{t+\mathrm{d}t}) = u(U_t) + u'(U_t)\,\mathrm{d}U_t + \frac{1}{2}u''(U_t)(\mathrm{d}U_t)^2.$$

Given the utility process Itô decomposition $\mathrm{d}U_t = \mu_t\,\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t$, the above results in

$$E_t\big[u(U_{t+\mathrm{d}t})\big] = u(U_t) + u'(U_t)\mu_t\,\mathrm{d}t + \frac{1}{2}u''(U_t)\Sigma_t'\Sigma_t\,\mathrm{d}t.$$

On the other hand, a first-order Taylor expansion gives

$$u\big(\nu_t(U_{t+\mathrm{d}t})\big) = u(U_t) + u'(U_t)\big(\nu_t(U_{t+\mathrm{d}t}) - U_t\big).$$

Combining the last two equations with the certainty equivalent definition (31) results in the certainty-equivalent expression (28) with the quadratic risk-aversion component

$$\mathcal{A}(\omega, t, U, \Sigma) = \frac{1}{2}a(U)\Sigma'\Sigma.$$

The corresponding aggregator (29) takes *Duffie–Epstein* form:

$$F(\omega, t, c, U, \Sigma) = f(\omega, t, c, U) - \frac{1}{2}a(U)\Sigma'\Sigma. \tag{32}$$

We refer to Duffie and Epstein (1992) for further analysis of this utility form. For example, they show that there is always an ordinally equivalent utility version with the same recursive representation but $a = 0$. The latter restriction can be analytically helpful, but minimizes the usefulness of the hierarchical separation of choice over deterministic plans and risk aversion of Proposition 20. If $a = 0$ and $F = f$ is linear in U, as in Example 13, then one obtains time-additive expected discounted utility.

Homothetic Duffie–Epstein utility is obtained if the aggregator takes the homogeneous form (24), for a proportional aggregator of the form (30) with $\mathcal{R}(\omega, t, \sigma) = (\gamma/2)\sigma'\sigma$ for some $\gamma \in \mathbb{R}_+$. In this case, the BSDE specifying the utility process $U = U(c)$ is

$$\frac{\mathrm{d}U_t}{U_t} = -\left(g\left(t, \frac{c_t}{U_t}\right) - \frac{\gamma}{2}\sigma_t'\sigma_t\right)\mathrm{d}t + \sigma_t'\,\mathrm{d}B_t, \quad U_T = c_T. \tag{33}$$

The coefficient γ can be obtained from the certainty equivalent (31) with

$$u(U) = \frac{U^{1-\gamma} - 1}{1-\gamma}, \tag{34}$$

in which case $\gamma = a(U)U$ is the coefficient of relative risk aversion of the von Neumann–Morgenstern utility u. Here and below, we interpret the function (34) with $\gamma = 1$ by taking the limit as $\gamma \to 1$, resulting in $u(U) = \log U$. Assuming it is state-independent, the function g entirely determines the agent's preferences over deterministic consumption plans. Given g, increasing the value of γ makes the agent more risk averse.

Example 26. The continuous-time version of Epstein–Zin utility is contained in the specification

$$g(\omega, t, x) = \alpha + \beta \frac{x^{1-\delta} - 1}{1-\delta},$$
$$\text{where } \alpha \in \mathbb{R},\ \beta \in (0, \infty),\ \delta \in [0, \infty). \tag{35}$$

Within this class, the utility is additive if and only if $\gamma = \delta$, a condition that ties the relative risk-aversion coefficient γ to a parameter that is determined entirely by the agent's preferences over deterministic plans. Assuming $\gamma = \delta$, let $b = \beta - (1-\gamma)\alpha$, and consider the ordinally equivalent utility process

$$V_t(c) = \frac{1}{\beta} \frac{U_t(c)^{1-\gamma} - 1}{1-\gamma} - \frac{\alpha}{\beta} \int_t^T e^{-b(s-t)}\, \mathrm{d}s.$$

An exercise using Itô's lemma shows that[14]

$$V_t(c) = E_t\left[\int_t^T e^{-b(s-t)} \frac{c_s^{1-\gamma} - 1}{1-\gamma}\, \mathrm{d}s + \frac{1}{\beta} e^{-b(T-t)} \frac{c_T^{1-\gamma} - 1}{1-\gamma} \right]. \tag{36}$$

4.3 Source-dependent risk aversion

As noted in the Introduction, it is of interest to consider risk aversion that can depend on the source of risk, for example, as an expression of aversion to ambiguity associated with a given source of risk. With a version of Proposition 20, Lazrak and Quenez (2003) made the important observation that the functional dependence of a general aggregator $F(t, c, U, \Sigma)$ on Σ allows the modeling of risk-aversion that varies with the direction of risk. Since Σ represents loadings to instantaneous linear factors, such directional risk aversion can be interpreted as source-dependent risk aversion. In this section and the following one, we motivate some special functional aggregator forms representing source-dependent risk aversion that were introduced in Schroder and Skiadas (2003) (for the homothetic case).

We begin with a simple extension of Duffie–Epstein utility that allows for source-dependent risk aversion, where each Brownian motion is viewed as a separate source of risk. In the Duffie–Epstein formulation, the certainty equivalent (31) is applied to the continuation utility, $U_{t+\mathrm{d}t} = U_t + \mu_t\, \mathrm{d}t + \Sigma_t'\, \mathrm{d}B_t$, which aggregates all sources of risk. Here we assume that the investor perceives and worries about the individual risk terms $\Sigma_t^1\, \mathrm{d}B_t^1, \ldots, \Sigma_t^d\, \mathrm{d}B_t^d$ separately,

[14] For $\gamma = 1$, this example's argument works only with $b = \beta$. It is shown in Schroder and Skiadas (2005b), however, that any V of the form (36) is ordinally equivalent to a homothetic Duffie–Epstein utility (33), with g specified as in (35) with $\alpha = 0$ and β a deterministic function of time (even if $\gamma = 1$ and $b \neq \beta$).

since they represent exposure to different sources of risk. We model this by postulating a twice continuously differentiable concave function $u:\mathbb{R}^{1+d}\to\mathbb{R}$ such that the time-t conditional certainty equivalent in the recursive specification (27) is defined by

$$u\big(\nu_t(U_{t+\mathrm{d}t}),0,\dots,0\big)=E_t\big[u\big(U_t+\mu_t\,\mathrm{d}t,\Sigma^1\,\mathrm{d}B^1,\dots,\Sigma^d\,\mathrm{d}B^d\big)\big]. \tag{37}$$

The first- and second-order partial derivatives of $u(x_0,x_1,\dots,x_d)$ with respect to x_i are denoted u_i and u_{ii}, respectively. We assume that u is strictly increasing in its first argument. The absolute risk aversion coefficient with respect to the ith risk source is defined by

$$a^i(U)=-\frac{u_{ii}(U,0,\dots,0)}{u_0(U,0,\dots,0)}. \tag{38}$$

We also define the diagonal matrix $A(U)=\operatorname{diag}[a^1(U),\dots,a^d(U)]$. Applying Itô's lemma and taking conditional expectations results in

$$\begin{aligned}&E_t\big[u\big(U_t+\mu_t\,\mathrm{d}t,\Sigma^1\,\mathrm{d}B^1,\dots,\Sigma^d\,\mathrm{d}B^d\big)\big]\\&\quad=u(U_t,0,\dots,0)+u_0(U_t,0,\dots,0)\bigg(\mu_t-\frac{1}{2}\Sigma_t'A(U_t)\Sigma_t\bigg)\,\mathrm{d}t.\end{aligned}$$

Similarly, we have the first-order Taylor expansion

$$\begin{aligned}u\big(\nu_t(U_{t+\mathrm{d}t}),0,\dots,0\big)&=u(U_t,0,\dots,0)+u_0(U_t,0,\dots,0)\\&\quad\times\big(\nu_t(U_{t+\mathrm{d}t})-U_t\big).\end{aligned}$$

Matching the lasts two expressions and simplifying results in the certainty-equivalent expression (28), and corresponding aggregator (29), with the quadratic risk-aversion component

$$\mathcal{A}(\omega,t,U,\Sigma)=\frac{1}{2}\Sigma'A(U)\Sigma.$$

The Duffie–Epstein case is obtained if $a^i=a$ for all i. Combining the above representation with the homothetic specification (24) results in a proportional aggregator of the form (30), where $\mathcal{R}$ is a quadratic form.

Remark 27. A simple extension is obtained if the Brownian motion in the above argument is replaced by a new Brownian motion $\bar{B}$, where $\mathrm{d}\bar{B}$ is a possibly spot-dependent rotation of $\mathrm{d}B$. More formally, we assume $\mathrm{d}\bar{B}_t=\Phi_t\,\mathrm{d}B_t$ for some $\Phi\in\mathcal{L}_2(\mathbb{R}^{d\times d})$ such that $\Phi_t'\Phi_t=I$. In this case, $U_{t+\mathrm{d}t}=U_t+\mu_t\,\mathrm{d}t+\bar{\Sigma}_t'\,\mathrm{d}\bar{B}_t$, where $\bar{\Sigma}_t=\Phi_t\Sigma_t$, and

$$\mathcal{A}(t,U_t,\Sigma_t)=\frac{1}{2}\bar{\Sigma}_t'A(U_t)\bar{\Sigma}_t=\frac{1}{2}\Sigma_t'\Phi_t'A(U_t)\Phi_t\Sigma_t.$$

In the Duffie–Epstein case, $\Phi'A\Phi=A$. With source-dependent risk aversion, however, the aggregator form changes with the Brownian motion rotation.

4.4 First-order risk aversion

Consider an investor who maximizes expected von Neumann–Morgenstern (vNM) utility in a single-period setting. If one were to zoom in a very small area of the graph of the vNM utility, one would see a straight line. This means that an investor is essentially risk-neutral toward the addition of sufficiently small risks to a given wealth level. As an implication, such an investor would seek some exposure to all investment opportunities of positive expected excess return, and would not completely insure a source of risk in actuarially unfavorable terms. These conclusions extend to the recursive utility formulations of the last two sections, as will become clear in the following section. In reality, we observe that investors do not necessarily participate in investment opportunities with positive Sharpe ratios, and they often pay actuarially unfavorable premia to completely insure some sources of risk (for example, against loss of individual items of negligible value relative to total wealth). While such behavior can relate to a number of issues, it is consistent with a certainty-equivalent specification exhibiting first-order risk aversion in the sense of Segal and Spivak (1990). In this subsection, we formulate recursive utility with a conditional certainty equivalent exhibiting first-order risk aversion, whose implication for portfolio choice is discussed in the following section.

In a static expected-utility setting, first-order risk aversion amounts to introducing a kink of the vNM utility around the given wealth level, hence removing local risk neutrality. Since a risk-averse vNM utility can have at most countably many kinks, the approach seems problematic. If one keeps track of different sources of risk, however, as in the source-dependent certainty equivalent introduced above, this problem does not arise. As in the last subsection, we assume the recursive utility specification (27) with the source-dependent certainty equivalent specification in (37), except that the function u in (37) is now replaced with the function

$$\hat{u}(x_0, x_1, \ldots, x_d) = u(x_0, x_1, \ldots, x_d) - \sum_{i=1}^{d} \delta^i(x_0)|x_i|\sqrt{\mathrm{d}t}.$$

The $\sqrt{\mathrm{d}t}$ scaling factor is necessary for a meaningful trade-off between the conditional mean of $\mathrm{d}U_t$, which is order $\mathrm{d}t$, and the conditional absolute variation of $\Sigma^i\,\mathrm{d}B^i$, which is order $\sqrt{\mathrm{d}t}$. We assume that each δ^i is differentiable and nonnegative valued, and that u is exactly as in the last subsection. Since $\hat{u}(U, 0, \ldots, 0) = u(U, 0, \ldots 0)$, the conditional certainty equivalent ν_t is specified by

$$\begin{aligned} u\big(\nu_t(U_{t+\mathrm{d}t}), 0, \ldots, 0\big) = {} & E_t\big[u\big(U_t + \mu_t\,\mathrm{d}t, \Sigma^1\,\mathrm{d}B^1, \ldots, \Sigma^d\,\mathrm{d}B^d\big)\big] \\ & - \sum_{i=1}^{d} E_t\big[\delta^i(U_t + \mu_t\,\mathrm{d}t)\big|\Sigma_t^i\,\mathrm{d}B_t^i\big|\sqrt{\mathrm{d}t}\,\big]. \end{aligned}$$

The left-hand side and the first term of the right-hand side in the above equation are computed exactly as in the last subsection. To compute the last term, we first note that $E_t|\mathrm{d}B_t^i| = \sqrt{2\mathrm{d}t/\pi}$ (since $\mathrm{d}B_t^i$ is normally distributed with zero mean and variance $\mathrm{d}t$). Using the first-order Taylor expansion $\delta^i(U_t + \mu_t\,\mathrm{d}t) = \delta^i(U_t) + \delta^{i\prime}(U_t)\mu_t\,\mathrm{d}t$ and the usual Itô calculus, we find

$$E_t\big[\delta^i(U_t + \mu_t\,\mathrm{d}t)\big|\Sigma_t^i\,\mathrm{d}B_t^i\big|\sqrt{\mathrm{d}t}\,\big] = u_0(U, 0, \ldots, 0)\kappa^i(U_t)\big|\Sigma_t^i\big|\,\mathrm{d}t,$$

where $\kappa^i(U) = \sqrt{2/\pi}\ \delta^i(U)/u_0(U, 0, \ldots, 0)$. Substituting these calculations in the above equation specifying ν_t results in the conditional certainty equivalent expression (28), and corresponding aggregator (29), with

$$\mathcal{A}(\omega, t, U, \Sigma) = \kappa(U)'|\Sigma| + \frac{1}{2}\Sigma' A(U)\Sigma,$$
$$\text{where } \kappa(U) = \big(\kappa^1(U), \ldots, \kappa^d(U)\big)' \text{ and } |\Sigma| = \big(\big|\Sigma^1\big|, \ldots, \big|\Sigma^d\big|\big)'.$$

For the homothetic specification (24), the proportional aggregator takes the form (30), where $\mathcal{R}(\omega, t, \sigma) = \kappa(1)'|\Sigma| + (1/2)\Sigma' A(1)\Sigma$. We revisit the homothetic case in the following section, where the effect of first-order risk aversion on portfolio choice is discussed.

A dual formulation of this subsection's utility corresponds to the "κ-ignorance" multiple-prior formulation of Chen and Epstein (2002). Further discussion of recursive utility duality can be found in El Karoui et al. (2001).

5 Scale-invariant solutions

In this section we study optimal strategies under the homothetic case of last section's utilities, thus taking advantage of the simplifications of scale invariance introduced in Section 3.4, as well as specific risk-aversion parameterizations.

The following condition is assumed to hold throughout the section.

Condition 28. Utility processes are valued in $I_U = (0, \infty)$, and are defined in terms of the functions[15] $g : [0, T] \times (0, \infty) \to \mathbb{R}$ and $\mathcal{R} : \Omega \times [0, T] \times \mathbb{R}^d \to \mathbb{R}$, where $\mathcal{R}(\omega, t, 0) = 0$. For any $c \in \mathcal{C}$, the utility process $U = U(c)$ solves, uniquely in $\mathcal{U}$, the BSDE

$$\frac{\mathrm{d}U_t}{U_t} = -\left(g\left(t, \frac{c_t}{U_t}\right) - \mathcal{R}(t, \sigma_t)\right)\mathrm{d}t + \sigma_t'\,\mathrm{d}B_t, \quad U_T = c_T. \tag{39}$$

For every time t, $g(t, \cdot)$ is differentiable and strictly concave, with derivative $g_x(t, \cdot)$ that maps $(0, \infty)$ onto $(0, \infty)$. Finally, g is sufficiently regular so

[15] The function g is assumed state-independent for economy of exposition. The optimality conditions, however, remain valid for a state-dependent g.

that, for any deterministic $c \in \mathcal{C}$, the ordinary differential equation $\mathrm{d}U/U = -g(t, c/U)\,\mathrm{d}t$, $U_T = 1$, has a unique deterministic solution U in $\mathcal{U}$.

The function g determines choice over deterministic plans. Given g, increasing $\mathcal{R}$ increases the investor's risk aversion, without changing the investor's preferences over deterministic plans. The restriction on g_x reflects our earlier assumption on F_c, and guarantees the strict positivity of an optimal consumption plan.

To state the simplified optimality conditions under the above specification, we introduce some notation. The functions $\mathcal{I}^g$, $g^* : [0, T] \times (0, \infty) \to (0, \infty)$ are defined by

$$g_x\big(t, \mathcal{I}^g(t, \lambda)\big) = \lambda \quad \text{and}$$
$$g^*(t, \lambda) = \max_{x \in \mathbb{R}_{++}} \big(g(t, x) - \lambda x\big) = g\big(t, \mathcal{I}^g(t, \lambda)\big) - \mathcal{I}^g(t, \lambda)\lambda.$$

For any strictly positive Itô process y (such as λ, U, or W), the notation σ^y is defined by the Itô decomposition

$$\frac{\mathrm{d}y}{y} = \ldots \mathrm{d}t + \sigma^{y\prime}\,\mathrm{d}B.$$

As discussed in Section 3.4, since utility is homogeneous of degree-one, at the optimum, the utility process U, the wealth process W, and the shadow-price-of-wealth process λ are related by $U = \lambda W$. The central part of the optimality conditions to follow will be a BSDE solved by $(\lambda, \sigma^\lambda)$. The form of this BSDE is specified by the utility parameters $(g, \mathcal{R})$, and the investment opportunity set parameters (r, μ^R, σ^R). The optimal strategy (ρ, ψ) is computed in terms of $(\lambda, \sigma^\lambda)$ by simple formulas. For the optimal consumption strategy, we note that

$$\rho = \frac{c}{W} = \frac{U}{W}\frac{c}{U} = \lambda x, \quad \text{where } x = \frac{c}{U}.$$

Since $\lambda = F_c(t, c, U, \Sigma) = g_x(t, x)$, it follows that

$$\rho_t = \lambda_t \mathcal{I}^g(t, \lambda_t). \tag{40}$$

The optimal consumption strategy is therefore determined entirely by λ and g. On the other hand, for last section's risk-aversion models, we will see that the optimal trading strategy ψ is determined entirely by σ^λ, the risk-aversion function $\mathcal{R}$, and the excess-return parameters (μ^R, σ^R).

5.1 Smooth quasi-quadratic proportional aggregator

The first specification we consider includes the homothetic version of last section's models of, possibly source-dependent, risk aversion with a smooth aggregator. The case of first-order risk aversion will be treated at the end of this section. Up to that point, we assume:

Condition 29 (Smooth quasi-quadratic proportional aggregator). Condition 28 holds with

$$\mathcal{R}(\omega, t, \sigma) = \frac{1}{2}\sigma' Q(\omega, t)\sigma, \tag{41}$$

for some bounded $Q : \Omega \times [0, T] \to \mathbb{R}^{d\times d}$, where $Q(\omega, t)$ is symmetric positive definite for all (ω, t).

In terms of the risk-aversion function $A(U)$ of Section 4.3, $Q = A(1)$, and therefore Q can be thought of as a relative risk-aversion process. In the Duffie–Epstein case, $Q = \gamma I$, where γ is a coefficient of relative risk aversion, common to all sources of risk. If Q is diagonal, then its ith diagonal element corresponds to relative risk aversion toward risk generated by the ith Brownian motion. Remark 27 leads us to consider nondiagonal positive definite specifications of Q. In last section's parametric formulations of risk aversion, the conditional certainty equivalent was defined in a spot-independent way in terms of the function u, resulting in Q being constant. The same analysis goes through, however, for a function u, and associated conditional certainty equivalent, that is spot-dependent, implying a stochastic risk-aversion process Q.

For every $(\omega, t) \in \Omega \times [0, T]$, the quadratic function $\mathcal{Q}(\omega, t, \cdot) : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\begin{aligned}\mathcal{Q}(t, z) = z' Q_t z - \big(\mu_t^R - \sigma_t^{R\prime}(Q_t - I)z\big)'\big(\sigma_t^{R\prime} Q_t \sigma_t^R\big)^{-1} \\ \times \big(\mu_t^R - \sigma_t^{R\prime}(Q_t - I)z\big), \quad z \in \mathbb{R}^d.\end{aligned}$$

Under Condition 29, the solution of the optimality conditions (stated in Condition 22) reduces to the following procedure:

1. Compute $(\lambda, \sigma^\lambda)$ by solving the BSDE:

$$\frac{\mathrm{d}\lambda_t}{\lambda_t} = -\left(r_t + g^*(t, \lambda) - \frac{1}{2}\mathcal{Q}(t, \sigma^\lambda)\right)\mathrm{d}t + \sigma_t^{\lambda\prime}\,\mathrm{d}B_t, \quad \lambda_T = 1. \tag{42}$$

2. Given the solution $(\lambda, \sigma^\lambda)$, the optimal consumption strategy ρ is given by Equation (40), and the optimal trading strategy is

$$\psi_t = \big(\sigma_t^{R\prime} Q_t \sigma_t^R\big)^{-1}\big(\mu_t^R - \sigma_t^{R\prime}(Q_t - I)\sigma_t^\lambda\big). \tag{43}$$

3. The wealth process W generated by the strategy (ρ, ψ) is computed from the budget equation (4). The optimal consumption plan financed by (ρ, ψ) is $c = \rho W$, and the utility process of c is $U = \lambda W$.

The proof of this claim (given in Schroder and Skiadas, 2003) is a matter of direct calculation using the specific aggregator form, and the key homogeneity condition $U = \lambda W$. The optimal trading strategy expression follows from Equations (23), which imply that $\mu^R + \sigma^{R\prime}(F_\Sigma + \sigma^\lambda) = 0$. In this equation we substitute $F_\Sigma = -Q\sigma^U$ (from the definition of F), $\sigma^U = \sigma^\lambda + \sigma^W$

(from $U = \lambda W$), and $\sigma^W = \sigma^R \psi$ (from the budget equation). Solving for ψ gives (43).

The optimal trading strategy of Equation (43) can deviate from an instantaneously mean–variance efficient solution for two possible reasons. One is source dependence of risk aversion, reflected in Q, and the other is the term involving σ^λ which arises from a stochastic investment opportunity set or stochastic risk aversion. Two special cases in which instantaneous mean–variance efficiency is recovered are given in the following examples. For expositional simplicity, we informally identify optimality with the above optimality conditions (ignoring the regularity assumptions required for actual equivalence).

Example 30 (Deterministic investment opportunity set and risk aversion). Suppose that the investment opportunity set parameters (r, μ^R, σ^R), and the risk-aversion process Q are all deterministic. Then the solution simplifies significantly by setting $\sigma^\lambda = 0$. That is, λ is a deterministic process solving the ODE

$$\frac{\mathrm{d}\lambda_t}{\lambda_t} = -\left(r_t + g^*(t, \lambda) + \frac{1}{2}\mu_t^{R\prime}\left(\sigma_t^{R\prime} Q_t \sigma_t^R\right)^{-1}\mu_t^R\right)\mathrm{d}t, \quad \lambda_T = 1.$$

Since λ is deterministic and g is assumed state-independent, the optimal consumption strategy $\rho = \lambda \mathcal{I}^g(\lambda)$ is also deterministic. The optimal trading strategy is $\psi = (\sigma^{R\prime} Q \sigma^R)^{-1}\mu^R$.

Suppose further that risk aversion is source-independent, and therefore $Q = \gamma I$ for some deterministic process γ. Then the optimal trading strategy $\psi = \gamma^{-1}(\sigma^{R\prime}\sigma^R)^{-1}\mu^R$ is instantaneously mean–variance efficient. Since ψ does not depend on g, it is the same as for the choice of g given in Example 26 with $\gamma = \delta$. In other words, the optimal trading strategy is the same as for time-additive power expected utility (considered by Merton, 1971). On the other hand, λ and the optimal consumption strategy depend on the specification of g. It is also worth noting that in the current special context the investment opportunity set enters the dynamics of λ only through the maximum squared conditional Sharpe ratio of Equation (3).

Example 31 (Robustly mean–variance efficient optimal trading strategies). Even under a stochastic investment opportunity set, the instantaneously mean–variance efficient strategy $\psi = (\sigma^{R\prime}\sigma^R)^{-1}\mu^R$ is optimal if $Q = I$ (the identity matrix). Moreover, for $Q = I$, the investment opportunity set enters the BSDE for λ only through λ and the maximum squared instantaneous Sharpe ratio of Equation (3). Combining time additivity with the assumption $Q = I$ implies that g is logarithmic (Example 26 with $\gamma = \delta = 1$), and therefore the optimal consumption strategy equals the utility discount rate as in Example 25. Without time-additivity, g is unrestricted. A discrete-time example of this type was first given by Giovannini and Weil (1989). The construction is further extended in Example 33 below.

As noted earlier, in a Markovian setting, a BSDE is characterized (under some regularity) by a corresponding PDE. The argument is outlined below for the BSDE (42) satisfied by λ.

Example 32 (Markovian solutions). Given is some underlying n-dimensional Markov process Z, uniquely solving the SDE

$$\mathrm{d}Z = a(t, Z)\,\mathrm{d}t + b(t, Z)'\,\mathrm{d}B, \quad Z_0 = z_0,$$

for some $z_0 \in \mathbb{R}^n$ and functions $a : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ and $b : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{d\times n}$. With some convenient abuse of notation, we assume that

$$r_t = r(t, Z_t) \quad \text{and} \quad \eta_t = \eta(t, Z_t),$$

for some functions $r : [0, T] \times \mathbb{R}^n \to \mathbb{R}$ and $\eta : [0, T] \times \mathbb{R}^n \to \mathbb{R}^d$. We conjecture that λ can be written as a function of time and the Markov state that is smooth enough to apply Itô's lemma. With the usual abuse of notation, we write $\lambda(\omega, t) = \lambda(t, Z(\omega, t))$. Omitting the argument (t, Z_t), and with subscripts of λ denoting partial derivatives, Itô's lemma implies:

$$\mathrm{d}\lambda = \left(\lambda_t + \lambda_z' a + \frac{1}{2}\operatorname{tr}\left[b\lambda_{zz}b'\right]\right)\mathrm{d}t + (b\lambda_z)'\,\mathrm{d}B.$$

Comparing the above dynamics to BSDE (42) suggests that $\sigma^\lambda = b\lambda_z/\lambda$ and λ solves the PDE

$$r + g^*(\lambda) - \frac{1}{2}\mathcal{Q}\left(b\frac{\lambda_z}{\lambda}\right) + \frac{\lambda_t}{\lambda} + a'\frac{\lambda_z}{\lambda} + \frac{1}{2}\operatorname{tr}\left[b\frac{\lambda_{zz}}{\lambda}b'\right] = 0,$$
$$\lambda(T, \cdot) = 1,$$

where r and $\mathcal{Q}$ are viewed as functions of the underlying Markov state, in the same notational pattern used earlier for λ. Reversing the above steps, one can construct a solution to BSDE (42) from a solution to the above PDE.

5.2 *Relating complete and incomplete market solutions*

Continuing with the assumption of a smooth quasiquadratic proportional aggregator (Condition 29), we consider some connections between complete and incomplete market solutions. (The type of market incompleteness discussed here leaves out the case of undiversifiable income risk. A tractable class of problems dealing with the latter is outlined in the final section.)

We recall that m is the number of traded risky assets. Given a matrix A of dimension $n \times k$, where $n \geqslant m$, we use the block matrix notation:

$$A = \begin{bmatrix} A_M \\ A_N \end{bmatrix}, \quad A_M \in \mathbb{R}^{m\times k},\ A_N \in \mathbb{R}^{(n-m)\times k}.$$

In particular, $R = R_M$, $\mu^R = \mu^R_M$, and $\sigma^{R\prime} = [\sigma^{R\prime}_M, \sigma^{R\prime}_N]$.

While the solution summarized by BSDE (42) is valid for incomplete markets, the role of nonmarketed uncertainty becomes clearer after passing to a new Brownian motion that generates the same filtration as B, and separates marketed and nonmarketed uncertainty. Informally, at each spot, the linear span of $\mathrm{d}R_M - \mu_M^R\,\mathrm{d}t$ can be obtained as the linear span of the first m elements of a rotated version of $\mathrm{d}B$. This transformation (stated formally in Schroder and Skiadas, 2003) corresponds to the type of spot-by-spot Brownian motion rotation of Remark 27, which preserves the quasiquadratic proportional aggregator structure. We therefore lose no generality in assuming that

$$\mathrm{d}R_M = \mu_M^R\,\mathrm{d}t + \sigma_M^{R\prime}\,\mathrm{d}B_M \quad \text{and} \quad \sigma_N^R = 0. \tag{44}$$

For the remainder of this section, we assume the normalized return structure (44), and we think of M and N as sets of indices corresponding to marketed and nonmarketed uncertainty, respectively. The processes r, μ_M^R, and σ_M^R need not, however, be adapted to the filtration generated by B_M.

A market-price-of-risk process in this context takes the form

$$\eta = \begin{bmatrix} \eta_M \\ \eta_N \end{bmatrix}, \quad \text{where } \eta_M = \left(\sigma_M^{R\prime}\right)^{-1}\mu_M^R. \tag{45}$$

The process η_M represents the *price of marketed risk*, while the unrestricted process η_N represents the *price of nonmarketed risk*. The latter parameterizes the set of every state price density π consistent with the given market:

$$\pi = \pi^M \xi^{\eta_N}, \quad \text{where } \frac{\mathrm{d}\pi^M}{\pi^M} = -r\,\mathrm{d}t - \eta_M'\,\mathrm{d}B_M$$
$$\text{and} \quad \frac{\mathrm{d}\xi^{\eta_N}}{\xi^{\eta_N}} = -\eta_N'\,\mathrm{d}B_N.$$

If π is a state price density that is also a utility supergradient density at an optimum, then the corresponding η_N reflects the shadow price of nonmarketed risk, in the following sense. Consider a hypothetical market completion in which risk generated by $\mathrm{d}B_N$ is priced by η_N. In such a market, the investor would find it optimal to not trade risk generated by $\mathrm{d}B_N$, since the original incomplete-markets strategy would still be optimal. Since the original strategy need not be optimal under any other choice of η_N, the incomplete-markets optimal utility is the minimum of optimal utilities over all market completions (parameterized by η_N). This connection between complete and incomplete market solutions is illustrated more concretely in Example 34 below, and extends to more general convex constraints (see Cvitanić and Karatzas, 1992 and Karatzas and Shreve, 1998 for the case of time-additive expected utility, and Schroder and Skiadas, 2003 and Appendix A of Schroder and Skiadas, 2005b for the case of recursive utility).

For expositional simplicity, in the remainder of this section we further assume that the relative risk aversion process Q assumes the block diagonal

structure

$$Q = \begin{bmatrix} Q_{MM} & 0 \\ 0 & Q_{NN} \end{bmatrix}, \tag{46}$$

where $Q_{MM} \in \mathcal{L}(\mathbb{R}^{m\times m})$ and $Q_{NN} \in \mathcal{L}(\mathbb{R}^{(d-m)\times(d-m)})$. In this context, the function $\mathcal{Q}$ of BSDE (42) of the optimality conditions can be written as

$$\begin{aligned}\mathcal{Q}(\sigma^\lambda) = {}& \sigma_N^{\lambda\prime} Q_{NN} \sigma_N^\lambda + 2\big(\eta_M + \sigma_M^\lambda\big)' \sigma_M^\lambda \\ & - \big(\eta_M + \sigma_M^\lambda\big)' Q_{MM}^{-1} \big(\eta_M + \sigma_M^\lambda\big).\end{aligned} \tag{47}$$

The corresponding optimal trading strategy is

$$\psi_M = \big(Q_{MM}\sigma_M^R\big)^{-1}\big(\eta_M - (Q_{MM} - I_{MM})\sigma_M^\lambda\big),$$

where I_{MM} is the $m \times m$ identity matrix.

Example 33 (Mean–variance efficiency). If $Q_{MM} = I_{MM}$, then ψ_M is instantaneously mean–variance efficient, an observation that extends Example 31.

Example 34 (Fictitious market completion and duality). Consider the above incomplete-market setting, with the normalized return dynamics (44), where $m < d$, and the block-diagonal Q in (46). Suppose that $(\lambda, \sigma^\lambda)$ solves the BSDE of the optimality conditions, (ρ, ψ_M) is the corresponding optimal strategy, and U is the corresponding optimal utility process. Given any choice of a nonmarketed-price-of-risk process η_N, we consider the complete market obtained by introducing $d - m$ fictitious assets, whose cumulative excess return process R_N follows the dynamics $\mathrm{d}R_N = \eta_N\,\mathrm{d}t + \mathrm{d}B_N$. The unique market-price-of-risk process in this fictitious complete market is given by (45). We let U^{η_N} denote the corresponding complete-market optimal utility process. Simple algebra shows that if one makes the specific selection

$$\eta_N = (Q_{NN} - I_{NN})\sigma_N^\lambda,$$

then $(\lambda, \sigma^\lambda)$ satisfies the BSDE of the optimality conditions in the fictitious complete market defined by this choice of η_N. Moreover, the corresponding optimal strategy in the fictitious complete market is (ρ, ψ), where (ρ, ψ_M) is the incomplete-market optimal strategy and $\psi_N = 0$. In other words, the above specification of η_N prices nonmarketed risk so that the investor finds it optimal to not trade the fictitious assets at all. As a consequence $U = U^{\eta_N}$. For any other choice of η_N, (ρ, ψ) need not be optimal in the fictitious complete market defined by η_N, and therefore $U \leqslant U^{\eta_N}$.

A different type of connection between incomplete and complete market solutions is given in the following example (which is generalized in Schroder and Skiadas, 2003). A particular case of the example shows that if the investor has the time-additive expected power utility (36) with $\gamma \in (0, 2)$, then the

solution to the investor's problem in an incomplete market is equivalent (in a sense clarified below) to the solution of the complete market problem obtained by pricing nontraded uncertainty risk-neutrally, and setting the investor's relative risk aversion toward nonmarketed uncertainty to $1/(2-\gamma)$. The original additive-utility problem with incomplete markets is therefore equivalent to a complete-markets problem with recursive utility.

Example 35 (Market-incompletess and source-dependent risk aversion). We further specialize the quasiquadratic form (41) of the proportional aggregator by assuming that

$$Q = \gamma I, \quad \text{where } \gamma \in (0, 2).$$

In Example 26 we saw that this class includes cases of Epstein–Zin utility, as well as time-additive expected discounted power utility. Let (ρ, ψ_M) be an incomplete-market optimal strategy, with corresponding shadow-price-of-wealth process λ, wealth process W, and utility process U.

We complete the market by introducing fictitious assets that are priced risk-neutrally; that is, the price-of-nonmarketed risk process is zero ($\eta_N = 0$). We let the corresponding excess return dynamics be given by $R_N = B_N$. In the resulting fictitious complete market, we consider the optimal strategy, not of the original investor, but rather of a fictitious investor whose proportional aggregator is

$$\bar{G}(t, c, \sigma) = g(t, c) - \frac{1}{2}\left(\gamma \sigma_M' \sigma_M + \frac{1}{2-\gamma}\sigma_N' \sigma_N\right). \tag{48}$$

In other words, the fictitious investor's relative risk aversion toward nonmarketed risk is modified from γ to $1/(2-\gamma)$. Let $(\bar{\rho}, \bar{\psi})$ be the optimal strategy of the fictitious investor in the fictitious complete market, and let $\bar{\lambda}$, $\bar{W}$ and $\bar{U}$ be the corresponding shadow-price-of wealth, wealth, and utility processes. Comparing optimality conditions, we observe that

$$\bar{\lambda} = \lambda, \quad \bar{\rho} = \rho, \quad \bar{\psi}_M = \psi_M,$$
$$\text{and} \quad \frac{\bar{W}_t}{W_t} = \frac{\bar{U}_t}{U_t} = \exp\left(\int_0^t \bar{\psi}_N' \, \mathrm{d}B_N\right).$$

The incomplete market solution can therefore be immediately recovered from the fictitious complete-market solution. This is true even though the specification of the fictitious-investor preferences does not depend on market prices!

5.3 Solutions based on quadratic BSDEs

In this subsection, we discuss scale-invariant formulations in which the BSDE satisfied by $\log(\lambda)$ takes a quadratic form. For certain specifications of the return dynamics, the quadratic BSDE solution can be expressed as a

quadratic function[16] of an exogenous state process, with deterministic coefficients that solve an ODE system. This type of solution is familiar in Finance mainly in the context of risk-neutral pricing (see, for example, Duffie, 2005 and Piazzesi, 2005), where the relevant BSDE is linear. Our application extends the solution method to quadratic BSDEs, where the quadratic term reflects risk aversion. For expositional simplicity, we outline below only some examples, referring to Schroder and Skiadas (2003, 2005a, 2005b) for more general treatments. Examples of this type of solution can also be found in Chacko and Viceira (2005), Kim and Omberg (1996), Liu (2005), Schroder and Skiadas (1999), and Wachter (2002).

Continuing with the assumption of a homothetic recursive utility, we adopt the utility specification in the intersection of Examples 25 and 26; that is, the proportional aggregator is of the form

$$G(t, x, \sigma) = \alpha + \beta \log(x) - \frac{\gamma}{2}\sigma'\sigma, \tag{49}$$

for some constants $\alpha \in \mathbb{R}$ and $\beta, \gamma \in \mathbb{R}_{++}$. The parameters (α, β) determine preferences over deterministic choices. Given (α, β), the parameter γ adjusts risk aversion.

Remark 36. The treatment in Schroder and Skiadas (2003) allows possible source-dependent risk aversion, and parameters α and β that are processes, the latter deterministic. The above specification for $\beta = 0$ results in a utility that is ordinally equivalent to expected power utility for terminal consumption. Even though we have not covered the case of no intermediate consumption, essentially the same analysis applies.

Recalling Example 25, the optimal strategy for the above utility specification is

$$\rho = \beta \quad \text{and} \quad \psi = \frac{1}{\gamma}\left(\sigma^{R\prime}\sigma^{R}\right)^{-1}\left(\mu^{R} - (\gamma - 1)\sigma^{R\prime}\sigma^{\lambda}\right).$$

A myopic solution results for $\gamma = 1$, corresponding to time-additive logarithmic utility (the intersection of Examples 25 and 31). The solution also simplifies if the investment opportunity set is deterministic, in which case $\sigma^{\lambda} = 0$ (Example 30). To compute the optimal strategy with a stochastic investment opportunity set and $\gamma \neq 1$, we need to determine $(\lambda, \sigma^{\lambda})$ by solving BSDE (42). Making the convenient change of variables

$$\ell_t = \log(\lambda_t),$$

[16] As explained in Schroder and Skiadas (2005b), the quadratic dependence on the state can be made affine by a suitable redefinition of the state process. A similar construction in a term-structure context appears in Cheng and Scaillet (2005).

we note that $g^*(t,\lambda) = \alpha - \beta + \beta\log(\beta) - \beta\ell_t$. Direct computation shows that BSDE (42) can be written as a quadratic BSDE to be solved for ℓ:

$$\mathrm{d}\ell_t = -\left(p_t - \beta\ell_t + h_t'\sigma_t^\ell + \frac{1}{2}\sigma_t^{\ell\prime} H_t \sigma_t^\ell\right)\mathrm{d}t + \sigma_t^{\ell\prime}\,\mathrm{d}B, \quad \ell_T = 0, \quad (50)$$

where

$$p = r + \alpha - \beta + \beta\log(\beta) + \frac{1}{2\gamma}\mu^{R\prime}\left(\sigma^{R\prime}\sigma^R\right)^{-1}\mu^R,$$
$$h = \frac{1-\gamma}{\gamma}\sigma^R\left(\sigma^{R\prime}\sigma^R\right)^{-1}\mu^R,$$
$$H = (1-\gamma)\left[I + \frac{1-\gamma}{\gamma}\sigma^R\left(\sigma_t^{R\prime}\sigma^R\right)^{-1}\sigma^{R\prime}\right].$$

A general set of conditions under which the above quadratic BSDE can be reduced to an ODE is given in Schroder and Skiadas (2003). We only consider here two representative examples. As in the last subsection, we assume the normalization $\mathrm{d}R = \mu^R\,\mathrm{d}t + \sigma_M^{R\prime}\,\mathrm{d}B_M$, and therefore the price-of-marketed-risk process is $\eta_M = (\sigma_M^{R\prime})^{-1}\mu^R$. We outline the form of the solution below, leaving the details as an exercise.

Example 37. Given is an underlying n-dimensional state vector Z following the dynamics

$$\mathrm{d}Z = (\mu - \theta Z)\,\mathrm{d}t + \Sigma'\,\mathrm{d}B,$$

for some $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{d\times n}$, and $\theta \in \mathbb{R}^{n\times n}$. The short-rate process and price-of-marketed-risk process are assumed to be given by

$$r = C_0^r + C_1^{r\prime}Z + \frac{1}{2}Z'C_2^rZ \quad \text{and} \quad \eta_M = C_0^\eta + C_1^{\eta\prime}Z,$$

where the coefficients C_i^r and C_i^η are all constants of conforming dimensions. In this case, we conjecture a solution to BSDE (50) of the form

$$\ell_t = C_0(t) + C_1(t)'Z_t + \frac{1}{2}Z_t'C_2(t)Z_t,$$

where the $C_i(t)$ are deterministic differentiable processes. Applying Itô's lemma to the above conjectured expression, collecting terms and comparing to the corresponding coefficients of BSDE (50) confirms that such a solution indeed solves BSDE (50), provided the coefficients C_i solve an ODE system.

Example 38. We modify the above example by assuming the dynamics

$$\mathrm{d}Z = (\mu - \theta Z)\,\mathrm{d}t + \Sigma'\,\mathrm{diag}\left(\sqrt{v + VZ}\right)\mathrm{d}B,$$
$$r = C_0^r + C_1^{r\prime}Z \quad \text{and} \quad \eta_M = \mathrm{diag}\left(\sqrt{v_M + V_MZ_t}\right)\varphi,$$

where $\operatorname{diag}(x)$ denotes the diagonal matrix with x on the diagonal, $\sqrt{x}$ denotes the vector with ith element $\sqrt{x_i}$, and $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{d\times n}$, $\theta \in \mathbb{R}^{n\times n}$, $C_0^r \in \mathbb{R}$, $C_1^r \in \mathbb{R}^n$, $\upsilon = [\upsilon_M', \upsilon_N']' \in \mathbb{R}^d$, $V = [V_M', V_N']' \in \mathbb{R}^{d\times n}$, $\varphi \in \mathbb{R}^m$. In this case, the conjectured solution takes the form

$$\ell_t = C_0(t) + C_1(t)'Z,$$

where again the $C_i(t)$ are differentiable deterministic processes. Arguing as in the last example, one obtains an ODE solved by C_1 alone, and another ODE (which uses the solution of the first ODE) that is satisfied by C_0. Given the pair (C_0, C_1) satisfying the ODE pair, the above affine expression defines a solution to BSDE (50).

5.4 *Solutions with first-order risk aversion*

The final set of scale-invariant solutions we consider utilizes the kinked proportional aggregator of Section 4.4, representing source-dependent first-order risk aversion. More specifically, we assume the following condition, using the notation

$$|x|' = \big(|x_1|, \ldots, |x_d|\big), \quad x \in \mathbb{R}^d.$$

Condition 39 (Quasi-quadratic proportional aggregator). Condition 28 holds with

$$\mathcal{R}(\omega, t, \sigma) = \kappa(\omega, t)'|\sigma| + \frac{1}{2}\sigma' Q(\omega, t)\sigma,$$

for some bounded processes $\kappa : \Omega \times [0, T] \to \mathbb{R}^d$ and $Q : \Omega \times [0, T] \to \mathbb{R}^{d\times d}$, where $Q(\omega, t)$ is *diagonal* and positive definite for all (ω, t).

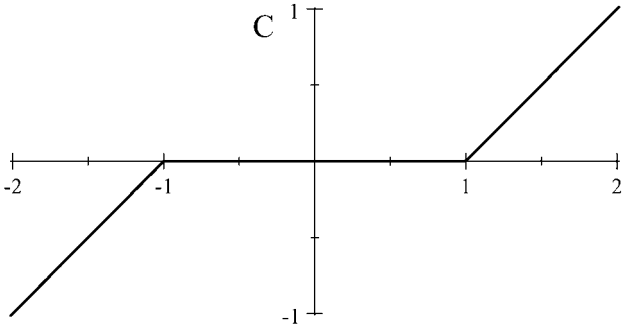
We adopt the notation and return normalization of Section 5.2. In particular, the excess return dynamics and the marketed-price-of-risk process are

$$\mathrm{d}R = \mu^R\,\mathrm{d}t + \sigma_M^{R\prime}\,\mathrm{d}B_M \quad \text{and} \quad \eta_M = \big(\sigma_M^{R\prime}\big)^{-1}\mu_M^R.$$

To formulate optimality conditions in this setting, we define, for any $\kappa \in \mathbb{R}_+$, the *collar function*

$$C(\alpha; \kappa) = \min\big\{\max\{0, \alpha - \kappa\}, \alpha + \kappa\big\}, \quad \alpha \in \mathbb{R},$$

plotted below for $\kappa = 1$:



The collar function will be applied to vectors coordinate by coordinate:

$$C(\alpha;\kappa) = \big(C(\alpha_1;\kappa_1),\ldots,C(\alpha_m;\kappa_m)\big)',$$
$$\text{for any } \alpha \in \mathbb{R}^m \text{ and } \kappa \in \mathbb{R}^m_+.$$

The BSDE for λ in this case is of the same form as in the smooth-quasi-quadratic case, except that $\mathcal{Q}$ is replaced with the function $\mathcal{K}:\Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$\mathcal{K}(z) = 2\kappa_N'|z_N| + z_N'Q_{NN}z_N + 2(\eta_M + z_M)'z_M - C(\eta_M + z_M;\kappa_M)'Q_{MM}^{-1}C(\eta_M + z_M;\kappa_M).$$

Expression (47) for $\mathcal{Q}$ is recovered if one sets $\kappa = 0$.

With the above notation and return normalization in place, the optimality conditions (Condition 22) under Condition 39 reduce to the following steps:

1. The shadow-price-of-wealth process λ solves the BSDE:
$$\frac{\mathrm{d}\lambda_t}{\lambda_t} = -\left(r_t + g^*(t,\lambda_t) - \frac{1}{2}\mathcal{K}\big(t,\sigma_t^\lambda\big)\right)\mathrm{d}t + \sigma_t^{\lambda\prime}\,\mathrm{d}B_t, \quad \lambda_T = 1.$$
2. Given the solution (λ,σ^λ) from step one, the optimal strategy is
$$\rho = \lambda\mathcal{I}^g(\lambda) \quad \text{and} \quad \psi = \big(\sigma_M^R\big)^{-1}\big[Q_{MM}^{-1}C\big(\eta_M + \sigma_M^\lambda;\kappa_M\big) - \sigma_M^\lambda\big].$$
3. The wealth process generated by the strategy (ρ,ψ) is computed from the budget equation, the corresponding optimal consumption plan is $c = \rho W$, and its utility process is $U = \lambda W$.

The proof of the above claim (given in Schroder and Skiadas, 2003) is a matter of direct calculation using the specific aggregator form, and the key homogeneity condition at the optimum: $U = \lambda W$. The latter also implies that if $\mathrm{d}U/U = \ldots\mathrm{d}t + \sigma^{U\prime}\,\mathrm{d}B$, then

$$\sigma_M^U = (Q_{MM})^{-1}C\big(\eta_M + \sigma_M^\lambda;\kappa_M\big).$$

Consequently, for any $i \in M$, σ^{Ui} vanishes whenever $\eta_M^i + \sigma_M^{\lambda i} \in [-\kappa_i, \kappa_i]$. Such perfect hedging of utility risk with respect to some source of risk is not encountered with an aggregator derived from a smooth certainty equivalent. The following example (from Schroder and Skiadas, 2003) extends Section 5.3 of Chen and Epstein (2002). Further examples of nonparticipation as an expression of source-dependent first-order risk aversion can be found in Epstein and Miao (2003) and Schroder and Skiadas (2003).

Example 40 (Deterministic investment opportunity set). Suppose that r, μ^R, σ^R, κ, and Q are all deterministic. Then the solution simplifies by setting $\sigma^\lambda = 0$. In particular, the optimal trading strategy is $\psi = (Q_{MM}\sigma_M^R)^{-1} \times C(\eta_M; \kappa_M)$. Let us further assume, for simplicity, that σ_{MM}^R is diagonal with positive diagonal. For any $i \in M$, $\psi_i = 0$ when $\eta_i \in [-\kappa_i, +\kappa_i]$; the agent will not participate in the market for risk i, unless its instantaneous expected return relative to its risk is sufficiently far from zero. This type of solution can be combined with different belief specifications, as in Remark 15, to obtain a richer set of optimal portfolio holdings. In particular, adding the term $b'\sigma$ to the proportional aggregator, for some (bounded) process b, means that the investor believes the market price of risk process to be $\eta - b$, rather than η, and therefore the investor will not participate in the market for risk source i if $\eta_i \in [b_i - \kappa_i, b_i + \kappa_i]$. If we further assume that $b_i = -\kappa_i$, then the optimal holding of asset i is $\psi_i = Q_{ii}^{-1}\mu_i^R/(\sigma_{ii}^R)^2$ when $\mu_i^R > 0$ (just as with $\kappa = b = 0$), but the agent will only short asset i when $\mu_i^R < -2\kappa_i\sigma_{ii}^R$. In other words, in this case, the optimal portfolio is identical to the Merton solution for positive expected excess returns, yet it is optimal for the investor to not go short for sufficiently small negative expected returns.

6 Extensions

This section concludes with two direct extensions of the main chapter's material, and a list of further topics and related references.

6.1 Convex trading constraints

We outline an extension of this chapter's arguments to include convex trading constraints, referring to Schroder and Skiadas (2003) for details. Examples of analysis of the Merton problem with constraints based on the Hamilton–Jacobi–Bellman approach include Zariphopoulou (1994) and Vila and Zariphopoulou (1997). Convex duality with trading constraints and additive utility is studied by He and Pearson (1991), Karatzas et al. (1991) (incomplete markets); Shreve and Xu (1992a, 1992b) (short-sale constraints); and Cvitanić and Karatzas (1992) (convex constraints). Related discussions with recursive preferences can be found in El Karoui et al. (2001), and Schroder and Skiadas

(2003, 2005b). We will not discuss duality here. Also not discussed here are constraints that prevent the investor from borrowing against future income, which is the focus of He and Pagès (1993), El Karoui and Jeanblanc-Picquè (1998), and Detemple and Serrat (2003).

We consider this chapter's setting with the additional constraint that the investor's trading strategy must be valued in some given convex set $K \subseteq \mathbb{R}^m$ at all times. For example, $K = \mathbb{R}^m_+$ represents the impossibility of short-selling. The definition of a feasible cash flow now includes the requirement that it can be financed by a K-valued trading strategy. We let $\delta_K(\varepsilon_t) = \sup\{k'\varepsilon_t\colon\ k \in K\}$ denote the support function of K.

We fix a feasible strategy (ρ, ψ) financing the consumption plan c. Given our new notion of feasibility, the definition of a state-price density at c is the same as before. The smaller the set K, the smaller the set of feasible incremental cash flows, and therefore the larger the set of state price densities at c. Under some regularity, state price dynamics are characterized in Schroder and Skiadas (2003) as

$$\begin{aligned}
&\frac{\mathrm{d}\pi_t}{\pi_t} = -\big(r_t + \delta_K(\varepsilon_t)\big)\,\mathrm{d}t - \eta_t'\,\mathrm{d}B_t,\\
&\varepsilon_t = \mu_t^R - \sigma_t^{R\prime}\eta_t, \quad \psi_t'\varepsilon_t = \delta_K(\varepsilon_t).
\end{aligned}$$

Proposition 3 still applies here, so combining the above dynamics with those of a utility supergradient density results in sufficient optimality conditions as a constrained FBSDE system.

As in the unconstrained case, scale invariance results in the uncoupling of the forward and backward components of the BSDE system. For example, consider a scale-invariant recursive utility with the smooth quasi-quadratic proportional aggregator of Condition 29, a specification that includes expected discounted power utility and Epstein–Zin utility (see Example 26). As shown in Schroder and Skiadas (2003), in this case the optimality conditions can be written as the constrained BSDE:

$$\begin{aligned}
&\frac{\mathrm{d}\lambda_t}{\lambda_t} = -\bigg(r_t + \delta_K(\varepsilon_t) + g^*(t,\lambda_t) - \frac{1}{2}\sigma_t^{\lambda\prime}Q_t\sigma_t^{\lambda}\\
&\qquad\qquad + \frac{1}{2}\psi_t'\sigma_t^{R\prime}Q_t\sigma_t^R\psi_t\bigg)\,\mathrm{d}t + \sigma_t^{\lambda\prime}\,\mathrm{d}B_t, \quad \lambda_T = 1,\\
&\psi_t = \big(\sigma_t^{R\prime}Q_t\sigma_t^R\big)^{-1}\big(\mu_t^R - \varepsilon_t - \sigma_t^{R\prime}(Q_t - I)\sigma_t^{\lambda}\big) \in K,\\
&\psi_t'\varepsilon_t = \delta_K(\varepsilon_t).
\end{aligned}$$

Example 41. Under Condition 29, a particularly simple expression for the optimal trading strategy is obtained if $K = \{k \in \mathbb{R}^m\colon\ \alpha \leqslant l'k \leqslant \beta\}$ where $l \in \mathbb{R}^m$ and α and β are valued in $[-\infty, +\infty]$. The case of a short-sale constraint on asset i corresponds to $\alpha = 0$, $\beta = \infty$, and l a vector of zeros except for a one in the ith position. The case of a cap on the proportion of wealth borrowed, possibly combined with a limit on short sales as a fraction of wealth, corresponds

to letting l be a vector of ones. We assume that K is nonempty, and define

$$\psi_t^* = A_t\big(\mu_t^R - \sigma_t^{R\prime}(Q_t - I)\sigma_t^\lambda\big), \quad A_t = \big(\sigma_t^{R\prime} Q_t \sigma_t^R\big)^{-1}.$$

The above expression gives the optimal trading strategy as a function of σ^λ in the unconstrained case ($\alpha = -\infty$, $\beta = \infty$). The (constrained) optimal trading strategy ψ and process ε in the dynamics of λ are given by

$$\psi_t = \psi_t^* - A_t\varepsilon_t, \quad \varepsilon_t = -(l'A_t l)^{-1}\big[\min\big\{\max\{l'\psi_t^*, \alpha\}, \beta\big\} - l'\psi_t^*\big]l. \tag{51}$$

These equations can in turn be used to complete the specification of the BSDE for λ, which can then be solved by some numerical method (for example, with numerical PDE methods in a Markovian setting).

6.2 *Translation-invariant formulations and nontradeable income*

A parallel theory to this chapter's scale-invariance argument is based on a notion of translation invariance in a setting that allows for a nontradeable income stream. This type of formulation is familiar in the subclass of problems with expected discounted exponential utility and Gaussian dynamics, as, for example, in Svensson and Werner (1993) and Musiela and Zariphopoulou (2004). We outline below a formulation with recursive utility, which is a special case of Schroder and Skiadas (2005a, 2005b) (where trading constraints, nonlinear wealth dynamics, and unpredictable return jumps are also considered).

We modify our earlier setting by assuming that the investor is endowed with a possibly nontradeable cash flow e, in addition to the initial wealth w_0. Consumption in this subsection is allowed to take negative values, and financial wealth can vanish. The representation of portfolios in terms of proportions of wealth is therefore unsuitable in our new setting. We correct this by defining a trading plan to be a process $\phi \in \mathcal{L}(\mathbb{R}^m)$, where ϕ_t^i represents a dollar amount invested in asset i at time t. The dollar amount invested in the money market at time t is $W_t - \sum_{i=1}^m \phi_t^i$, where W_t represents total time-t financial wealth (excluding e). Ignoring some integrability requirements, a *plan* is a triple (c, ϕ, W) of a consumption plan c, a trading plan ϕ, and a wealth process W. The plan (c, ϕ, W) is *feasible* if it satisfies the budget equation:

$$W_0 = w, \quad \mathrm{d}W_t = (r_t W_t + e_t - c_t)\,\mathrm{d}t + \phi_t'\,\mathrm{d}R_t, \quad c_T = W_T + e_T. \tag{52}$$

The derivation and form of the optimality conditions as a FBSDE system in this setting is similar to this chapter's main analysis, as explained in Schroder and Skiadas (2005a, 2005b).

We place restrictions on the market and preferences in terms of a strictly positive (bounded) cash flow γ, that is fixed throughout. On the market side, we assume there is a tradeable fund that generates γ as a dividend stream. We refer to this fund as the "γ-fund," and we let Γ_t and κ_t be its time-t value and

value allocation, respectively.[17] The γ-fund budget equation is

$$\mathrm{d}\Gamma_t = (r_t\Gamma_t - \gamma_t)\,\mathrm{d}t + \Gamma_t\kappa_t'\,\mathrm{d}R_t, \quad \Gamma_T = \gamma_T.$$

For example, if r and γ are deterministic, the γ-fund can be implemented entirely through the money market (with $\kappa = 0$). If either r or γ is stochastic, one can assume that risky asset one is a share in the γ-fund, and therefore $\kappa = (1, 0, \ldots, 0)$.

On the preference side, we assume that the investor's time-zero utility function is *translation-invariant with respect to* γ, meaning that, for any consumption plans a and b,

$$U_0(a) = U_0(b) \quad \text{implies} \quad U_0(a + k\gamma) = U_0(b + k\gamma) \quad \text{for all } k \in \mathbb{R}.$$

If utility is normalized so that the investor is indifferent between the consumption plan c and the consumption plan $U_0(c)\gamma$, the above property can equivalently be stated as quasilinearity with respect to γ; that is, $U_0(c + k\gamma) = U_0(c) + k$ for any consumption plan c and scalar k. For recursive utility, the latter restriction is essentially equivalent to the BSDE form:

$$\mathrm{d}U_t = -G\left(t, \frac{c_t}{\gamma_t} - U_t, \Sigma_t\right)\mathrm{d}t + \Sigma_t'\,\mathrm{d}B_t, \quad U_T = \frac{c_T}{\gamma_T}, \tag{53}$$

for a possibly state-dependent function G that we call an *absolute aggregator*. For concreteness, we combine the above representation with our earlier formulation of possibly source-dependent risk aversion with a smooth conditional certainty equivalent, resulting in the quasi-quadratic absolute aggregator specification

$$G(t, x, \Sigma) = g(t, x) - \frac{1}{2}\Sigma' Q_t \Sigma. \tag{54}$$

In the remainder of this subsection, we assume this absolute aggregator form, where $g(t, x)$ is strictly increasing, concave, and differentiable in x, the partial derivative $g_x(t, \cdot)$ maps $\mathbb{R}$ onto $\mathbb{R}$, and Q is a (bounded) process valued in the space of positive-definite $d \times d$ matrices.

Example 42 (Expected discounted exponential utility). Let β be any (say bounded) process, and suppose the utility process V of the plan c is well defined by

$$V_t = E_t\left[\int_t^T -\exp\left(-\int_t^s \beta_u\,\mathrm{d}u - \frac{c_s}{\gamma_s}\right)\mathrm{d}s - \exp\left(-\int_t^T \beta_u\,\mathrm{d}u - \frac{c_T}{\gamma_T}\right)\right].$$

[17] In Schroder and Skiadas (2005a) $\varrho = \Gamma\kappa$ was set, without loss in generality, equal to a constant for simplicity of exposition. Their analysis applies essentially unchanged with ϱ stochastic, as assumed here and in Schroder and Skiadas (2005b) (where the exposition is simplified by taking κ to be constant).

Then the ordinally equivalent utility process $U_t = -\log(-V_t)$ solves BSDE (53) with the absolute aggregator (54), where $Q(\omega, t) = 1$ and $g(\omega, t, x) = \beta(\omega, t) - \exp(-x)$.

Analogously to the scale-invariance argument, translation-invariance with respect to γ uncouples the FBSDE of the first-order conditions. Intuitively, if the agent's problem is solved at some information spot at a given financial wealth level, it is also solved at all wealth levels, since the agent can always invest any additional wealth to the γ-fund while preserving optimality.

More specifically, at the optimum, the utility process U, the wealth process W, and the shadow-price-of-wealth process λ, are related by

$$U_t = \frac{1}{\Gamma_t}(Y_t + W_t) \quad \text{and} \quad \lambda_t = \frac{1}{\Gamma_t},$$

where the process Y solves the quadratic BSDE

$$\begin{aligned} \mathrm{d}Y_t &= -\left(e_t + p_t - r_t Y_t + \Sigma_t^{Y\prime} h_t + \frac{1}{2}\Sigma_t^{Y\prime} H_t \Sigma_t^Y\right)\mathrm{d}t + \Sigma_t^{Y\prime}\,\mathrm{d}B_t, \\ Y_T &= e_T, \end{aligned}$$

with

$$\begin{aligned} p &= \Gamma g^*\left(\frac{\gamma}{\Gamma}\right) + \frac{\Gamma}{2}\left(\mu^R - \sigma_t^{R\prime}\sigma_t^R\kappa\right)'\left(\sigma^{R\prime} Q \sigma^R\right)^{-1}\left(\mu^R - \sigma_t^{R\prime}\sigma_t^R\kappa\right), \\ h &= -\sigma^R\kappa - Q\sigma^R\left(\sigma^{R\prime} Q\sigma^R\right)^{-1}\left(\mu^R - \sigma_t^{R\prime}\sigma_t^R\kappa\right), \\ H &= \frac{1}{\Gamma}\left(Q\sigma^R\left(\sigma^{R\prime} Q\sigma^R\right)^{-1}\sigma^{R\prime} Q - Q\right). \end{aligned}$$

The optimal plan trading plan ϕ and consumption plan c can be written as

$$\begin{aligned} &\phi = \phi^0 + U\Gamma\kappa, \quad c = \gamma U + \gamma g_x^{-1}\left(\frac{\gamma}{\Gamma}, \frac{\sigma^R\phi^0 + \Sigma^Y}{\Gamma}\right), \\ &\quad \text{where } \phi^0 = \left(\sigma^{R\prime} Q\sigma^R\right)^{-1}\left[\Gamma\left(\mu^R - \sigma^{R\prime}\sigma^R\kappa\right) - \sigma^{R\prime} Q\Sigma^Y\right]. \end{aligned}$$

Just as with the quadratic BSDE case of the scale-invariant formulation, for a certain class of price dynamics the solution reduces to an ODE system. We refer to Schroder and Skiadas (2005a, 2005b) for examples and extensions (some of which are outlined below).

6.3 *Other directions*

We conclude with a list of selected topics on dynamic portfolio theory and a highly biased small sample of associated references that can be consulted for further leads to a large related literature. Brandt (forthcoming) reviews the econometrics of portfolio choice.

Nonlinear wealth dynamics. Cuoco and Cvitanić (1998), El Karoui et al. (2001), and Schroder and Skiadas (2005b) characterize optimality with wealth dynamics that can allow nonlinearities reflecting, for example, market impact or differential borrowing and lending rates. The last reference includes the extension of this chapter's scale/translation invariance arguments to this case.

Discontinuous information. Merton's original work includes examples of discontinuous information generated by Poisson jumps. The extension of Merton's work to Lévy type processes using the Hamilton–Jacobi–Bellman approach is presented in the monograph by Øksendal and Sulem (2005). This chapter's arguments are extended in Schroder and Skiadas (2005b) so that the filtration is generated by Brownian motions as well as marked point processes. The above references provide links to several other papers on this topic.

Habit formation. Asset pricing models with habit formation include Sundaresan (1989), Constantinides (1990), and Detemple and Zapatero (1991). Duffie and Skiadas (1994) defined recursive utility with habit formation, and computed its gradient density. The latter can be combined with this chapter's state price dynamics to formulate optimality conditions as a FBSDE system. Schroder and Skiadas (2002) showed that, by redefining consumption, a formulation with linear habit formation can be transformed to an equivalent one without habit formation. This technique can be used to mechanically translate this chapter's solutions (assuming either a deterministic short-rate process or complete markets) to corresponding solutions that incorporate linear habit formation. The same argument applies with durability of consumption.

Nontradeable income. We have seen that the optimality conditions given a nontradeable income simplify in the translation-invariant formulation, which implies constant absolute risk aversion. More general models of nontradeable income must deal with a fully coupled FBSDE system. The Merton problem with nontradeable income and additive utility has been analyzed in terms of the Hamilton–Jacobi–Bellman approach by Duffie and Zariphopoulou (1993), Duffie et al. (1997), and Koo (1998). Related theoretical results with nontradeable income and additive utilities include Cuoco (1997), Kramkov and Schachermeyer (1999, 2003), Cvitanić et al. (2001), and Hugonnier and Kramkov (2002).

Endogenous labor supply and retirement. Bodie et al. (1992), Bodie et al. (2004), Dybvig and Liu (2005), Farhi and Panageas (2005), and Liu and Neis (2002), among others, have analyzed the lifetime consumption-portfolio problem with endogenous labor supply and/or retirement. Recursive utility formulations in this area are yet to be developed.

Transaction costs. The Merton analysis has been extended to include proportional transaction costs by Davis and Norman (1990), Shreve and Soner (1994),

Liu and Loewenstein (2002), and others. Grossman and Laroque (1990) and Cuoco and Liu (2000) studied problems in which transaction costs apply to changes in the stock of a durable good. Proportional transaction costs preserve scale invariance, motivating the use of expected discounted power utility in the above papers. Fixed transaction costs on the other hand destroy scale invariance. For this reason existing analytically tractable formulations with fixed transaction costs are based on translation invariance, so far only with additive exponential utility, as in Vayanos (1998) and Liu (2004). Optimality conditions with both proportional and fixed transaction costs with i.i.d. returns are given in Øksendal and Sulem (2002). This is only a small sample of a large literature dealing with some form of transaction costs. I am not aware of any related theoretical results with recursive utility.

Acknowledgements

I am grateful to Mark Schroder for the many years of collaborative research on which this article is based, as well as corrections to this chapter. I also thank Darrell Duffie, Flavio de Andrade, Ali Lazrak, Hong Liu, Jacob Sagi, George Skoulakis, Jeremy Staum, and Jared Williams for valuable feedback. I am responsible for all errors. The latest version of this article can be found at http://www.kellogg.nwu.edu/faculty/skiadas/home.htm.

References

Anderson, E., Hansen, L., Sargent, T. (2000). Robustness, detection and the price of risk. *Working paper*, Department of Economics, University of Chicago.

Arrow, K.J. (1965). *Aspects of the Theory of Risk Bearing*. Yrjo Jahnssonin Saatio, Helsinki.

Arrow, K.J. (1970). *Essays in the Theory of Risk Bearing*. North-Holland, London.

Bally, V., Pages, G. (2002). A quantization algorithm for solving discrete time multidimensional optimal stopping problems. *Bernoulli* 9, 1003–1049.

Bodie, Z., Merton, R.C., Samuelson, W.F. (1992). Labor supply flexibility and portfolio choice in a life cycle model. *Journal of Economic Dynamics and Control* 16, 427–449.

Bodie, Z., Detemple, J.B., Otruba, S., Walter, S. (2004). Optimal consumption-portfolio choices and retirement planning. *Journal of Economic Dynamics and Control* 28, 1115–1148.

Bouchard, B., Elie, R. (2005). Discrete time approximation of decoupled forward–backward SDE with jumps. *Working paper*, LPMA, CNRS, UMR 7599, Université Paris 6 and CREST.

Bouchard, B., Touzi, N. (2004). Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations. *Stochastic Processes and their Applications* 111, 175–206.

Brandt, M.W. (forthcoming). Portfolio choice problems. In: Ait-Sahalia, Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*, Elsevier/North-Holland, New York.

Campbell, J., Viceira, L. (2002). *Strategic Asset Allocation*. Oxford Univ. Press, New York.

Chacko, G., Viceira, L. (2005). Dynamic consumption and Portfolio Choice with Stochastic Volatility in Incomplete Markets. *Review of Financial Studies* 18, 1369–1402.

Chen, Z., Epstein, L. (2002). Ambiguity, risk, and asset returns in continuous time. *Econometrica* 70, 1403–1443.

Cheng, P., Scaillet, O. (2005). Linear-quadratic jump-diffusion modeling with application to stochastic volatility. *Working paper*, HEC Geneva, Switzerland.

Chevance, D. (1997). Numerical methods for backward stochastic differential equations. In: Rogers, L., Talay, D. (Eds.), *Numerical Methods in Finance*. Cambridge Univ. Press, Cambridge, UK, pp. 232–244.

Constantinides, G.M. (1990). Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy* 98, 519–543.

Cox, J., Huang, C.-F. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory* 49, 33–83.

Cuoco, D. (1997). Optimal consumption and equilibrium prices with portfolio constraints and stochastic income. *Journal of Economic Theory* 72, 33–73.

Cuoco, D., Cvitanić, J. (1998). Optimal consumption choices for a large investor. *Journal of Economic Dynamics and Control* 22, 401–436.

Cuoco, D., Liu, H. (2000). Optimal consumption of a divisible durable good. *Journal of Economic Dynamics and Control* 24, 561–613.

Cvitanić, J., Karatzas, I. (1992). Convex duality in constrained portfolio optimization. *The Annals of Applied Probability* 2, 767–818.

Cvitanić, J., Schachermayer, W., Wang, H. (2001). Utility maximization in incomplete markets with random endowment. *Finance and Stochastics* 5, 259–272.

Daniel, K., Hirshleifer, D., Teoh, S.H. (2002). Investor psychology in capital markets: Evidence and policy implications. *Journal of Monetary Economics* 49, 139–209.

Davis, M., Norman, A.R. (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research* 15, 676–713.

Detemple, J., Serrat, A. (2003). Dynamic equilibrium with liquidity constraints. *Review of Financial Studies* 16, 597–629.

Detemple, J., Zapatero, F. (1991). Asset prices in an exchange economy with habit formation. *Econometrica* 59, 1633–1657.

Douglas Jr., J., Ma, J., Protter, P. (1996). Numerical methods for forward–backward stochastic differential equations. *Annals of Applied Probability* 6, 940–968.

Duffie, D. (2001). *Dynamic Asset Pricing Theory*, third ed. Princeton Univ. Press, Princeton, New Jersey.

Duffie, D. (2005). Credit risk modeling with affine processes. *Journal of Banking and Finance* 29, 2751–2802.

Duffie, D., Epstein, L. (1992). Stochastic differential utility. *Econometrica* 60, 353–394.

Duffie, D., Skiadas, C. (1994). Continuous-time security pricing: A utility gradient approach. *Journal of Mathematical Economics* 23, 107–131.

Duffie, D., Zariphopoulou, T. (1993). Optimal investment with undiversifiable income risk. *Mathematical Finance* 3, 135–148.

Duffie, D., Fleming, V., Soner, M., Zariphopoulou, T. (1997). Hedging in incomplete markets with HARA utility. *Journal of Economic Dynamics and Control* 21, 753–782.

Dybvig, P.H., Liu, H. (2005). Lifetime consumption and investment: Retirement and constrained borrowing. *Working paper*, Olin School of Business, Washington University, St. Louis, MO.

El Karoui, N., Jeanblanc-Picquè, M. (1998). Optimization of consumption with labor income. *Finance and Stochastics* 2, 409–440.

El Karoui, N., Mazliak, L. (Eds.) (1997). *Backward Stochastic Differential Equations*. Addison–Wesley/Longman, Essex, UK.

El Karoui, N., Peng, S., Quenez, M.-C. (1997). Backward stochastic differential equations in finance. *Mathematical Finance* 7, 1–71.

El Karoui, N., Peng, S., Quenez, M.-C. (2001). A dynamic maximum principle for the optimization of recursive utilities under constraints. *Annals of Applied Probability* 11, 664–693.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75, 643–669.

Epstein, L. (1992). Behavior under risk: Recent developments in theory and applications. In: Laffont, J.-J. (Ed.), *Advances in Economic Theory*. Cambridge Univ. Press, Cambridge, UK.

Epstein, L., Miao, J. (2003). A two-person dynamic equilibrium under ambiguity. *Journal of Economic Dynamics and Control* 27, 1253–1288.

Epstein, L., Schneider, M. (2003). Recursive multiple priors. *Journal of Economic Theory* 113, 1–31.

Epstein, L., Zin, S. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica* 57, 937–969.

Farhi, E., Panageas, S. (2005). Saving and investing for early retirement: A theoretical analysis. *Working paper*, Department of Economics, MIT, Cambridge, MA.

Fleming, W.H., Soner, H.M. (1993). *Controlled Markov Processes and Viscosity Solutions*. Springer, New York.

Giovannini, A., Weil, P. (1989). Risk aversion and intertemporal substitution in the capital asset pricing model. *NBER working paper No. 2824*, Cambridge, MA.

Gobet, E., Lemor, J., Warin, X. (2005). A regression-based Monte Carlo method to solve backward stochastic differential equations. *Annals of Applied Probability* 15, 2172–2202.

Gollier, C. (2001). *The Economics of Risk and Time*. MIT Press, Cambridge, MA.

Grossman, S.J., Laroque, G. (1990). Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58, 25–51.

Hamadene, S. (1996). Équations différentielles stochastiques rétrogrades : Le cas localement Lipschitzien. *Annales de l'Institut Henri Poincaré* 32, 645–659.

Hansen, L., Sargent, T., Turmuhambetova, G., Williams, N. (2001). Robustness and uncertainty aversion. *Working paper*, Department of Economics, University of Chicago.

He, H., Pagès, H. (1993). Labor income, borrowing constraints and equilibrium asset prices: A duality approach. *Economic Theory* 3, 663–696.

He, H., Pearson, N. (1991). Consumption and portfolio policies with incomplete markets and short-sale constrains: The infinite dimensional case. *Journal of Economic Theory* 54, 259–304.

Hugonnier, J., Kramkov, D. (2002). Optimal investment with random endowments in incomplete markets. *Working paper*, HEC Montreal and Carnegie Mellon University.

Karatzas, I., Shreve, S. (1988). *Brownian Motion and Stochastic Calculus*. Springer, New York.

Karatzas, I., Shreve, S. (1998). *Methods of Mathematical Finance*. Springer, New York.

Karatzas, I., Lehoczky, J., Shreve, S. (1987). Optimal portfolio and consumption decisions for a 'small investor' on a finite horizon. *SIAM Journal of Control and Optimization* 25, 1557–1586.

Karatzas, I., Lehoczky, J., Shreve, S., Xu, G. (1991). Martingale and duality methods for utility maximization in an incomplete market. *SIAM Journal of Control and Optimization* 29, 702–730.

Kim, T., Omberg, E. (1996). Dynamic nonmyopic portfolio behavior. *Review of Financial Studies* 9, 141–162.

Klibanoff, P., Marinacci, M., Mukerji, S. (2002). A smooth model of decision making under ambiguity. *Working paper*, MEDS, Kellogg School of Management, Northwestern University.

Kobylanski, M. (2000). Backward stochastic differential equations and partial differential equations with quadratic growth. *The Annals of Probability* 28, 558–602.

Koo, H. (1998). Nontraded assets in incomplete markets. *Mathematical Finance* 8, 49–65.

Korn, R. (1997). *Optimal Portfolios*. World Scientific, River Edge, NJ.

Kramkov, D., Schachermayer, W. (1999). The asymptotic elasticity of utility functions and optimal investment in incomplete markets. *Annals of Applied Probability* 9, 904–950.

Kramkov, D., Schachermayer, W. (2003). Necessary and sufficient conditions in the problem of optimal investment in incomplete markets. *Annals of Applied Probability* 13, 1504–1516.

Kreps, D., Porteus, E. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.

Lazrak, A., Quenez, M.C. (2003). A generalized stochastic differential utility. *Mathematics of Operations Research* 28, 154–180.

Lemor, J., Gobet, E., Warin, X. (2006). Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations. *Bernoulli* 12, 889–916.

Lepeltier, J.-P., San Martín, J. (1997). Backward stochastic differential equations with continuous coefficient. *Statistics and Probability Letters* 32, 425–430.

Lepeltier, J.-P., San Martín, J. (1998). Existence for BSDE with superlinear–quadratic coefficient. *Stochastics and Stochastics Reports* 63, 227–240.

Lepeltier, J.-P., San Martín, J. (2002). On the existence or non-existence of solutions for certain backward stochastic differential equations. *Bernoulli* 8, 123–137.

Liu, H. (2004). Optimal consumption and investment with transaction costs and multiple risky assets. *Journal of Finance* 59, 289–338.
Liu, J. (2005). Portfolio selection in stochastic environments. *Working paper*, UCLA.
Liu, H., Loewenstein, M. (2002). Optimal portfolio selection with transaction costs and finite horizons. *Review of Financial Studies* 15, 805–835.
Liu, J., Neis, E. (2002). Endogenous retirement and portfolio choice. *Working paper*, UCLA.
Ma, J., Yong, J. (1999). *Forward–Backward Stochastic Differential Equations and Their Applications*. Springer-Verlag, Berlin/Heidelberg.
Ma, J., Protter, P., Yong, J. (1994). Solving forward–backward stochastic differential equations explicitly—A four step scheme. *Probability Theory and Related Fields* 98, 339–359.
Ma, J., Protter, P., San Martin, J., Torres, S. (2002). Numerical method for backward stochastic differential equations. *Annals of Applied Probability* 12, 302–316.
Maenhout, P. (1999). Robust portfolio rules and asset pricing. *Working paper*, INSEAD.
Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7, 77–91.
Merton, R. (1969). Lifetime portfolio selection under uncertainty: The continuous time case. *Review of Economics and Statistics* 51, 247–257.
Merton, R. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3, 373–413;
Erratum *Journal of Economic Theory* 6 (1973) 213–214.
Merton, R. (1990). *Continuous Time Finance*. Blackwell, Malden, MA.
Musiela, M., Zariphopoulou, T. (2004). An example of indifference prices under exponential preferences. *Finance and Stochastics* 8, 229–239.
Narens, L. (1985). *Abstract Measurement Theory*. MIT Press, Cambridge, MA.
Obstfeld, M. (1994). Risk-taking, global diversification, and growth. *American Economic Review* 84, 1310–1329.
Øksendal, B., Sulem, A. (2002). Optimal consumption and portfolio with fixed and proportional transaction costs. *SIAM Journal of Control and Optimization* 40, 1765–1790.
Øksendal, B., Sulem, A. (2005). *Applied Stochastic Control of Jump Diffusions*. Springer, New York.
Pardoux, E., Peng, S. (1990). Adapted solution of a backward stochastic differential equation. *Systems and Control Letters* 14, 55–61.
Piazzesi, M. (2005). *Affine term structure models*. *Handbook of Financial Econometrics*. Elsevier/North-Holland, Amsterdam/New York.
Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica* 32, 122–136.
Schroder, M., Skiadas, C. (1999). Optimal consumption and portfolio selection with stochastic differential utility. *Journal of Economic Theory* 89, 68–126.
Schroder, M., Skiadas, C. (2002). An isomorphism between asset pricing models with and without linear habit formation. *Review of Financial Studies* 15, 1189–1221.
Schroder, M., Skiadas, C. (2003). Optimal lifetime consumption-portfolio strategies under trading constraints and generalized recursive preferences. *Stochastic Processes and Their Applications* 108, 155–202.
Schroder, M., Skiadas, C. (2005a). Lifetime consumption-portfolio choice under trading constraints and nontradeable income. *Stochastic Processes and Their Applications* 115, 1–30.
Schroder, M., Skiadas, C. (2005b). Optimality and state pricing in constrained financial markets with recursive utility under continuous and discontinuous information. *Mathematical Finance*, in press.
Segal, U., Spivak, A. (1990). First-order versus second-order risk aversion. *Journal of Economic Theory* 51, 111–125.
Sethi, S. (1997). *Optimal Consumption and Investment with Bankruptcy*. Kluwer Academic Publishers, Norwell, MA.
Shreve, S., Soner, H.M. (1994). Optimal investment and consumption with transaction costs. *Annals of Applied Probability* 4, 609–692.
Shreve, S., Xu, G. (1992a). A duality method for optimal consumption and investment under short-selling prohibition. I. General market coefficients. *Annals of Applied Probability* 2, 87–112.
Shreve, S., Xu, G. (1992b). A duality method for optimal consumption and investment under short-selling prohibition. II. Constant market coefficients. *Annals of Applied Probability* 2, 314–328.

Skiadas, C. (1998). Recursive utility and preferences for information. *Economic Theory* 12, 293–312.
Skiadas, C. (2003). Robust control and recursive utility. *Finance and Stochastics* 7, 475–489.
Sundaresan, S. (1989). Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2, 73–89.
Svensson, L. (1989). Portfolio choice with non-expected utility in continuous time. *Economic Letters* 30, 313–317.
Svensson, L., Werner, I. (1993). Nontraded assets in incomplete markets: Pricing and portfolio choice. *European Economic Review* 37, 1149–1168.
Uppal, R., Wang, T. (2003). Model misspecification and under-diversification. *Journal of Finance*.
Vayanos, D. (1998). Transaction costs and asset prices: A dynamic equilibrium model. *Review of Financial Studies* 11, 1–58.
Vila, J.-L., Zariphopoulou, T. (1997). Optimal consumption and portfolio choice with borrowing constraints. *Journal of Economic Theory* 77, 402–431.
von Neumann, J., Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton Univ. Press, Princeton, NJ.
Wachter, J. (2002). Portfolio and consumption decisions under mean-reverting returns: An exact solution in complete markets. *Journal of Financial and Quantitative Analysis* 37, 63–91.
Wakker, P.P. (1989). *Additive Representations of Preferences*. Kluwer, Dordrecht.
Yong, J., Zhou, X.Y. (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer-Verlag, New York.
Zariphopoulou, T. (1994). Consumption-investment models with constraints. *SIAM Journal on Control and Optimization* 32, 59–85.
Zariphopoulou, T., Tiu, C.-I. (2002). Optimization in incomplete markets with recursive utility. *Working paper*, University of Texas, Austin.
Zhang, J. (2004). A numerical scheme for BSDEs. *Annals of Applied Probability* 14, 459–488.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15020-9

Chapter 20

Optimization Methods in Dynamic Portfolio Management

John R. Birge

Graduate School of Business, University of Chicago, Chicago, IL 60637, USA
E-mail: john.birge@ChicagoGSB.edu

Abstract

This chapter describes various methods for solving optimal portfolio and asset-liability management models as discrete-time stochastic programs. The focus is on solution methods that consider investment decisions concerning large numbers of alternative assets and liabilities with general constraints, return distributions, transaction costs, and taxes. Each method relies on a form of approximation, which we use for classification. We discuss results concerning the approximations, relative computational advantages of the approaches, and potential areas for future research.

1 Introduction

Dynamic portfolio management arises in virtually every area of financial engineering, e.g., in creating hedges for derivative pricing, in determining capital reserves and credit positions in risk management, and in determining asset allocations to meet streams of liabilities. While continuous-time models provide solution structure and often lead to useful analytical results, practical limitations, such as transaction costs and various constraints, usually necessitate discrete-time models that sacrifice simplifying structure for general application. These general models, as dynamic stochastic programs (or stochastic dynamic programs), then assume some form of approximation to obtain tractable computation. This chapter describes the form of those approximations, the relationship between the approximations and optimization algorithms, and open issues for future research.

The following section provides the general formulation that will appear here. Section 3 considers a variety of approaches to approximation of the problem. Section 4 discusses optimization methodology that applies to the various approximation schemes. Section 5 presents conclusions and areas for future research.

2 Formulation

In this section, we first present a broad general formulation and then provide specifications for the dynamic portfolio management problem. Additional details on this general formulation appear in the textbook, Birge and Louveaux (1997), and the surveys by Dempster (1980) and Wets (1990). This formulation combines state and control variables into a single state vector in contrast to many control-theoretic expositions [see, for example, the discussion in Varaiya and Wets (1989)]. The formulation also assumes that the data process is observable and has a known distribution, again, in contrast to other models, such as Bayesian decision models (see, for example, Berger, 1985), although extensions to Bayesian approaches are direct (and considered in Section 5). While other models can generally be captured through the variable, constraint, or objective structure of the model described below, the emphasis here is (as generally in stochastic programming) on high-dimensional decisions that require significant computational effort for optimization, thereby motivating a focus on relevant optimization methodology.

For this general model, we assume a data process, $\omega := \{\omega_t\colon\ t = 0, 1, 2, \ldots\}$ in a (canonical) probability space (Ω, Σ, μ). We also assume a decision process $x := \{x_t\colon\ t = 0, 1, 2, \ldots\}$ such that x is a measurable function $x\colon\ \omega \mapsto x(\omega)$. As in most stochastic programming situations, we assume the decision process space is the space of essentially bounded functions, $L^n_\infty := L_\infty(\Omega \times \mathbb{N}, \Sigma \times \mathcal{P}(\mathbb{N}), \mu \times \#; \Re^n)$, where $\mathcal{P}$ is the power set and $\#$ is counting measure. [Other spaces are, of course, possible, but spaces such as L_p, $p < \infty$, described, for example, in Eisner and Olsen (1975), are more difficult for defining constraint qualifications, particularly due to their lack of interior points for positive cones.] The norm on this (vector) sequence space is defined by

$$\|x\| := \sup_n \operatorname{ess\,sup} \big|x_n(\omega)\big|.$$

The data process has an associated filtration $\mathbb{F} := \{\Sigma_t\}_{t=0}^\infty$, where $\Sigma_t := \sigma(\omega^t)$ is the σ-field of the history process $\omega^t := \{\omega_0, \ldots, \omega_t\}$ and the Σ_t satisfy $\{0, \Omega\} \subset \Sigma_0 \subset \cdots \subset \Sigma$. The history of the decision process is defined similarly, $x^t = (x_0, \ldots, x_t)$.

A fundamental property of the decision process at time t is that it must only depend on the data up to time t only, i.e., x_t must be Σ_t-measurable, or $\mathbf{x}_t(\omega) = \mathrm{E}\{\mathbf{x}_t(\omega) \mid \Sigma_t\}$ a.s., $t = 0, 1, 2, \ldots$, where $\mathrm{E}\{\cdot \mid \Sigma_t\}$ is conditional expectation with respect to the σ-field Σ_t. In stochastic programming, this condition is called the nonanticipative property (and is also known as implementable or that x_t is Σ_t adapted). We write the nonanticipative condition as a constraint using the projection operator, $\Pi_t : z \mapsto \pi_t z := \mathrm{E}\{z \mid \Sigma_t\},\ t = 0, 1, 2, \ldots$, on L^n_∞, as

$$(I - \Pi_t)x_t = 0, \quad t = 0, 1, 2, \ldots. \tag{2.1}$$

If we let $\mathcal{N}$ denote the closed linear subspace of nonanticipative processes in L^n_∞ and denote by $\Pi := (\Pi_0, \Pi_1, \ldots)$ the projection operator from L^n_∞

onto $\mathcal{N}$, our general optimization model is to find

$$\inf_{x\in\mathcal{N}} \mathrm{E}\sum_{t=0}^{\infty} f_t\big(\omega, x_t(\omega), x_{t+1}(\omega)\big), \tag{2.2}$$

where "E" denotes expectation with respect to the probability measure μ on Σ, assumed completed with respect to μ. For many applications, we might only be concerned with a finite horizon H and would truncate the sum to $t = H$. Following the convention in this literature, we use the notation $\mathbf{x}_t$ and $\mathbf{f}_t$ to denote respectively x_t and f_t as functions of ω, i.e., as random entities. Problem (2.2) then becomes

$$\inf_{\mathbf{x}\in\mathcal{N}} \mathrm{E}\sum_{t=0}^{\infty} \mathbf{f}_t(\mathbf{x}_t, \mathbf{x}_{t+1}), \tag{2.3}$$

where we can then write the objective as $F(\mathbf{x}) := \mathrm{E}\sum_{t=0}^{\infty} \mathbf{f}_t(\mathbf{x}_t, \mathbf{x}_{t+1})$. We assume in (2.3) that the objective components $\mathbf{f}_t$ are proper convex normal integrands (see Rockafellar, 1976).

Most portfolio problems would involve maximization of a concave, utility function in place of the convex objective in (2.3), which only requires a change of sign in the objective. For the general problem, we will continue with minimization and convex integrands again to be consistent with the majority of the literature in this area. In the utility function framework, a question arises over the ability of the time-additive form in (2.3) to satisfy various preference axioms. The objective in (2.3) can, however, be defined to meet a broad range of objectives by appropriate definition of $\mathbf{x}_t$ to include the history process and by allowing $\mathbf{f}_t$ to depend generally on the history of actions and the resolution of uncertainty. In particular, when $\mathbf{f}_t$ includes a product of functions of previous consumption and future wealth, the objective can fit the temporal utility form in Kreps and Porteus (1978). [For a discussion of alternative utility forms in continuous time, see the chapter in this volume by Skiadis (2007).]

For specifying the portfolio management problem, we suppose that x_t includes components $y_t(i)$ for allocations in asset i, $i = 1, \ldots, K$; actions, $b_t(i)$, for the amount bought of asset i; $s_t(i)$, for the amount sold of asset i; and $c_t(j)$, $j = 1, \ldots, m$ for the amount consumed of product or service category j (where, for example, for individuals, consumption utility in a category, such as housing, may depend on an asset allocation, such as residential real estate). The history process would also determine parameters such as $l_t(\omega^t)$, the net external cash flow at t (generally a liability in asset-liability management models); $d_t(i)$, $i = 1, \ldots, K$, the vector of cash dividends (perhaps negative) for each asset i; $r_t(i)$, $i = 1, \ldots, K$, the vector of returns (net dividends) of each asset i, and $\alpha^{\pm}(i)$, $i = 1, \ldots, K$, the vector of transaction costs for buying ($\alpha^{+}(i)$) and selling ($\alpha^{-}(i)$) each asset i. With these definitions and a consumption utility $U_t(c_t, c_{t+1}, \omega^t)$ that depends on previous and current consumption (to model

persistence and habit formation), we would have, for $x_t = (y_t, b_t, s_t, c_t)$,

$$f_t(x_t, x_{t+1}, \omega^t) = \begin{cases} -U_t(c_t, c_{t+1}, \omega^t) \\ \quad \text{if } \operatorname{diag}(r(\omega^t))y_t = y_{t+1} + b_{t+1} - s_{t+1}, \\ \quad (e + \alpha^+)^T b_{t+1} + e^T c_{t+1} + l_t(\omega^t) - d_t(\omega^t)^T y_t \\ \qquad - (e - \alpha^-)^T s_{t+1} = 0, \\ +\infty \quad \text{otherwise.} \end{cases} \tag{2.4}$$

This definition can also be extended to include additional portfolio constraints that might include other risk characteristics (that restrict y_{t+1}), trading restrictions (that may limit b_{t+1} or s_{t+1}), and liability funding restrictions (that may require multiple liability categories including, for example, that certain categories are restricted in some countries as only payable with cash or dividend income). Taxes can also be included, but this generally requires maintaining the tax basis for all assets, which rapidly increases the size of the state space. While allowing for all of these possibilities, we will assume that we can represent f_t as a convex objective function subject to linear constraints:

$$f_t(x_t, x_{t+1}, \omega^t) = \begin{cases} g_t(x_t, x_{t+1}, \omega^t) & \text{if } B_t(\omega^t)x_t + A_{t+1}x_{t+1} \\ & \qquad = h_{t+1}(\omega^t),\ x_{t+1} \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.5}$$

This formulation does not restrict the general classification (since we can still place any other restrictions in the definition of g_t), but it provides a convenient format for several methods discussed below.

For convenience, we may also consider problem (2.3) as a dynamic program. For that format, let $V_t(x_t, \omega^t)$ be the value function in state x_t with history ω^t, defined as:

$$V_t(x_t, \omega^t) = \sup_{x_{t+1}} \left[f_t(x_t, x_{t+1}, \omega^t) + E_{\omega_{t+1}|\omega^t}\left[V_{t+1}(x_{t+1}, \omega^{t+1}) \right] \right], \tag{2.6}$$

with a given terminal value, $V_H(x_H, \omega^H)$.

In some cases, we may also use the implicit formulation of the model in which x_t is defined for each realization of ω^t. In these cases, we would generally assume a finite set of possible history processes at t as scenarios, $1, \ldots, N_t$. Associated with each scenario i at t is an ancestor scenario $a(i)$ at $t-1$ and a set of descendant scenarios, $\mathcal{D}(i)$, at $t+1$. In many cases, it is also convenient to consider a set of relevant parameters, ξ_t, that are determined by ω^t and form a random vector $\boldsymbol{\xi}_t$.

This format fits the classical investment-consumption problem in discrete time [as, for example, in Samuelson (1969)]. Extensions to continuous time (e.g., Merton, 1969) are also possible, but the focus here is on numerical methods that would in turn require some form of discretization. The discussion below pertains to general models in this form. When specific characteristics for portfolio problems are relevant, the notation above will be used.

3 Approximation methods

Problems of form (2.3) are naturally quite computationally intense. From the computational complexity view, for example, these problems are PSPACE-hard (Dyer and Stougie, 2006) and require exponential effort in the horizon H for provably tight approximations with high probability (Swamy and Shmoys, 2005 and Shmoys and Swamy, 2006). Even the common-case of a two-stage ($H = 2$) problem with the common mean-variance objective is NP-hard (Ahmed, 2006). Worst-case results can, therefore, be quite disappointing, but specific forms, such as the portfolio representation in (2.4), may offer opportunities for reliable approximation and efficient computation. Approximation is generally required in any event and forms the theme of this section.

This variety of approximation approaches considered here are as follows.

1. *Time, state, and path aggregation or scenario generation and reduction*: These methods essentially start with a large (often continuous) set of possibilities and then combine (or select) them to form more tractable representations;
2. *Value function approximation*: These methods focus on the form of the value function V_t with some simplified representation;
3. *Policy restriction*: These approaches restrict the set of alternative controls to a simplified form that allows for efficient computation;
4. *Constraint relaxation and dualization*: These approaches relax constraints or look at dual forms, generally not guaranteeing implementable policies but perhaps giving bounds or guidelines for implementable policies;
5. *Monte Carlo methods*: These methods rely on sampling results and can often be applied to any of the previous methods as well.

3.1 Time, path, and state aggregation or scenario generation and reduction

Approximation of (2.3) can take many forms. The fixing in time of decisions (that may start as continuous controls) represents one form of approximation that can be viewed as *aggregation* across time. Aggregation can also be applied to states [especially in using the dynamic programming view in (2.6)]. Stochastic programs of this form also often involve the representation of sample paths or scenarios of potential outcomes that require sample selection or aggregation to derive tractable formulations. This section discusses each of these forms of problem reduction as used in dynamic portfolio optimization.

3.1.1 Time aggregation

Aggregation of time periods involves replacing k periods of decisions $s, \ldots, s + k$ with a single period decision, where $x_{s+1}, \ldots, x_{s+k+1}$ is replaced by an aggregate decision X_{s+1} and the objective sum, $\sum_{i=s+1}^{s+k} f_i(x_i, x_{i+1})$, is replaced by $F_s(X_s, X_{s+1})$. The overall approach then includes an aggregation step, $(x_{s+1}, \ldots, x_{s+k+1}) \to X_s$ and $\sum_{i=s+1}^{s+k} f_i(x_i, x_{i+1}) \to F_s(X_s, X_{s+1})$, followed potentially by disaggregation, $X_s \to (x_{s+1}, \ldots, x_{s+k+1})$, with objective

measurement using the original objective function with the disaggregated solution (or some bound on that value).

Various results provide bounds on optimal solutions from the solutions of the aggregated and disaggregated solutions. The bounds either use duality results or bounds on state-to-state transitions (generally in a setting with a finite number of states and actions). In the general linear constraint case, the duality-based bounds rely on known penalties for constraint violation or bounds on the associated Lagrange multipliers for each constraint. Typically, X_s represents an average over the lower-level decision vectors, $X_s = \sum_{l=s+1}^{s+k+1} x_l/k$ and the objective is a sum or multiple of the original objective values, e.g., $F_s(X_s, X_{s+1}) = f_s(X_s, X_{s+1}) + \sum_{l=s+1}^{s+k} f_l(X_s, X_{s+1})$.

Linear constraints as in (2.5) can also be aggregated directly to produce an aggregate constraint in defining F_s. For example, when the objective in period s includes a discount factor, ρ^s, the set of constraints can be aggregated using a discount factor weighting (see, e.g., Grinold, 1986 and Birge, 1985a). The linear constraints, $B_l x_l + A_{l+1} x_{l+1} = h_{l+1}$ for $l = s+1, \ldots, s+k+1$, are replaced by $\tilde{B}_s X_s + \tilde{A}_{s+1} X_{s+1} = \tilde{h}_{s+1}$, where $\tilde{B}_s = \sum_{l=s}^{s+k} \rho^{l-s} B_l$, $\tilde{A}_{s+1} = \sum_{l=s+1}^{s+k+1} \rho^{l-s-1} A_l$, and $\tilde{h}_{s+1} = \sum_{l=s+1}^{s+k+1} \rho^{l-s-1} h_l$. This form of aggregation is especially useful in controlling for the end effects of truncating a horizon for a long or infinite-horizon problem. With assumptions on the form of the objective, bounds arise. For convex objectives, for example, an averaging aggregation implies by Jensen's inequality that the solution of the aggregated problem yields a lower bound on $V_0(x_0)$. Maximum penalties for constraint violations or other objective function properties produce dual formulations with bounded multipliers that can yield upper bounds on $V_0(x_0)$. Descriptions of these bounding procedures appear in Birge (1985a), Wright (1994), and Kuhn (2005).

3.1.2 Path aggregation

In conversions from continuous-time to discrete-time models, time aggregation is often referred to as time discretization. Implementation also generally requires discretization or aggregation of sample paths. This process can involve Monte Carlo sampling as described in Section 3.5 below, or deterministic selection of sample paths, as in quasi-Monte Carlo methods (see, e.g., Niederreiter, 1978 and Glasserman, 2004), or bounding approximations (see, e.g., Birge and Wets, 1986 and Birge and Louveaux, 1997) that may also be derived from convexity and duality results as in the time aggregation process given above (Birge, 1985a; Kuhn, 2005).

Lower bounds (for minimization problems) again result from Jensen-like inequalities on the expectation of convex functions of random variables when path aggregation corresponds to expectations (and properly weighted conditional expectations) of original paths. For example, if $x_t^*(\omega)$ is an optimal solution to (2.2) and the aggregation at time t corresponds to a partition of Σ_t into N_t subsets, $S_t(1), \ldots, S_t(N_t)$, such that $p_t(i) = \text{Prob}\{S_t(i)\}$, $x_t(i) =$

$E[x_t^*(\omega) \mid \omega \in S_t(i)]$, and f_t is a convex function of random parameters $\xi_t(\omega)$ such that $E[\xi_t(\omega) \mid \omega \in S_t(i)] = \xi_t(i)$, then $E_{\Sigma_t}[f_{t-1}(x_{t-1}, x_t(\omega), \xi_t(\omega))x] \geqslant \sum_i^{N_t} p_t(i) f_{t-1}(x_{t-1}, x_t(i), \xi_t(i))$ for any x_{t-1}. Applying this inequality sequentially then yields that conditional expectations of optimal $\mathbf{x}_t^*$ produce a lower objective that is then feasible in a path-aggregated formulation. Optimizing that aggregate formulation then produces a lesser lower bound.[1]

Finite upper bounds (for minimization objectives) are possible in this case as well when the solutions and random parameters, or the objective gradient (and constraint violation penalties) are bounded (i.e., grow at most linearly at a known rate beyond some point). Relaxations of this assumption are also possible but require higher moment bounds, since a higher-order penalty function may lead only to trivial bounds otherwise. The general idea is again to use convexity properties (or equivalently duality to construct feasible solutions to a dual problem). A basic useful result for these bounds is the following that appears in Birge and Wets (1986) and Birge and Louveaux (1997).

Theorem 3.1. *Suppose that $\xi \mapsto g(x, \xi)$ is convex and Ξ is compact. For all $\xi \in \Xi$, let $\phi(\boldsymbol{\xi}, \cdot)$ be a probability measure on* $(\operatorname{ext} \Xi, \mathcal{E})$*, such that*

$$\int_{e \in \operatorname{ext} \Xi} e\phi(\boldsymbol{\xi}, \mathrm{d}e) = \boldsymbol{\xi}, \tag{3.7}$$

and $\omega \mapsto \phi(\xi(\omega), A)$ is measurable for all $A \in \mathcal{E}$; then

$$E\big(g(x)\big) \leqslant \int_{e \in \operatorname{ext} \Xi} g(x, e)\lambda(\mathrm{d}e), \tag{3.8}$$

where λ is the probability measure on $\mathcal{E}$ defined by

$$\lambda(A) = \int_{\Omega} \phi\big(\xi(\omega), A\big)P(d\omega). \tag{3.9}$$

This result is extended to noncompact Ξ by also considering the value of the objective g along extreme directions of Ξ. The bounding problem results from substituting the original measure on $\Xi_t(\omega)$, the random vector in period t, with the measure λ on the extreme points (and directions) of Ξ_t in each period t. This approach can also be employed on some partition of Ξ_t, resulting in an improving bounding approximation. Partitions, such as simplices, that maintain a manageable set of extremal values are particularly efficient [see, e.g., the *barycentric* methods in Frauendorfer (1992)].

[1] This form of aggregation also appears in revenue management models where the approximate x_t represents the vector of expected allocations to different customer classes and is constrained by both expected future demand and an overall capacity constraint. The resulting formulation is known as bid–price control (Talluri and Van Ryzin, 1998) and can also be interpreted as a linear value function approximation (Adelman, 2007).

Bounds of this form also can be viewed as general semi-parametric schemes that assume some limited moment information about the underlying distributions and then construct solutions to the resulting moment problems to obtain bounds in the form of limiting distributions (see, e.g., Birge and Wets, 1986). These forms of extreme value and conditional expectation substitutions for path aggregation then lead to deterministic approximations which produce bounds on the overall value of the original formulation in (2.2). The tightness of these bounds generally depends on the relative degree of nonlinearity of the value function. Functions with close to linear behavior yield tighter approximations while significant nonlinear response yields quite loose approximation.

3.1.3 State aggregation

A related approach to the path and time aggregation approaches mentioned above is to aggregate states directly in the dynamic programming formulation (2.6). The most basic approach is to assume $x_t \in X_t$, to form some partition of X_t into N_t subsets and to assign $x_t(i)$, $i = 1, \ldots, N_t$ for each of the subsets. This general approach can be given the same interpretation as the path aggregation approaches mentioned above and may lead to similar overall bounding results.

Other possibilities for state reduction in the context of the portfolio optimization are possible when the optimal policy only depends on a lower dimensional representation of the general state space. For example, in the case without transaction costs, the policy may only depend on wealth. In that case, the dynamic programming form (2.6) is reduced from K dimensions of the variable y to a single dimension. In a more realistic setting with transaction costs, if the objective function is independent of wealth, the state may be reduced to the proportion held in each asset class [as, for example, in Papi and Sbaraglia (2006)]. This reduction by a single dimension is, however, only useful when the dimension is already quite low. In more general cases, state aggregation involves some loss of optimality.

Other bounds based on the dynamic programming state aggregation are possible using bounds on the state-to-state transition objective contributions (see, e.g., Bean et al., 1987 and Shen and Caines, 2002). Other state aggregation methods rely on forms of interpolation between explicit states that represent grid points for the approximation. Bounds for this form of approximation generally use properties of the value function (such as derivative bounds) that may not hold for constrained problems (as here) where derivatives are not usually continuous.

3.1.4 Scenario generation and reduction

The general approaches given in the previous sub-sections relate to state and path reductions that enable bounding approximations. In many cases, obtaining deterministic bounds of this type, particularly in very high dimensional problems, may lead either to intractably large problems or excessively loose bounds. Other procedures to generate representative samples (scenarios) of

sample paths are often more heuristic in nature (without provable deterministic error estimates), but still yield empirically effective results. This section describes some of these basic approaches.

The general term for this approach is scenario generation, which refers to generating a (finite) set of sample paths that form the basis of a tractable formulation of (2.2). These scenarios may also have some statistical representation (that is described later), but, in many implementations, are found by a deterministic process meant to produce a simplified representation of some unknown distribution.

This general approach for designing trees of scenarios appears in papers including Høyland and Wallace (2001), Kouwenberg (2001), Gondzio and Kouwenberg (2001), and Pflug (2001). These approaches generally advocate fitting moments of the distributions or, in the case of Pflug, a transportation metric on the distance between the underlying and approximate distribution. The result is effectively fitting the trees used for solution to the overall dynamic distributions. In portfolio optimization problems for asset-liability management, scenario generation requires some care to avoid creating arbitrage opportunities within the resulting trees. Strategies to avoid these problems appear in Klaassen (1998) and the constrained optimization approach presented by Pflug (2001).

These general approaches can also be augmented by scenario reduction techniques that start with large trees and attempt to find close trees with fewer branches using different metrics (such as the transportation metric) on the distance between distributions. These approaches appear, e.g., in Dupačová et al. (2003) and Heitsch and Römisch (2003). They also have some justification in sensitivity results that bound the distance between optimal solutions of stochastic programs with differing underlying distributions [see, e.g., Römisch (2003) for an overview and Römisch and Wets (2006) for recent results with convex objectives]. General extensions of these results for the multi-stage form in (2.2) are the subject of ongoing research.

3.2 Value function approximation

Many of the approaches to approximating the dynamic programming form in (2.6) amount to approximations of the value function V_t. These approaches may involve approximation by a set of basis functions (e.g., separable and piecewise linear as in Tsitsiklis and Van Roy, 1997, 2001; Birge and Louveaux, 1997, Chapter 11.3; Frantzeskakis and Powell, 1993), by general spline approximations as described in Trick and Zin (1997) and Judd (1998), and by outer linearizations (e.g., Louveaux, 1980; Birge, 1985a, 1985b; Birge and Rosa, 1996). The latter approaches rely on convexity properties of the objective function.

For the general portfolio problems, convexity of the objective in the state variables generally follows from assumptions on the utility and the constraint representation. For example, convex transaction costs and von Neumann–Morgenstern utilities yield a convex form of objective in (2.4). The form of

U_t in this case can be quite general (as long as it is concave). In practice, the specific functional form appears to have relatively little overall impact on the portfolio solution as long as the relative risk aversion in the utility is captured adequately (Kallberg and Ziemba, 1983).

The general value function approximation with linear basis functions approximates $V_t(x_t, \omega^t)$ by $\hat{V}_t(x_t, \omega^t) = \sum_{j=1}^{p} \phi_t^j(x_t, \xi_t(\omega^t))$, where ϕ_t^j has a simple structure, such as separable linear or piecewise linear, that allows for efficient computation and integration. The approaches build on some discrete representation of ξ_t, use Monte Carlo methods to generate samples from $\boldsymbol{\xi}_t$, or assume a functional relationship (especially separability) in ξ_t to obtain the expectation of $\hat{V}_t$. Lower and upper bounds are again possible using similar arguments to those for path aggregation.

Outer approximation or cutting plane methods use local information of V_t to produce global approximations under the convexity assumption. The general approach is to assume a lower-bounding convex approximation $V_{t+1}^l(x_{t+1}, \omega^{t+1}) \leqslant V_{t+1}(x_{t+1}, \omega^{t+1})$ to solve for a given x_t^k and ω^t to find:

$$V_t^l(x_t^k, \omega^t) = \inf_{x_{t+1}} \left[f_t(x_t^k, x_{t+1}, \omega^t) + E_{\omega_{t+1}|\omega^t} \left[V_{t+1}^l(x_{t+1}, \omega^{t+1}) \right] \right]. \tag{3.10}$$

Using convexity, this yields a global approximation for all x_t given ω^t such that

$$V_t(x_t, \omega^t) \geqslant V_t^l(x_t^k, \omega^t) + \eta_t^{kT}(x_t - x_t^k), \tag{3.11}$$

where η_t^k is a subgradient of V^l at x_t^k given ω^t. In the case of serial independence (where the distribution of ω_{t+1} does not depend on ω^t as is often assumed for random elements determining return distributions), solving for $V_t^l(x_t^k)$ is independent of ω^t and yields a global bound for all ω^t.

The bounding result in (3.11) is the basis for the nested decomposition method (e.g., Birge, 1985b and Birge and Louveaux, 1997, Chapter 11) that is used widely for portfolio optimization and asset-liability management models (e.g., Cariño et al., 1994 and Dempster et al., 2003). The method converges by generating increasingly tight bounds V_t^l whenever the updated solution to (3.11) improves on the previous value of $V_t^l(x_t^k)$. If no improvement occurs, then another x_t^k can be chosen or the bound can be tested at $t-1$ (or $t+1$). When no improvement is possible at any t, the method has converged.

3.3 Policy restriction

Another approach that is often useful for approximations is to limit the set of possible actions or policies x_t that may be taken at each stage of the process. In asset-liability management models, this may, for example, represent having a default borrowing or investment strategy that is taken whenever net cash flows are nonzero. The result of this approach then is that a given investment

strategy is followed regardless of actual cash flow realizations and then penalties are applied using the default policies. This approach then can produce a separable response function (upper bounding for minimization objectives) as described for value function approximation.

An example in asset-liability management appears in Kusy and Ziemba (1986). The result is that x_1 represents a target investment profile (in this case, targets for bank deposit and loan holdings) and all future uncertainty is represented in a second-stage with short-term borrowing and investment strategies carried out through all subsequent periods. This produces a two-stage model for which efficient solution methods apply.

This approach also appears in Dempster and Thompson (2002) who also show how to derive general policy rules that can be tested on large-scales samples of a full multi-period model. This general idea also allows for a limited number of full-scale optimization or re-balancing points as a form of partial time aggregation. Restricted policies at intermediate times between re-balancing periods then allow for capturing of the model dynamics.

Other restricted policy methods include fixed mix optimization schemes that build on the continuous time stationary solution in which fixed proportions of assets are held in each asset category (as in Merton, 1969). Without transaction costs and nonstationary dynamics, such solutions may be optimal. Finding the best fixed-mix allocation in general, however, becomes a nonconvex optimization problem, which may be amenable to global optimization methods (see, e.g., Maranas et al., 1997).

The use of continuous-time optimization approaches as a guide also appears in other approaches, such as Davis and Norman (1990), who use the result with proportional transaction costs that an optimal portfolio policy consists of a no-trade region around an optimal proportional allocation and that, whenever the portfolio values reach the boundary of the no-trade region, the portfolio should be re-balanced to the optimal proportional allocation. In very low dimensions, this approach is computational tractable, but the general discovery of the no-trade region boundaries becomes quite complex. Recent results using Monte Carlo methods combined with boundary estimates (e.g., Muthuraman and Zha, 2006), however, offer some promise in this direction.

Other possibilities for generalizing continuous-time results include relaxing the form of the transaction costs. Morton and Pliska (1995), for example, show that assuming a proportion of the entire portfolio value is lost with each transaction (instead of a proportion of the amount traded as in reality) yields a computable solution with an ellipsoidal region around the no-transaction-cost optimum (or Merton point) that is asymptotically optimal (Atkinson and Wilmott, 1995). Korn (2004) shows that this methodology can also be used to bound actual proportional transaction costs and proposes an approximation using the Morton–Pliska method.

3.4 Constraint relaxation and dualization

As noted earlier, methods that allow for constraint relaxation (and penalization) can often produce upper bounds on an original maximization objective problem. Most of these relaxations focus on relaxation of allocation bounds (e.g., no short positions), incomplete markets, elimination of transaction costs, and other complications that invalidate general continuous-time solutions. The approaches in Haugh et al. (2006) and Haugh and Kogan (2007, this volume) represent an example of this procedure that relies on using a dual solution that corresponds to using multipliers on the relevant constraints. By obtaining dual feasibility, these methods obtain lower bounding solutions. A key to their efficient implementation is that the relaxed problem and the generation of feasible dual solutions are solved simply, for example, using structural results from the continuous time solution. These methods are effectively a type of Lagrangian method, in which relaxed constraints appear with a suitable multiplier in the objective, that is described in the next section.

3.5 Monte Carlo methods

Most of the approaches given above have natural extensions that include Monte Carlo methods. The solution of the sampled problem and statistical properties of bounding estimates appears in Blomvall and Shapiro (2006). Value function approximation based on polynomial basis functions and Monte Carlo methods are the basis of the approach in Brandt et al. (2005) and van Binsbergen and Brandt (2006). Direct solution of the continuous-time formulation also forms the basis for the methods considered in Detemple et al. (2007, this volume). Since these methods depend highly on specific problem structure, we will not describe them in more detail here, although their use of continuous-time solution structure could be valuable for more general problem solutions as well.

Procedures to incorporate Monte Carlo sampling into outer approximation methods appear in Pereira and Pinto (1991), Dantzig and Infanger (1991), Higle and Sen (1996), and Donohue and Birge (2006). In general, the Monte Carlo methods rely on asymptotic properties of the sampled problems (as in Blomvall and Shapiro, 2006). If a solution x^N is generated with N samples of paths ω^H in (2.2), then the general result is that $x^N \to D(x^*)$ where $D(x^*)$ is a suitably defined distribution around an optimal solution x^*. The rate of this convergence depends on both N and H through the number of distinct branches at each time $t = 1, \ldots, H$. In the case of serial independence, the bias of the number of stages can be reduced.

The methods in Pereira and Pinto (1991) and Donohue and Birge (2006) also use an assumption of serial independence to allow for efficient solution. These methods assume that a lower bounding approximation is available V_t^l by solving an exhaustive sample of the next period stochastic parameters ξ_{t+1} beginning with $t + 1 = H$. Since the lower bounding approach with convex

objective produces global approximations, this procedure can be continued through all stages t producing lower bounds for any selection of states x_t. The resulting lower bounding approximations are then used to produce upper bounding forward estimates by sampling full sample paths and using the resulting objective values. The methods terminate when the upper bounding forward sample and lower bound are sufficiently close.

4 Solution methods

The previous section described various methods for constructing approximations of the general optimization problem in (2.2). This section describes various optimization procedures for solving the overall optimization problem that use the structure of the problem above. The methods include general active set methods, interior point methods, decomposition methods, and Lagrangian methods.

4.1 Active set methods

The goal in active set methods is to take advantage of sparsity structure of the matrices generated by the optimality conditions for a fixed active set of constraints. These approaches have been applied to linear versions of (2.2) (e.g., Kall, 1979; Strazicky, 1980; Birge, 1985b). Specialized linear algebra procedures for the structure in (2.2) can lead to computational efficiencies, but many commercial optimization codes currently include efficient linear algebra techniques that achieve similar performance gains.

4.1.1 Interior point methods

Interior point methods point methods include linear algebra operations that require solution with differently structured (from an active set basis) matrices that may be dense in direct application. The factorization scheme given by Birge and Qi (1988) avoids the dense matrices from direct approaches and achieves a polynomial complexity result that increases linearly in the number of samples. Large-scale implementations with interior point methods appear in Yang and Zenios (1997) and Czyzyk et al. (1995). Other possibilities include methods using a symmetric indefinite, augmented system as in Berger et al. (1995).

4.2 Decomposition approaches

The general idea behind decomposition methods is the use of the outer approximations given above. This general approach began as a method called L-shaped by Van Slyke and Wets (1969) that is a form of Benders' (1962) decomposition. That approach is essentially a Dantzig–Wolfe decomposition (inner linearization) (1960) of the dual to the linear two-stage form of (2.2). As

noted, the method easily generalizes to multiple stages (Birge, 1985b), where it is known as the nested L-shaped or Benders' decomposition method.

4.3 Lagrangian-based approaches

Lagrangian-based methods are essentially relaxation strategies that take complicating constraints into the objective with a multiplier that is then adjusted to obtain a dual optimal (or approximate) solution. The complicating constraints can be those associated with nonanticipativity, with state constraints (such as short sales as noted above for the relaxation in Haugh et al., 2006), or any combination that allows rapid solution of the relaxed problem. For the case of relaxing state constraints, for example, continuous-time optimal solutions are often easily derived. In the case of relaxing nonanticipativity constraints, the optimization can decompose into problems for separate sample paths, enabling a combination of simulation of separate sample paths and optimization over each individual path.

If the nonanticipativity constraints are relaxed, then the solution reduces to separate problems for each realization of the underlying parameters, $\boldsymbol{\xi}$. To see how the procedure develops, assume a multiplier, π [defined on an appropriate dual space as in, for example, Rockafellar and Wets (1976)], to obtain a dual problem to (2.2):

$$\max_{\pi(\xi)} w = \theta(\pi), \tag{4.12}$$

where

$$\theta(\pi) = \inf_{x\in X} z = \mathrm{E}\left[\sum_{t=0}^{H} f_{t+1}\big(x^t(\boldsymbol{\xi}), x_{t+1}(\boldsymbol{\xi}), \boldsymbol{\xi}\big)\right] + \mathrm{E}\left[\sum_{t=0}^{H} \pi^t(\boldsymbol{\xi})(I - \Pi_t)x^t(\boldsymbol{\xi})\right], \tag{4.13}$$

where X represents all constraints in $X_t(\boldsymbol{\xi})$ and π^t corresponds to the first t period components of π.

The general idea in these methods is to ascend in the dual to a maximum, which, under appropriate regularity conditions, corresponds to a minimum in the primal. Assuming that (4.12) always has a unique solution, a basic method is the following (where we assume a finite number N of sample paths and $\Pi_{\mathcal{N}}$ is projection on the nonanticipative subspace $\mathcal{N}$).

Lagrangian dual ascent method

- Step 0. Set π^0, $\nu = 0$ and go to Step 1.
- Step 1. Given $\pi = \pi^\nu$ in (4.12), find a solution, $(x_1^\nu, \ldots, x_N^\nu)$.
- Step 2. If $x_k^\nu - \Pi_{\mathcal{N}} x^\nu = 0$, $k = 1, \ldots, N$, stop (with optimality); otherwise, let $\hat{\pi}_k = x_k^\nu - \Pi_{\mathcal{N}} x^\nu$ and go to Step 3.

- Step 3. Let λ^ν minimize $\theta(\pi^\nu + \lambda\hat{\pi})$ over $\pi^\nu + \lambda\hat{\pi} \geqslant 0, \lambda \geqslant 0$. Let $\pi^{\nu+1} = \pi^\nu + \lambda^\nu\hat{\pi}$, $\nu = \nu + 1$, and go to Step 1.

With the unique solution assumption, this method always produces an ascent direction in θ. The algorithm either converges finitely to an optimal solution or produces an infinite sequence with all limit points optimal assuming a bounded set of optima. When (4.12) has multiple optima, a nondifferentiable procedure (i.e., subgradient method) must be used. In this case, the maximum norm subgradient assures ascent or various bundle type methods (see, e.g., Lemaréchal, 1978 and Kiwiel, 1983) are possible.

For computational efficiency in the dual ascent procedure, the number of dual iterations must be small compared to the number of function evaluations that might be required by directly solving (2.2). Time may be saved by operating on the dual instead of the primal (by avoiding the linking constraints), but many iterations might be required. Since this method uses a single-point linearization of θ that may slow convergence, other Lagrangian approaches to (2.2) use more global or at least second-order information.

Rockafellar and Wets (1986) suggest a procedure that applies to a special case of (2.2) in which f_0 is a convex quadratic function over a convex region and f_1 is a quadratic function subject to linear constraints. In a general augmented Lagrangian approach (see, e.g., Bertsekas, 1982) for this problem, a penalty $\frac{r}{2}\|x_k - \Pi_{\mathcal{N}}x\|^2$ is added to each term k of $\theta(\pi)$ and iterations include a fixed step size such that $\pi_k^{\nu+1} = \pi_k^\nu + rp_k(x_k - \Pi_{\mathcal{N}}x)$. An advantage (as noted in Dempster, 1988) of this approach is that it allows Newton type steps by maintaining a nonsingular Hessian and achieving an improved convergence rate.

Dempster suggests adding a new variable, x_0, substituting for $\Pi_{\mathcal{N}}x$ to solve for:

$$\hat{\theta}(\pi) = \min_{x,y} f_0(x_0) + \sum_{k=1}^{N} p_k \Big[f_1(x_{1k}, x_{2k}, \xi_k) + \pi_k^\nu \cdot (x_{1k} - x_0) + \frac{r}{2}\|x_{1k} - x_0\|^2 \Big]. \tag{4.14}$$

In this approach, iterations alternate between searches in x_0 and then separable optimizations for each k. In this way, the augmented Lagrangian approach of solving (4.14) achieves improved overall convergence.

This method is similar to the progressive hedging algorithm (PHA) of Rockafellar and Wets (1991), which achieves full separation of the separate scenario problems for each iteration, resulting in considerably less work per iteration but perhaps more iterations. PHA offers computational advantages particularly for structured problems (see, e.g., Mulvey and Vladimirou, 1991). A related approach, the extension of the row action algorithm of Censor and Lent (1981) by Nielsen and Zenios (1993a, 1993b), also has particular efficiencies for network constraints, which appear in portfolio problems. The key to

these methods is that individual subproblem structure is maintained throughout the algorithm.

The general structure of the augmented Lagrangian methods and PHA allows for a variety of conditions on the objective functions (and implicitly defined constraints). Augmented Lagrangian methods also may apply to problem with integer variables as well (e.g., for fixed transaction costs or limits on the number of nonzero positions). In these cases, PHA may not converge (see Takriti et al., 1996), but PHA may still obtain good solutions quite efficiently.

Mulvey and Ruszczyński (1995) also develop a variant of the augmented Lagrangian method called diagonal quadratic approximation (DQA) that fixes nonseparable terms to allow separation. Their approach places the nonanticipativity constraints in a permutation order, $\sigma_k, k = 1, \ldots, N$, as $x_{1k} - x_{1\sigma(k)}, k = 1, \ldots, N$ in the two-stage case and approximates the $\|x_{1k} - x_{1\sigma(k)}\|^2$ terms in the objective with a current iterate, $\hat{x}_{1k}$. The result is that the original augmented Lagrangian problem then decomposes again into separate subproblems (allowing parallel computation) for each k. For two stages, the formulation is the following:

$$\begin{aligned} \inf z_k = {} & f_0(x_{1k}) + f_1(x_{1k}, x_{2k}, \xi_k) + (\pi_k - \pi_{\sigma^{-1}(k)}) \cdot (x_{1k}) \\ & + \frac{r}{2}\left[\|x_{1k} - \hat{x}_{1\sigma(k)}\|^2 + \|x_{1k} - \hat{x}_{1\sigma^{-1}(k)}\|^2\right], \end{aligned} \tag{4.15}$$

where $\sigma^{-1}(k)$ refers to the scenario j such that $\sigma(j) = k$.

5 Extensions and conclusions

This chapter has focused on methods that apply to general portfolio problems that appear in contexts such as asset-liability management. The treatment in the approaches here generally involves situations in which distributions are known but Bayesian assumptions, such as those in Pástor (2000), Pástor and Stambaugh (2000), can readily be incorporated in the approaches above, as long as the expected objective functional remains convex in the actions in x_t. If the actions affect learning (e.g., when the price of a nonmarket asset may not be known until it is purchased), then the convexity relationship may no longer hold, causing complications for the optimization procedures here.

Other forms of distributional assumptions include max–min (or min–max) objectives or *robust optimization* procedures (e.g., Goldfarb and Iyengar, 2003; Tütüncü and Koenig, 2004; and Garlappi et al., 2007) and optimization methods built on robust estimation (DeMiguel and Nogales, 2006). These approaches have generally been applied to static optimization models (or applied myopically in a dynamic environment), although the principles can be applied in the general setting of dynamic portfolio optimization as well. In robust optimization, the objective is effectively modified to represent a extremum over a set of distributions as in a max–min utility representation (e.g., Ghilboa and Schmeidler, 1989). The extreme case of this is that any set of distributions with

the same support are considered. Robust estimators also incorporate an inner optimization to obtain an implied distribution that fits observations while not allowing contamination to dominate the estimation. These methods can be incorporated into the optimization procedures above, but again with the caveat on convexity preservation.

The discussion in this chapter is to show that a variety of approximation techniques and computational methods can be applied to dynamic portfolio optimization with general constraints and objectives. The methods builds on basic principles in representing sample path distributions, the relationship between distributions and optimal values, the effect of restriction and relaxation, and the use of problem structure in optimization. Many opportunities exist for further results that relate continuous-time solutions to their discrete-time counterparts, that consider the effects of estimation and model uncertainty on optimization, and that adapt optimization procedures to distribution representation and estimation.

Acknowledgements

This work was supported in part by the National Science Foundation under Grants DMI-0200429 and 0422937 and by The University of Chicago Graduate School of Business.

References

Adelman, D. (2007). Dynamic bid prices in revenue management. *Operations Research* 55 (4), in press.

Ahmed, S. (2006). Convexity and decomposition of mean-risk stochastic programs. *Mathematical Programming, Series A* 106, 433–446.

Atkinson, C., Wilmott, P. (1995). Portfolio management with transaction costs: An asymptotic analysis of the Morton and Pliska model. *Mathematical Finance* 5, 357–367.

Bean, J.C., Birge, J.R., Smith, R.L. (1987). Aggregation in dynamic programming. *Operations Research* 35, 215–220.

Benders, J.F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik* 4, 238–252.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Berger, A.J., Mulvey, J.M., Rothberg, E., Vanderbei, R.J. (1995). Solving multistage stochastic programs using tree dissection, *Technical Report SOR 92-5*, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ.

Bertsekas, D.P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York.

Birge, J.R. (1985a). Aggregation in stochastic linear programming. *Mathematical Programming* 31, 25–41.

Birge, J.R. (1985b). Decomposition and partitioning methods for multi-stage stochastic linear programs. *Operations Research* 33, 989–1007.

Birge, J.R., Louveaux, F. (1997). *Introduction to Stochastic Programming*. Springer, New York.

Birge, J.R., Qi, L. (1988). Computing block-angular Karmarkar projections with applications to stochastic programming. *Management Science* 34, 1472–1479.

Birge, J.R., Rosa, C.H. (1996). Parallel decomposition of large-scale stochastic nonlinear programs. *Annals of Operations Research* 64, 39–65.

Birge, J.R., Wets, R.J.-B. (1986). Designing approximation schemes for stochastic optimization problems, in particular, for stochastic programs with recourse. *Mathematical Programming Study* 27, 54–102.

Blomvall, J., Shapiro, A. (2006). Solving multistage asset investment problems by the sample average approximation method. *Mathematical Programming, Series B* 108, 571–595.

Brandt, M.W., Goyal, A., Santa-Clara, P., Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *Review of Financial Studies* 18, 831–873.

Cariño, D.R., Kent, T., Meyers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T. (1994). The Russel–Yasuda Kasai model: An asset/liability model for a Japanese insurance company using multistage stochastic programming. *Interfaces* 24, 29–49.

Censor, Y., Lent, A. (1981). An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications* 343, 321–353.

Czyzyk, J., Fourer, R., Mehrotra, S. (1995). A study of the augmented system and column-splitting approaches for solving two-stage stochastic linear programs by interior-point methods. *ORSA Journal on Computing* 7, 474–490.

Dantzig, G.B., Infanger, G. (1991). Large-scale stochastic linear programs—Importance sampling and Benders decomposition. In: Brezinski, C., Kulisch, U. (Eds.), *Computational and Applied Mathematics*, vol. I. North-Holland, Amsterdam, pp. 111–120.

Dantzig, G.B., Wolfe, P. (1960). The decomposition principle for linear programs. *Operations Research* 8, 101–111.

Davis, M.H.A., Norman, A.R. (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research* 15, 676–713.

DeMiguel, V., Nogales, F.J. (2006). Portfolio selection with robust estimates of risk. *London Business School working paper*.

Dempster, M.A.H. (1980). Introduction to stochastic programming. In: Dempster, M.A.H. (Ed.), *Stochastic Programming*. Academic Press, New York, pp. 3–59.

Dempster, M.A.H. (1988). On stochastic programming II: Dynamic problems under risk. *Stochastics* 25, 15–42.

Dempster, M.A.H., Thompson, G.W.P. (2002). Dynamic portfolio replication using stochastic programming. In: Dempster, M.A.H. (Ed.), *Risk Management: Value at Risk and Beyond*. Cambridge University Press, Cambridge, UK.

Dempster, M.A.H., Germano, M., Medova, E.A., Villaverde, M. (2003). Global asset liability management. *British Actuarial Journal* 9, 137–216.

Detemple, J., Garcia, R., Rindisbacher, M. (2007). Simulation methods for optimal portfolios. In: Birge, J.R., Linetsky, V. (Eds.), Financial Engineering. In: Handbooks in Operations Research and Management Science, vol. 15, Elsevier, Amsterdam, this volume.

Donohue, C.J., Birge, J.R. (2006). The abridged nested decomposition method for multistage stochastic linear programs with relatively complete recourse. *Algorithmic Operations Research* 1, 18–28.

Dupačová, J., Gröwe-Kuska, N., Römisch, W. (2003). Scenario reduction in stochastic programming: An approach using probability metrics. *Mathematical Programming, Series A* 95, 493–511.

Dyer, M., Stougie, L. (2006). Computational complexity of stochastic programming problems. *Mathematical Programming, Series A* 106, 423–432.

Eisner, M., Olsen, P. (1975). Duality for stochastic programming interpreted as l. p. in L_p-*space. SIAM Journal of Applied Mathematics* 28, 779–792.

Frantzeskakis, L.F., Powell, W.B. (1993). Bounding procedures for multistage stochastic dynamic networks. *Networks* 23, 575–595.

Frauendorfer, K. (1992). *Stochastic Two-Stage Programming*. *Lecture Notes in Economics and Mathematical Systems*, vol. 392. Spinger-Verlag, Berlin.

Garlappi, L., Uppal, R., Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies* 20, 41–81.

Ghilboa, I., Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–153.
Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer, New York.
Goldfarb, D., Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research* 28, 1–38.
Gondzio, J., Kouwenberg, R. (2001). High-performance computing for asset-liability management. *Operations Research* 49, 879–891.
Grinold, R.C. (1986). Infinite horizon stochastic programs. *SIAM Journal of Control Optimization* 24, 1246–1260.
Haugh, M.B., Kogan, L. (2007). Duality theory and approximate dynamic programming for pricing American options and portfolio optimization. In: Birge, J.R., Linetsky, V. (Eds.), Financial Engineering. In: Handbooks in Operations Research and Management Science, vol. 15, Elsevier, Amsterdam, this volume.
Haugh, M.B., Kogan, L., Wang, J. (2006). Evaluating portfolio policies: A duality approach. *Operations Research* 54, 405–418.
Heitsch, H., Römisch, W. (2003). Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications* 24, 187–206.
Higle, J., Sen, S. (1996). *Stochastic Decomposition*. Kluwer Academic Publishers, Dordrecht.
Høyland, K., Wallace, S.W. (2001). Generating scenario trees for multistage decision problems. *Management Science* 47, 295–307.
Judd, K.L. (1998). *Numerical Methods in Economics*. MIT Press, Cambridge, MA.
Kall, P. (1979). Computational methods for solving two-stage stochastic linear programming problems. *Journal of Applied Mathematics and Physics* 30, 261–271.
Kallberg, J.G., Ziemba, W.T. (1983). Comparison of alternative utility functions in portfolio selection problems. *Management Science* 29, 1257–1276.
Kiwiel, K.C. (1983). An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming* 27, 320–341.
Klaassen, P. (1998). Financial asset-pricing theory and stochastic programming models for asset/liability management: A synthesis. *Management Science* 44, 31–48.
Korn, R. (2004). Realism and practicality of transaction cost approaches in continuous-time portfolio optimisation: The scope of the Morton–Pliska approach. *Mathematical Methods of Operations Research* 60, 165–174.
Kouwenberg, R. (2001). Scenario generation and stochastic programming models for asset liability management. *European Journal of Operational Research* 134, 279–292.
Kreps, D., Porteus, E. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46, 185–200.
Kuhn, D. (2005). Aggregation and discretization in multistage stochastic programming. Institute for Operations Research and Computational Finance, University of St. Gallen, Switzerland.
Kusy, M., Ziemba, W.T. (1986). A bank asset and liability management model. *Operations Research* 34, 356–376.
Lemaréchal, C. (1978). Bundle methods in nonsmooth Optimization. In: Lemaréchal, C., Mifflin, R. (Eds.), *Nonsmooth Optimization. Proc. IIASA Workshop*. Pergamon, Oxford/Elmsford, NY, pp. 79–102.
Louveaux, F.V. (1980). A solution method for multistage stochastic programs with recourse with application to an energy investment problem. *Operations Research* 28, 889–902.
Maranas, C.D., Androulakis, I.P., Floudas, C.A., Berger, A.J., Mulvey, J.M. (1997). Solving long-term financial planning problems via global optimization. *Journal of Economic Dynamics and Control* 21, 1405–1425.
Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics* 51, 247–257.
Morton, A.J., Pliska, S.R. (1995). Optimal portfolio management with fixed transaction costs. *Mathematical Finance* 5, 337–356.
Mulvey, J.M., Ruszczyński, A. (1995). A new scenario decomposition method for large scale stochastic optimization. *Operations Research* 43, 477–490.

Mulvey, J.M., Vladimirou, H. (1991). Applying the progressive hedging algorithm to stochastic generalized networks. *Annals of Operations Research* 31, 399–424.

Muthuraman, K., Zha, H. (2006). Simulation based portfolio optimization for large portfolios with transaction costs. *Mathematical Finance* 16, 301–335.

Niederreiter, H. (1978). Quasi-Monte Carlo methods and pseudorandom numbers. *Bulletin of the American Mathematical Society* 84, 957–1041.

Nielsen, S.S., Zenios, S.A. (1993a). Proximal minimizations with *D*-functions and the massively parallel solution of linear stochastic network programs. *International Journal of Supercomputing and Applications* 7, 349–364.

Nielsen, S.S., Zenios, S.A. (1993b). A massively parallel algorithm for nonlinear stochastic network problems. *Operations Research* 41, 319–337.

Papi, M., Sbaraglia, S. (2006). Optimal asset-liability management with constraints: A dynamic programming approach. *Applied Mathematics and Computation* 173, 306–349.

Pástor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance* 55, 179–223.

Pástor, L., Stambaugh, R.F. (2000). Comparing asset pricing models: An investment perspective. *Journal of Financial Economics* 56, 335–381.

Pereira, M.V.F., Pinto, L.M.V.G. (1991). Multistage stochastic optimization applied to energy planning. *Mathematical Programming* 52, 359–375.

Pflug, G. (2001). Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming, Series B* 89, 251–271.

Rockafellar, R.T. (1976). Integral Functionals, Normal Integrands and Measurable Selections. *Lecture Notes in Mathematics*, vol. 543. Springer, Berlin.

Rockafellar, R.T., Wets, R.J.-B. (1976). Stochastic convex programming: Basic duality. *Pacific Journal of Mathematics* 63, 173–195.

Rockafellar, R.T., Wets, R.J.-B. (1986). A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming. *Mathematical Programming Study* 28, 63–93.

Rockafellar, R.T., Wets, R.J.-B. (1991). Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research* 16, 119–147.

Römisch, W. (2003). Stability of stochastic programming problems. In: Ruszczyński, A., Shapiro, A. (Eds.), *Stochastic Programming, Handbooks of Operations Research and Management Science*, vol. 10. Elsevier, Amsterdam, pp. 483–554.

Römisch, W., Wets, R.J.-B. (2006). Stability of ϵ-approximate solutions to convex stochastic programs. *Humboldt University working paper*.

Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming. *The Review of Economics and Statistics* 51, 239–246.

Shen, G., Caines, P.E. (2002). Hierarchically accelerated dynamic programming for finite-state machines. *IEEE Transactions on Automatic Control* 47, 271–283.

Shmoys, D.B., Swamy, C. (2006). An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *Journal of the ACM* 53, 978–1031.

Skiadis C. (2007). Dynamic portfolio choice and risk aversion. In: Birge, J.R., Linetsky, V. (Eds.), Financial Engineering. In: Handbooks in Operations Research and Management Science, vol. 15, Elsevier, Amsterdam, this volume.

Stravicky, B. (1980). Some results concerning an algorithm for the discrete recourse problem. In: Dempster, M.A.H. (Ed.), *Stochastic Programming*. Academic Press, New York.

Swamy, C., Shmoys, D.B. (2005). Sampling-based approximation algorithms for multi-stage stochastic optimization. In: *Proceedings of FOCS 2005*. IEEE Computer Society, Los Alamitos, CA, pp. 357–366.

Takriti, S., Birge, J.R., Long, E. (1996). A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems* 11, 1497–1508.

Talluri, K.T., Van Ryzin, G.J. (1998). An analysis of bid–price controls for network revenue management. *Management Science* 44, 1577–1593.

Trick, M.A., Zin, S.E. (1997). Spline approximations to value functions: Linear programming approach. *Macroeconomic Dynamics* 1, 255–277.

Tsitsiklis, J.N., Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control* 42, 674–690.

Tsitsiklis, J.N., Van Roy, B. (2001). Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks* 12, 694–703.

Tütüncü, R.H., Koenig, M. (2004). Robust asset allocation. *Annals of Operations Research* 132, 157–187.

van Binsbergen, J.H., Brandt, M.W. (2006). Optimal asset allocation in asset liability management, Duke University, *Fuqua School working paper*.

Van Slyke, R., Wets, R.J.-B. (1969). L-shaped linear programs with application to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics* 17, 638–663.

Varaiya, P., Wets, R.J.-B. (1989). Stochastic dynamic optimization approaches and computation. In: Iri, M., Tanabe, K. (Eds.), *Mathematical Programming: Recent Developments and Applications*. Kluwer, Dordrecht, pp. 309–332.

Wets, R.J.-B. (1990). Stochastic programming. In: Nemhauser, G.L., Rinnooy Kan, A.H.G., Todd, M.J. (Eds.), *Optimization, Handbooks in Operations Research and Management Science*, vol. 1. North-Holland, Amsterdam.

Wright, S.E. (1994). Primal–dual aggregation and disaggregation for stochastic linear programs. *Mathematics of Operations Research* 19, 893–908.

Yang, D.F., Zenios, S.A. (1997). A scalable parallel interior point algorithm for stochastic linear programming and robust optimization. *Computational Optimization and Applications* 7, 143–158.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15021-0

Chapter 21

Simulation Methods for Optimal Portfolios

Jérôme Detemple

Boston University School of Management, 595 Commonwealth ave., Boston, MA 02215, USA and CIRANO
E-mail: detemple@bu.edu

René Garcia

Department of Economics at the Université de Montréal, CIRANO and CIREQ
E-mail: rene.garcia@umontreal.ca

Marcel Rindisbacher

Rotman School of Management, University of Toronto and CIRANO
E-mail: marcel.rindisbacher@rotman.utoronto.ca

Abstract

This chapter surveys and compares Monte Carlo methods that have been proposed for the computation of optimal portfolio policies. The candidate approaches include the Monte Carlo Malliavin derivative (MCMD) method proposed by Detemple et al. [Detemple, J.B., Garcia, R., Rindisbacher, M. (2003). A Monte-Carlo method for optimal portfolios. *Journal of Finance* 58, 401–446], the Monte Carlo covariation (MCC) method of Cvitanic et al. [Cvitanic, J., Goukasian, L., Zapatero, F. (2003). Monte Carlo computation of optimal portfolio in complete markets. *Journal of Economic Dynamics and Control* 27, 971–986], the Monte Carlo regression (MCR) method of Brandt et al. [Brandt, M.W., Goyal, A., Santa-Clara, P., Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *Review of Financial Studies* 18, 831–873] and Monte Carlo finite difference (MCFD) methods. The asymptotic properties of the various portfolio estimators obtained are described. A numerical illustration of the convergence behavior of these estimators is provided in the context of a dynamic portfolio choice problem with exact solution. MCMD is shown to dominate other approaches.

1 Introduction

The optimal allocation of wealth among various assets is an important issue that has been of long-standing interest both for academics and practitioners. The workhorse models in the field, have been, for almost 50 years, based on the

mean–variance analysis developed by Markowitz (1952). This simple framework brought to light the fundamental notion of a mean–variance trade-off associated with the choice of different securities or portfolios. This popular notion and the associated portfolio rules remain, to this day, at the core of decisions taken and practical recommendations formulated by investment firms and financial advisors.

Yet, mean–variance portfolio rules have been known to be flawed for over 3 decades. In a seminal contribution, Merton (1971) identified the main problem, their failure to account for stochastic shifts in the investment opportunity set (i.e. in means and variances).[1] While of little consequence for very short term investors, this failure proves important for economic units with long horizons. Indeed, only a very particular class of long term investors, namely those with unit relative risk aversion (logarithmic utility), will find it optimal to behave as short term mean–variance optimizers. Generic long term investors follow amended portfolio rules that include intertemporal hedging terms, in addition to mean–variance components. The benefit of those hedging terms is intuitively clear: in a stochastically changing environment it pays to take intertemporal links into account and hedge against variations in means and variances.

Numerical methods for computing these hedging terms and the associated optimal portfolios have notably lagged behind. Much of the earlier literature has indeed searched for closed form solutions in the context of simple parametric models, with limited assets and state variables and simple dynamics. The earliest attempt to numerically solve a nontrivial portfolio choice problem can perhaps be attributed to Brennan et al. (1997), who examine a model with 3 assets and 4 state variables. Their approach uses numerical methods for partial differential equations (PDEs) and is based on the dynamic programming characterization of the optimal solution developed by Merton. Their study reveals the importance of the dynamic portfolio choice problem and highlights some of the difficulties that need to be overcome.

Rapid developments have followed. Simulation methods were first proposed by Detemple et al. (2003), who exploit a portfolio formula based on Malliavin calculus derived by Ocone and Karatzas (1991) for Itô price processes. Their basic method, labeled Monte Carlo Malliavin derivative (MCMD), involves the simulation of state variables and Malliavin derivatives to compute the expectations arising in the portfolio components. A variation of the method applies a change of variables (a Doss transformation) and simulates the transformed variables and their Malliavin derivatives to compute the relevant expressions. This Monte Carlo Malliavin derivative method with Doss transformation (MCMD-Doss) is proposed and studied in Detemple et al. (2003, 2005a, 2005c). An alternative, suggested by Cvitanic et al. (2003), uses an approximation of the optimal portfolio rule, based on the covariation between wealth and the underlying Brownian motions, as the basis for Monte

[1] See also Breeden (1979).

Carlo simulation. This simple approach, called the Monte Carlo covariation (MCC) method, is easy to implement as it only involves the simulation of the primitive state variables. Another approximation, that combines dynamic programming with regressions and simulations, is advocated by Brandt et al. (2005). It relies on an approximation of the optimality conditions for the portfolio and uses a regression-simulation method to evaluate conditional expectations in the coefficients of the approximate portfolio conditions. This Monte Carlo regression (MCR) scheme is reminiscent of the regression method developed by Longstaff and Schwartz (2001) for American options' valuation. Monte Carlo finite difference (MCFD) methods complete the list of simulation approaches that have been proposed to date for optimal portfolio calculations. This approach exploits the link between Malliavin derivatives and tangent processes (loosely speaking derivatives with respect to initial conditions) and evaluates the relevant derivatives using finite differences. MCFD approaches are described in Detemple et al. (2005d) and evaluated along with MCMD and MCC in the context of risk management problems.

This chapter surveys the recent literature on simulation methods for optimal portfolios. The various methods, informally described above, are presented in details and discussed. A numerical study is performed to evaluate their relative performances. MCMD is shown to dominate other candidate approaches.

Section 2 outlines the consumption-portfolio choice problem in a setting with complete markets and von Neumann–Morgenstern preferences and presents several representations formulas for its solution. Simulation methods for optimal portfolio calculations are reviewed in Section 3. Asymptotic properties of portfolio estimators are examined in Section 4 and a numerical study of the convergence behavior of the various methods is conducted in Section 5. Concluding remarks and avenues for future work are outlined in the last section. Appendix A presents elementary rules of Malliavin calculus that are needed to derive formulas underlying some of the methods. Proofs are collected in Appendix B.

2 The consumption-portfolio choice problem

We formulate a continuous time consumption-portfolio choice model in the tradition of Merton (1971). A finitely-lived investor operates in a frictionless economy in which asset prices and state variables follow a joint diffusion process. The investor's planning horizon is $[0, T]$.

2.1 The financial market

The financial market has d risky assets (stocks) and 1 locally riskless

$$\mathrm{d}S_{it} = S_{it}\big[\big(\mu_i(t, Y_t) - \delta_i(t, Y_t)\big)\,\mathrm{d}t + \sigma_i(t, Y_t)'\,\mathrm{d}W_t\big]; \quad S_{i0} \text{ given}, \tag{2.1}$$

where μ_i represents the return's drift, δ_i the dividend yield and σ_i' the $1 \times d$ vector of volatility coefficients. The coefficients of (2.1) depend on a $k \times 1$ vector of state variables $Y = (Y_1, \ldots, Y_k)'$. The interest rate on the riskless asset, $r(t, Y_t)$, also depends on the state variables. To simplify notation we will write μ_t for the $d \times 1$ vector of expected risky asset returns at date t, δ_t for the $d \times 1$ vector of dividend yields, σ_t for the $d \times d$ matrix of return volatilities and r_t for the interest rate. We assume that σ is invertible at all times (i.e. the market is complete).

The price system (2.1) induces a unique d-dimensional vector of market prices of risk $\theta_t = (\theta_{1t}, \ldots, \theta_{dt})'$ defined by $\theta_t \equiv \sigma_t^{-1}(\mu_t - r_t 1_d)$ where $1_d = (1, \ldots, 1)'$ is the d-dimensional vector of ones. The market prices of risk represent the premia implicitly assigned by the financial market to the sources of uncertainty (the Brownian motions) affecting the economy. The state price density (SPD), $\xi_v \equiv \exp(-\int_0^v (r_s + \frac{1}{2}\theta_s'\theta_s)\,\mathrm{d}s - \int_0^v \theta_s'\,\mathrm{d}W_s)$, is the stochastic discount factor that matters to find the value at date 0 of cash flows received at $v \geqslant 0$. The relative state price density (RSPD), $\xi_{t,v} \equiv \exp(-\int_t^v (r_s + \frac{1}{2}\theta_s'\theta_s)\,\mathrm{d}s - \int_t^v \theta_s'\,\mathrm{d}W_s) = \xi_v/\xi_t$, is the stochastic discount factor that matters to find the value at date t of cash flows received at $v \geqslant t$.

2.2 State variables

The state variables $Y = (Y_1, \ldots, Y_k)'$ affect the coefficients of asset returns and the riskfree rate (i.e. the opportunity set). The list of state variables can include the market prices of risk and the interest rate (e.g. $Y_1 = r$ and $Y_j = \theta_j$, $j = 2, \ldots, d+1$). Additional variables that could be relevant include dividend–price ratios, measures of firm sizes and measures of sales or revenues. State variables are assumed to evolve according to

$$\mathrm{d}Y_t = \mu^Y(t, Y_t)\,\mathrm{d}t + \sigma^Y(t, Y_t)\,\mathrm{d}W_t; \quad Y_0 \text{ given}, \tag{2.2}$$

where $\mu^Y(t, Y_t)$ is the $k \times 1$ vector of drift coefficients and $\sigma^Y(t, Y_t)$ is a $k \times d$ matrix of volatility coefficients.

2.3 Consumption, portfolios and wealth

The investor under consideration consumes and allocates his/her wealth among the different assets available. Let X_t be wealth at date t. Consumption is c_t and π_t is the $d \times 1$ vector of wealth proportions invested in the risky assets (thus $1 - \pi_t' 1_d$ is the proportion invested in the riskless asset). The evolution of wealth is governed by the stochastic differential equation

$$\mathrm{d}X_t = (X_t r_t - c_t)\,\mathrm{d}t + X_t \pi_t' \big[(\mu_t - r_t 1_d)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t\big] \tag{2.3}$$

subject to the initial condition $X_0 = x$.

2.4 Preferences

Preferences are assumed to have the time-separable von Neumann–Morgenstern representation. A consumption-terminal wealth plan (c, X_T) is ranked according to the criterion

$$\mathbf{E}\left[\int_0^T u(c_v, v)\,\mathrm{d}v + U(X_T, T)\right] \tag{2.4}$$

where the utility functions $u:[d_1, \infty) \times [0, T] \to \mathbb{R}$ and $U:[d_2, \infty) \to \mathbb{R}$ are strictly increasing, strictly concave and differentiable over their respective domains. We also assume that the limiting conditions $\lim_{c\to d_1} u'(c, t) = \lim_{X\to d_2} U'(X, T) = \infty$ and $\lim_{c\to\infty} u'(c, t) = \lim_{X\to\infty} U'(X, T) = 0$ hold for all $t \in [0, T]$. If domains include $\mathbb{R}_+ \times [0, T]$ (i.e. $d_1, d_2 \leqslant 0$) no further restrictions are imposed. If $[d_1, \infty)$ is a proper subset of $\mathbb{R}_+$ (i.e. $d_1 > 0$) we extend the function u to $\mathbb{R}_+ \times [0, T]$ by setting $u(c, t) = -\infty$ for $c \in \mathbb{R}_+ \backslash [d_1, \infty)$ and for all $t \in [0, T]$. We proceed in the same manner to extend U if $[d_2, \infty)$ is a proper subset of $\mathbb{R}_+$.

This class of utility functions includes the HARA specification

$$u(c, t) = \frac{1}{1-R}(c + A)^{1-R},$$

where $R > 0$. If A is positive the utility function $u(c, t)$ is defined over the domain $[d_1, \infty) = [-A, \infty)$ and satisfies all the required conditions. If $A < 0$ the function has the required properties over the subset $[d_1, \infty) = [-A, \infty) \subset \mathbb{R}_+$. The function is then extended by setting $u(c, t) = -\infty$ for $c \leqslant d_1$. This particular HARA specification corresponds to a model with subsistence consumption $-A$.

Under these assumptions the respective inverses $I:\mathbb{R}_+ \times [0, T] \to [d_1, \infty)$ and $J:\mathbb{R}_+ \to [d_2, \infty)$ of the marginal utility functions $u'(c, t)$ and $U'(X, T)$ exist and are unique. They are also strictly decreasing with limiting values $\lim_{y\to 0} I(y, t) = \lim_{y\to 0} J(y, T) = \infty$ and $\lim_{y\to\infty} I(y, t) = d_1$, $\lim_{y\to\infty} J(y, T) = d_2$.

2.5 The dynamic consumption-portfolio choice problem

The investor seeks to maximize expected utility

$$\max_{(c,\pi)} \mathbf{E}\left[\int_0^T u(c_v, v)\,\mathrm{d}v + U(X_T, T)\right] \tag{2.5}$$

subject to the following constraints

$$\mathrm{d}X_t = (r_t X_t - c_t)\,\mathrm{d}t + X_t \pi_t'\big[(\mu_t - r_t 1_d)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t\big]; \quad X_0 = x, \tag{2.6}$$

$$c_t \geqslant 0, \qquad X_T \geqslant 0 \tag{2.7}$$

for all $t \in [0, T]$. The first constraint, (2.6), describes the evolution of wealth given a consumption-portfolio policy (c, π). The next one (2.7) captures the physical restriction that consumption and bequest cannot become negative. This constraint ensures that wealth, that is the present value of future consumption, cannot become negative.

2.6 *Optimal consumption, portfolio and wealth*

Standard results of Pliska (1986), Karatzas et al. (1987) and Cox and Huang (1989) (see Karatzas and Shreve, 1998) can be invoked to show that the optimal consumption policy is

$$c_t^* = I(y^*\xi_t, t)^+ = \max\bigl\{I(y^*\xi_t, t), 0\bigr\}, \tag{2.8}$$

$$X_T^* = J(y^*\xi_T, T)^+ = \max\bigl\{J(y^*\xi_T, T), 0\bigr\} \tag{2.9}$$

where the constant y^* is the unique solution of the static budget constraint

$$\mathbf{E}\left[\int_0^T \xi_v I(y\xi_v, v)^+ \,\mathrm{d}v + \xi_T J(y\xi_T, T)^+\right] = x \tag{2.10}$$

with $x \geqslant \max\{\mathbf{E}[\int_0^T \xi_v d_1 \,\mathrm{d}v + \xi_T d_2], 0\}$.

The resulting wealth process is the present value of optimal future consumption and is therefore given by

$$X_t^* = \mathbf{E}_t\left[\int_t^T \xi_{t,v} I(y^*\xi_v, v)^+ \,\mathrm{d}v + \xi_{t,T} J(y^*\xi_T, T)^+\right] \equiv \mathbf{E}_t[F_{t,T}] \tag{2.11}$$

for $t \in [0, T]$, where $F_{t,T} \equiv \int_t^T \xi_{t,v} I(y^*\xi_v, v)^+ \,\mathrm{d}v + \xi_{t,T} J(y^*\xi_T, T)^+$. The associated optimal portfolio can be expressed as

$$X_t^* \pi_t^* = X_t^* (\sigma_t')^{-1} \theta_t + \xi_t^{-1} (\sigma_t')^{-1} \phi_t, \tag{2.12}$$

where ϕ is the predictable process in the representation of the martingale $M_t = \mathbf{E}_t[F] - \mathbf{E}[F]$ with $F \equiv F_{0,T} = \int_0^T \xi_t c_t^* \,\mathrm{d}t + \xi_T X_T^*$.

2.7 *The optimal portfolio: an explicit formula*

To find a more explicit expression for the optimal portfolio it remains to identify the process ϕ in (2.12). The Clark–Ocone formula (see Appendix A) becomes instrumental for that purpose: it identifies the integrand in the representation of the martingale M and enables us to express the optimal portfolio in terms of the parameters of the model (i.e. the structure of F). Ocone and Karatzas (1991) establish this portfolio formula for general models with Itô

price processes. The specialization to diffusions can be found in Detemple et al. (2003).

Applying the Clark–Ocone formula and using the rules of Malliavin calculus shows that

$$\phi_t = \mathbf{E}_t[\mathcal{D}_t F]$$

where

$$\mathcal{D}_t F = \mathbf{E}_t\left[\int_t^T Z_1(y^*\xi_v, v)\mathcal{D}_t\xi_v \,\mathrm{d}v + Z_2(y^*\xi_T, T)\mathcal{D}_t\xi_T\right] \tag{2.13}$$

with

$$\begin{aligned} Z_1(y^*\xi_v, v) &= I(y^*\xi_v, v)^+ + y^*\xi_v I'(y^*\xi_v, v)1_{\{I(y^*\xi_v,v)\geqslant 0\}} \\ &= c_v^*\left(1 - \frac{1}{R_u(c_v^*, v)}\right), \end{aligned} \tag{2.14}$$

$$\begin{aligned} Z_2(y^*\xi_T, T) &= J(y^*\xi_T, T)^+ + y^*\xi_T J'(y^*\xi_T, T)1_{\{J(y^*\xi_T,T)\geqslant 0\}} \\ &= X_T^*\left(1 - \frac{1}{R_U(X_T^*, T)}\right). \end{aligned} \tag{2.15}$$

In these expressions $I'(y^*\xi_v, v)$, $J'(y^*\xi_T, T)$ are the derivatives with respect to the first argument $y^*\xi$ of the inverse marginal utility functions and $R_u(x, v) = -u_{cc}(x, v)x/u_c(x, v)$, $R_U(x, T) = -U_{XX}(x, T)x/U_X(x, T)$ are relative risk aversion coefficients.

From the definition of the stochastic discount factor ξ in Section 2.1 we obtain

$$\mathcal{D}_t\xi_v \equiv -\xi_v\left(\int_t^v (\mathcal{D}_t r_s + \theta_s'\mathcal{D}_t\theta_s)\,\mathrm{d}s + \int_t^v \mathrm{d}W_s' \cdot \mathcal{D}_t\theta_s + \theta_t'\right).$$

The chain rule of Malliavin calculus then gives $\mathcal{D}_t\xi_v = -\xi_v(H_{t,v}' + \theta_t')$ with

$$H_{t,v}' = \int_t^v \big(\partial r(s, Y_s) + \theta_s'\partial\theta(s, Y_s)\big)\mathcal{D}_t Y_s \,\mathrm{d}s + \int_t^v \mathrm{d}W_s' \cdot \partial\theta(s, Y_s)\mathcal{D}_t Y_s \tag{2.16}$$

and where $\mathcal{D}_t Y_s$ satisfies the stochastic differential equation

$$\begin{aligned} \mathrm{d}\mathcal{D}_t Y_s &= \left[\partial\mu^Y(s, Y_s)\,\mathrm{d}s + \sum_{j=1}^d \partial\sigma_j^Y(s, Y_s)\,\mathrm{d}W_s^j\right]\mathcal{D}_t Y_s; \\ \mathcal{D}_t Y_t &= \sigma^Y(t, Y_t). \end{aligned} \tag{2.17}$$

In this expression the notation $\partial f(Y)$ stands for the $p\times k$-dimensional Jacobian matrix of a p-dimensional vector function f with respect to the k-dimensional

vector Y. Substituting (2.13)–(2.17) back in (2.12), collecting terms and simplifying leads to our explicit portfolio formula (for details see the proof of Proposition 1 in Appendix B). Our next proposition summarizes the results

Proposition 1. *Consider the dynamic consumption-portfolio problem (2.6)–(2.7). The optimal consumption policy is*

$$c_v^* = I(y^*\xi_v, v)^+, \qquad X_T^* = J(y^*\xi_T, T)^+. \tag{2.18}$$

The optimal portfolio policy has the decomposition $X_t^*\pi_t^* = X_t^*[\pi_{1t}^* + \pi_{2t}^*]$ where π_{1t}^* is the mean–variance demand and π_{2t}^* the intertemporal hedging demand. The two components are

$$X_t^*\pi_{1t}^* = -\mathbf{E}_t[D_{t,T}](\sigma_t')^{-1}\theta_t, \tag{2.19}$$

$$X_t^*\pi_{2t}^* = -(\sigma_t')^{-1}\mathbf{E}_t[G_{t,T}], \tag{2.20}$$

where

$$\begin{aligned} D_{t,T} \equiv & \int_t^T \xi_{t,v}(y^*\xi_v)I'(y^*\xi_v, v)1_{\{I(y^*\xi_v,v)\geqslant 0\}}\,\mathrm{d}v \\ & + \xi_{t,T}(y^*\xi_T)J'(y^*\xi_T, T)1_{\{J(y^*\xi_T,T)\geqslant 0\}}, \end{aligned} \tag{2.21}$$

$$G_{t,T} \equiv \int_t^T \xi_{t,v}Z_1(y^*\xi_v, v)H_{t,v}\,\mathrm{d}v + \xi_{t,T}Z_2(y^*\xi_T, T)H_{t,T} \tag{2.22}$$

and where $Z_1(y^*\xi_v, v)$ and $Z_2(y^*\xi_T, T)$ are given in (2.14)–(2.15), the random variable $H_{t,v}$ is defined in (2.16) and the Malliavin derivative of the state variables, $\mathcal{D}_t Y_s$, satisfies the stochastic differential equation (2.17). The multiplier y^* solves the nonlinear equation (2.10). Optimal wealth is $X_t^* = \mathbf{E}_t[F_{t,T}]$.

The portfolio decomposition described in this proposition reflects two investment motives. The first one, which underlies the mean–variance demand π_1, is driven by the trade-off between risk and return embedded in asset returns. This motive, originally identified by Markowitz (1952), has played an important role in portfolio theory and remains at the core of practical implementations. The second one, underlying the demand component π_2, is a hedging motive prompted by stochastic fluctuations in the opportunity set. This intertemporal motive, identified by Merton (1971), is a fundamental aspect of optimal dynamic portfolio policies whose implementation has become a focus of current practice.

For later developments we record the special case of constant relative risk aversion in the following corollary:

Corollary 1. *Suppose that the investor exhibits constant relative risk aversion R and has subjective discount factor $\eta_t \equiv \exp(-\beta t)$ where β is a constant rate.*

The optimal consumption policy is given by $c_v^* = (y^*\xi_v/\eta_v)^{-1/R}$ *and* $X_T^* = (y^*\xi_T/\eta_T)^{-1/R}$. *The optimal portfolio is* $X_t^*\pi_t^* = X_t^*[\pi_{1t}^* + \pi_{2t}^*]$ *where*

$$X_t^*\pi_{1t}^* = \frac{X_t^*}{R}(\sigma_t')^{-1}\theta_t, \tag{2.23}$$

$$X_t^*\pi_{2t}^* = -X_t^*\rho(\sigma_t')^{-1}\frac{\mathbf{E}_t[\int_t^T \xi_{t,v}^{\rho}\eta_{t,v}^{1/R}H_{t,v}\,\mathrm{d}v + \xi_{t,T}^{\rho}\eta_{t,T}^{1/R}H_{t,T}]}{\mathbf{E}_t[\int_t^T \xi_{t,v}^{\rho}\eta_{t,v}^{1/R}\,\mathrm{d}v + \xi_{t,T}^{\rho}\eta_{t,T}^{1/R}]} \tag{2.24}$$

with $\rho = 1 - 1/R$ *and* $H_{t,v}$ *defined in* (2.16).

For general model structures the conditional expectations appearing in the formulas of Proposition 1 and Corollary 1 cannot be calculated in more explicit form. Numerical methods must then be used in order to implement the optimal portfolio policies. The complexity inherent in the random variables $\xi_{t,v}$, $H_{t,v}$ appearing in the expressions obtained, and in particular their path-dependent nature, naturally suggests the use of Monte Carlo simulation for computation purposes.

2.8 *Malliavin derivative representation and dynamic programming*

The classic approach to the consumption-portfolio choice problem in a Markovian setting was pioneered by Merton (1971) and is based on dynamic programming principles. Let $V(t, X^*, Y)$ be the value function associated with the problem. The optimal consumption and terminal wealth policies and the optimal portfolio are expressed in terms of the derivatives $V_t, V_x, V_y, V_{xx}, V_{xy}, V_{yy}$ of the value function as

$$c_t^* = I\big(V_x(t, X_t^*, Y_t), t\big)^+, \qquad X_T^* = J\big(V_x(T, X_T^*, Y_T), T\big)^+, \tag{2.25}$$

$$\begin{aligned} X_t^*\pi_t^* = {} & \frac{V_x(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)}\big(\sigma(t, Y_t)'\big)^{-1}\theta(t, Y_t) \\ & + \big(\sigma(t, Y_t)'\big)^{-1}\sigma^Y(t, Y_t)'\frac{V_{yx}(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)}. \end{aligned} \tag{2.26}$$

The value function solves the partial differential equation (PDE)

$$\begin{aligned} & u\big(I(V_x, t)^+, t\big) - V_x I(V_x, t)^+ + V_t + V_y\mu^Y + \frac{1}{2}\mathrm{trace}\big\{V_{yy}\sigma^Y\big(\sigma^Y\big)'\big\} \\ & \quad - \frac{1}{2}V_{xx}\|\psi'\|^2 = 0, \end{aligned} \tag{2.27}$$

where $\psi \equiv \frac{V_x}{-V_{xx}}\theta + (\sigma^Y)'\frac{V_{yx}}{-V_{xx}}$, subject to the boundary conditions $V(T, x, y) = U(x, T)$ and $V(t, 0, y) = \int_t^T u(0, s)\,\mathrm{d}s + U(0, T)$.

Our next result draws the link between Merton's solution and the probabilistic representation obtained in Proposition 1.

Proposition 2. *The state price density is proportional to the wealth derivative of the value function*

$$y^*\xi_t = V_x(t, X_t^*, Y_t), \quad or \quad \xi_t = \frac{V_x(t, X_t^*, Y_t)}{V_x(0, X_0^*, Y_0)} \tag{2.28}$$

ensuring that the optimal consumption and terminal wealth policies in (2.18) and (2.25) are identical. The scaling factors in the mean–variance and hedging demands (2.19)–(2.20) are given by

$$-\mathbf{E}_t[D_{t,T}] = \frac{V_x(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)}, \tag{2.29}$$

$$-\mathbf{E}_t[G_{t,T}] = \sigma^Y(t, Y_t)' \frac{V_{xy}(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)}. \tag{2.30}$$

Formulas (2.18)–(2.22) are alternative representations of the solution as expressed in (2.25)–(2.27).

Proposition 2 shows that our previous expressions for the optimal policies, (2.18)–(2.22), are probabilistic representations of the formulas derived by Merton. These representation are in the spirit of Feynman–Kac as they express the elements of the HJB equation in terms of conditional expectations of random variables. Note, in particular, that $-\mathbf{E}_t[D_{t,T}]$ in (2.29) is simply the probabilistic representation of the coefficient of "absolute risk tolerance" of the indirect utility function V (the value function).[2]

The results in Proposition 2 shed light on the relation between the value function and the fundamentals of the model, namely preferences and the state price density. For instance, it is well known that the hedging motive vanishes (at all times and in all states) if and only if the vector of cross partial derivatives of the value function, V_{xy}, is identically equal to zero. The Malliavin derivative representation in (2.30) and Equation (2.22) show that this condition is satisfied if and only if the processes (r, θ) are deterministic and/or the investor displays myopic behavior ($R_u = R_U = 1$).

2.9 Malliavin derivative and tangent process

For interpretation and computational purposes it is also instructive to rewrite the portfolio policy in terms of the derivative of the state variables with respect to their initial values. This derivative, called the tangent process (or first variation process), is described in Appendix A (see Section A.8).

The tangent process of Y, denoted by $\nabla_{t,y}Y \equiv \{\nabla_{t,y}Y_v\colon v \in [t, T]\}$, captures the change in the future values of Y following an incremental perturbation of

[2] The coefficient of absolute risk aversion $A(x)$ of a utility function u is $A(x) \equiv R(x)/x$ where $R(x) = -u''(x)x/u'(x)$ is relative risk aversion. Absolute risk tolerance is $1/A(x)$.

the position $Y_t = y$ at time t. In particular, for $v \geqslant t$, $\nabla_{t,y} Y_v$ is the variation in Y_v due to the initial perturbation. The tangent process is easy to characterize when Y solves an SDE. In fact, one can verify that $\nabla_{t,y} Y$ solves the SDE

$$d(\nabla_{t,y} Y_v) = \left(\partial \mu^Y(v, Y_v)\, \mathrm{d}v + \sum_{j=1}^{d} \partial \sigma_j^Y(v, Y_v)\, \mathrm{d}W_v^j \right) \nabla_{t,y} Y_v;$$
$$\nabla_{t,y} Y_t = I_k, \tag{2.31}$$

where I_k is the k-dimensional identity matrix. A comparison of (2.31) with (2.17) shows that the equation for the tangent process differs from the one for the Malliavin derivative only through the initial condition ($\nabla_{t,y} Y_t = I_k$ versus $\mathcal{D}_t Y_t = \sigma^Y(t, Y_t)$). It follows immediately that the relationship

$$\mathcal{D}_t Y_t = \nabla_{t,y} Y_v \sigma^Y(t, Y_t) \tag{2.32}$$

holds. The tangent process can be viewed as a normalized version of the Malliavin derivative. Conversely, the Malliavin derivative is a linear transformation of the tangent process.

Relationship (2.32) between the two notions enables us to rewrite the hedging term (2.20) in the form

$$X_t^* \pi_{2t}^* = -\big(\sigma_t(t, Y_t)'\big)^{-1} \sigma^Y(t, Y_t)' \mathbf{E}_t\big[G_{t,T}(\Phi)\big] \tag{2.33}$$

where

$$G_{t,T}(\Phi) \equiv \int_t^T \xi_{t,v} Z_1(y^* \xi_v, v) \Phi_{t,v}\, \mathrm{d}v + \xi_{t,T} Z_2(y^* \xi_T, T) \Phi_{t,T} \tag{2.34}$$

and

$$\Phi'_{t,v} \equiv \int_t^v \big(\partial r(s, Y_s) + \theta'_s \partial \theta(s, Y_s)\big) \nabla_{t,y} Y_s\, \mathrm{d}s + \int_t^v \mathrm{d}W'_s \cdot \partial \theta(s, Y_s) \nabla_{t,y} Y_s. \tag{2.35}$$

To derive this representation we used $H'_{t,v} = \Phi'_{t,v} \sigma^Y(t, Y_t)$. For computations it is also useful to note that $\Phi_{t,v} = -\nabla_{t,y} \log(\xi_{t,v})$: the functional $\Phi_{t,v}$ is the variation of $-\log(\xi_{t,v})$ for a perturbation in the position of the state variables $Y_t = y$ at time t. Finally, one can write the general representation

$$X_t^* \pi_{2t}^* = -\big(\sigma_t(t, Y_t)'\big)^{-1} \sigma^Y(t, Y_t)' \mathbf{E}_t\big[(\nabla_{t,y} F_{t,T})'\big] \tag{2.36}$$

where the functional $F_{t,T}$ is as defined in (2.11) and $(\nabla_{t,y}F_{t,T})' = G_{t,T}(\Phi)$. A comparison of (2.33) with (2.30) also shows that

$$\frac{V_{xy}(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)} = -\mathbf{E}_t\left[\int_t^T \xi_{t,v} Z_1(y^*\xi_v, v)\Phi_{t,v}\,\mathrm{d}v + \xi_{t,T} Z_2(y^*\xi_T, T)\Phi_{t,T}\right]. \tag{2.37}$$

This relation captures the intuitive notion that the hedging coefficient is related to the impact of a perturbation in the state variables at date t on the optimal wealth. This effect is precisely the expectation on the right-hand side of (2.37).

3 Simulation methods for portfolio computation

This section reviews various Monte Carlo methods that have been proposed for the computation of asset allocation rules.

3.1 Monte Carlo Malliavin derivatives (Detemple et al., 2003)

This simulation approach, developed by Detemple et al. (2003), is directly based on the formulas described in Section 2.7. Suppose that we are in the general context of Proposition 1 where the multiplier y^* for the static budget constraint cannot be solved explicitly from (2.10). Consider first the case where y^* has already been calculated by solving (2.10) numerically. In that case the method proceeds by rewriting the hedging demand in Proposition 1 as

$$X_t^*\pi_{2t}^* = -(\sigma_t')^{-1}\mathbf{E}_t[G_{t,T}] \tag{3.1}$$

where $G_{t,T} \equiv G_{t,T}^c + G_{t,T}^x$, with

$$G_{t,s}^c \equiv \int_t^s \xi_{t,v} Z_1(y^*\xi_v, v) H_{t,v}\,\mathrm{d}v \quad \text{and}$$
$$G_{t,T}^x \equiv \xi_{t,T} Z_2(y^*\xi_T, T) H_{t,T}. \tag{3.2}$$

To calculate $X_t^*\pi_{2t}^*$ write the random variables in the hedges in the form of a joint system $V_{t,s}' \equiv (Y_s', \mathrm{vec}(\mathcal{D}_t Y_s)', K_{t,s}, H_{t,s}', (G_{t,s}^c)')$, where $\mathrm{vec}(\cdot)$ denotes the operator stacking the columns of a matrix one below the other, and where

$$K_{t,v} \equiv \int_t^v \left(r_s + \frac{1}{2}\theta_s'\theta_s\right)\mathrm{d}s + \int_t^v \theta_s'\,\mathrm{d}W_s,$$

$$H'_{t,v} \equiv \int_t^v \partial r(s, Y_s)\mathcal{D}_t Y_s \,\mathrm{d}s + \int_t^v \theta'_s \partial\theta(s, Y_s)\mathcal{D}_t Y_s \,\mathrm{d}s + \int_t^v \mathrm{d}W'_s \cdot \partial\theta(s, Y_s)\mathcal{D}_t Y_s$$

and $\xi_{t,v} = \exp(-K_{t,v})$. An application of Itô's lemma shows that

$$\mathrm{d}K_{t,s} = \left(r_s + \frac{1}{2}\theta'_s\theta_s\right)\mathrm{d}s + \theta'_s\,\mathrm{d}W_s, \tag{3.3}$$

$$\mathrm{d}H'_{t,s} = \partial r(s, Y_s)\mathcal{D}_t Y_s\,\mathrm{d}s + \big(\mathrm{d}W_s + \theta(s, Y_s)\,\mathrm{d}s\big)'\partial\theta(s, Y_s)\mathcal{D}_t Y_s, \tag{3.4}$$

$$\mathrm{d}G^c_{t,s} = \xi_{t,s} Z_1(y^*\xi_s, s) H_{t,s}\,\mathrm{d}s, \tag{3.5}$$

where $(Y_s, \mathcal{D}_t Y_s)$ satisfy (2.2), (2.17). Initial conditions are $H_{t,t} = 0_d$, $G^c_{t,t} = 0_d$, where 0_d denotes the d-dimensional null vector, and $K_{t,t} = 0$.

Next, simulate M trajectories of V using (3.3)–(3.5), (2.2) and (2.17). To do this select a discretization scheme, such as the Euler scheme, the Milshtein scheme or any other higher order procedure and let N be the number of discretization points of the time interval $[0, T]$ chosen. This simulation produces M estimates $\{V^{N,i}_{t,s}\colon\ s \in [t, T]\}$, $i = 1, \ldots, M$, of the trajectories $\{V_{t,s}\colon\ s \in [t, T]\}$. Given that y^* is already known (through prior computation) the terminal values of the simulated processes can be used to construct M estimates of the random variables $G_{t,T}$. Averaging over these M values yields the estimate

$$\widehat{X^*_t \pi^*_{2t}} = -(\sigma'_t)^{-1}\frac{1}{M}\sum_{i=1}^{M} G^{N,i}_{t,T}$$

of the hedging demand.[3]

Suppose now that the multiplier y^* is unknown. In this case a two-stage simulation procedure can be employed to calculate the hedging demands. The first stage mixes iteration and simulation to calculate y^*. Fix a candidate multiplier y. Based on this choice simulate $(K_{0,s}, F^c_{0,s})$ where

$$F^c_{0,s} = \int_0^s \xi_v I(y\xi_v, v)^+\,\mathrm{d}v$$

in order to estimate the cost of consumption (the left-hand side of (2.10)). If the value obtained exceeds resources (initial wealth x) raise the candidate y

[3] The computation of the mean–variance component $X^*_t\pi^*_{1t}$ is carried out along the same lines. The evaluation of this demand component is straightforward due to the simple structure of the term $\mathbf{E}_t[D_{t,T}]$.

and repeat the calculation. In the opposite case reduce the candidate y. Repeat until the difference falls below some preselected threshold. The second stage parallels the procedure outlined above for the case of a known y^*.

Various procedures can be employed to accelerate the iterative search in stage 1. Schemes available include the Newton–Raphson procedure, the bracketing method, the bisection method, the secant method, the false position method, Ridder's method and the method of van Wijngaarden–Dekker–Brent (see Press et al., 1992 for details).

3.2 The Doss transformation (Detemple et al., 2003, 2005a)

The computation of the Malliavin derivatives in the portfolio formula can also be performed using a change of variables, commonly called a "Doss transformation." This change of variables, examined in Detemple et al. (2003, 2005a), leads to a characterization of Malliavin derivatives involving the solution of an ordinary differential equation (ODE). To simplify matters we assume $k = d$ (for more general cases see Detemple et al., 2005a).

Consider now a multivariate diffusion satisfying the restrictions

Condition 1. *The coefficients of the diffusion* (2.2) *have the following properties*:

1. *Differentiability*: $\mu^Y \in C([0, T] \times \mathbb{R}^d)$, $\sigma_j^Y \in C([0, T] \times \mathbb{R}^d)$,
2. *Boundedness*: $\mu^Y(t, 0)$ *and* $\sigma_j^Y(t, 0)$ *are bounded for all* $t \in [0, T]$, *and*
3. *Invertibility*:
   (a) $\partial_2 \sigma_j^Y \sigma_i^Y = \partial_2 \sigma_i^Y \sigma_j^Y$ (*i.e. the vector field generated by the columns of* σ *is abelian*),
   (b) $\mathrm{rank}(\sigma) = d$, *a.e.*

Under the provisions of Condition 1 there exists an invertible function $\Gamma : [0, T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ solving the total differential equation

$$\partial_2 \Gamma(t, z) = \sigma\big(t, \Gamma(t, z)\big); \qquad \Gamma(t, 0) = 0 \quad \text{for all } t \in [0, T] \tag{3.6}$$

and a d-dimensional process Z satisfying

$$\mathrm{d}Z_v = A(v, Z_v)\,\mathrm{d}v + \mathrm{d}W_v; \qquad \Gamma(0, Z_0) = Y_0 \tag{3.7}$$

where

$$\begin{aligned} A(t, z) \equiv{}& \sigma\big(t, \Gamma(t, z)\big)^{-1} \\ &\times \left[\mu^Y(t, y) - \frac{1}{2}\sum_{j=1}^{d} \partial_y \sigma_j^Y(t, y)\sigma_j(t, y)\right]_{|y=\Gamma(t,z)} \\ &- \partial_1 \Gamma(t, z), \end{aligned} \tag{3.8}$$

such that

$$\mathcal{D}_t Y_v = \sigma\big(v, \Gamma(v, Z_v)\big)\mathcal{D}_t Z_v \quad \text{for all } v \geqslant t, \tag{3.9}$$

$$\mathrm{d}\mathcal{D}_t Z_v = \partial_2 A(v, Z_v)\mathcal{D}_t Z_v \,\mathrm{d}v; \qquad \lim_{v\to t} \mathcal{D}_t Z_v = I_d. \tag{3.10}$$

This final expression (3.9)–(3.10) for the Malliavin derivative $\mathcal{D}_t Y_v$ does not involve stochastic integrals. An Euler approximation based on (3.6)–(3.10) will therefore converge faster (see Detemple et al., 2005a, 2005c).

Property 3(a) of Condition 1 is always satisfied for univariate diffusions. For multivariate diffusions, it represents a commutativity condition. It is, in fact, the same commutativity condition that is needed to implement the Milshtein scheme in the case of multivariate diffusions, without resorting to further subdiscretizations of the time interval (see Detemple et al., 2005c for details).

The MCMD-Doss estimator for the optimal portfolio, is obtained by using (3.6)–(3.10) to calculate the components of the portfolio policy.

3.3 *Monte Carlo covariation* (Cvitanic et al., 2003)

Another simulation-based approach, proposed by Cvitanic et al. (2003), is based on an approximation of the volatility coefficient of the optimal wealth process. The optimal portfolio, being a linear transformation of the volatility of the wealth process, can be estimated from this approximation.

The limits

$$X_t^* \pi_t^{*\prime} \sigma_t = \lim_{h\to 0} \frac{1}{h} \mathbf{E}_t\big[F_{t+h,T}(W_{t+h} - W_t)'\big], \tag{3.11}$$

$$X_t^* \pi_t^{*\prime} \sigma_t = \lim_{h\to 0} \frac{1}{h} \mathbf{E}_t\left[F_{t,T} \frac{(W_{t+h} - W_t)'}{\xi_{t,t+h}}\right], \tag{3.12}$$

$$X_t^* \pi_t^{*\prime} \sigma_t = X_t^* \theta_t' + \lim_{h\to 0} \frac{1}{h} \mathbf{E}_t\big[F_{t,T}(W_{t+h} - W_t)'\big] \tag{3.13}$$

with $F_{t,T}$ as defined in (2.11), can serve as foundations for the approach (see Appendix B for derivations). Approximations of the optimal portfolio are obtained by fixing a discretization h and setting

$$X_t^* \pi_t^{*\prime} \sigma_t \simeq \frac{1}{h} \mathbf{E}_t\big[F_{t+h,T}(W_{t+h} - W_t)'\big], \tag{3.14}$$

$$X_t^* \pi_t^{*\prime} \sigma_t \simeq \frac{1}{h} \mathbf{E}_t\left[F_{t,T} \frac{(W_{t+h} - W_t)'}{\xi_{t,t+h}}\right], \tag{3.15}$$

$$X_t^* \pi_t^{*\prime} \sigma_t \simeq X_t^* \theta_t' + \frac{1}{h} \mathbf{E}_t\big[F_{t,T}(W_{t+h} - W_t)'\big]. \tag{3.16}$$

The conditional expectations on the right-hand sides of (3.14), (3.15) and (3.16) are then computed by simulation of the relevant processes and averaging over independent replications. The procedure originally developed by CGZ relies on (3.14) or (3.15). It is based on models with constant relative risk aversion and either terminal wealth [estimator (3.15)] or intermediate consumption [estimator (3.14)], but not both. These are subcases of the setting in Corollary 1 for which the multiplier can be eliminated, resulting in (2.24). Formula (3.16)

is an alternative approximation that isolates the volatility of discounted wealth related to the volatility of the state price density. Implementation of these approximations for general preferences requires a preliminary stage to compute y^*.[4]

The procedure is easy to implement, as it does not require the simulation of auxiliary processes such as Malliavin derivatives. Nevertheless, it is based on an approximation (as h is fixed) of the optimal policy, and this will affect the convergence properties of the method. We refer to this method as MCC (Monte Carlo covariation). MCC estimators based on (3.14), (3.15) or (3.16) are numerically different. This difference only disappears in the limit as h vanishes.

3.4 Monte Carlo finite difference (MCFD)

The Monte Carlo finite difference (MCFD) method computes the hedging terms based on a version of the formulas (2.33)–(2.35) involving tangent processes. In essence the method calculates a tangent process by simulating the underlying process using perturbed initial values and then taking a finite difference approximation of the relevant derivative. This computation can be performed path-by-path. Conditional expectations involving tangent processes can then be calculated by averaging the random variables of interest over all the trajectories.[5]

Three versions of the formula involving tangent processes can serve as starting points for implementation. The first one consists of Equations (2.33)–(2.35) where $\Phi_{t,v}$ is expressed in terms of the tangent process $\nabla_{t,y}Y$ of the state variables. The second version consists of Equations (2.33)–(2.34) where $\Phi_{t,v} = -\nabla_{t,y}\log(\xi_{t,v})$ is expressed in terms of the variation of the log-RSPD. The last one is the general representation (2.36) based on the variation $\nabla_{t,y}F_{t,T}$ of the functional $F_{t,T}$.

Finite difference approximations of the relevant tangent processes are

$$\nabla_{t,y_j}^{\tau_j,\alpha_j} Y_v = \frac{1}{\tau_j}\big(Y_v(Y_t + \alpha_j\tau_j e_j) - Y_v\big(Y_t - (1-\alpha_j)\tau_j e_j\big)\big),$$

$$\Phi_{t,v}^{\tau_j,\alpha_j} \equiv -\nabla_{t,y_j}^{\tau_j,\alpha_j}\log\big(\xi_{t,v}(Y)\big) = -\frac{1}{\tau_j}\big(\log\big(\xi_{t,T}(Y_t + \alpha_j\tau_j e_j)\big) - \log\big(\xi_{t,T}\big(Y_t - (1-\alpha_j)\tau_j e_j\big)\big)\big),$$

[4] Models with constant relative risk averse utility functions, but different risk aversion coefficients for the utility of terminal wealth and the utility of intermediate consumption, fall outside the scope of Corollary 1. For those settings a preliminary stage is also needed to compute the budget constraint multiplier y^*.

[5] Finite difference methods have been used extensively to solve PDEs or ODEs in applications such as option pricing and asset allocation. The interest of combining these methods with Monte Carlo simulation, to handle certain financial applications, has only been noted recently.

$$\nabla_{t,y_j}^{\tau_j,\alpha_j} F_{t,T} = \frac{1}{\tau_j}\big(F_{t,T}(Y_t + \alpha_j \tau_j e_j) - F_{t,T}\big(Y_t - (1-\alpha_j)\tau_j e_j\big)\big),$$

where $\alpha_j \in [0,1]$, $\tau_j > 0$, and $e_j \equiv [0,\ldots,0,1,0,\ldots,0]$ is the jth unit vector. Different choices of α_j result in different types of finite difference approximations. The selection $\alpha_j = 1$ corresponds to a single forward difference, $\alpha_j = 0$ to a single backward difference and $\alpha_j = 1/2$ to a central difference approximation of the tangent process of interest.

To simplify notation we write $\nabla_{t,y}^{\tau,\alpha} Y_v$, $\Phi_{t,v}^{\tau,\alpha}$, $\nabla_{t,y}^{\tau,\alpha} F_{t,T}$ for the vectors of tangent processes, where $\tau = (\tau_1,\ldots,\tau_k)$ and $\alpha = (\alpha_1,\ldots,\alpha_k)$. As $\tau \to 0$ the limits

$$\begin{aligned}
&\nabla_{t,y}^{\tau,\alpha} Y_v \to \nabla_{t,y} Y_v,\\
&\Phi_{t,v}^{\tau,\alpha} = -\nabla_{t,y}^{\tau,\alpha} \log\big(\xi_{t,v}(Y)\big) \to \Phi_{t,v} = -\nabla_{t,y} \log\big(\xi_{t,v}(Y)\big),\\
&\nabla_{t,y}^{\tau,\alpha} F_{t,T} \to \nabla_{t,y} F_{t,T}
\end{aligned}$$

hold (**P**-a.s.). Under regularity conditions permitting the exchange of limits and conditional expectations we can write

$$X_t^* \pi_{2t}^* = -\big(\sigma_t(t,Y_t)'\big)^{-1} \sigma^Y(t,Y_t)' \mathbf{E}_t\big[G_{t,T}(\Phi)\big]$$

with

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] = \lim_{\tau\to 0} \mathbf{E}_t\big[G_{t,T}\big(\Phi_{t,v}^{\tau,\alpha}\big)\big] \tag{3.17}$$

or

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] = \mathbf{E}_t[\nabla_{t,y} F_{t,T}] = \lim_{\tau\to 0} \mathbf{E}_t\big[\nabla_{t,y}^{\tau,\alpha} F_{t,T}\big]. \tag{3.18}$$

Writing $\Phi'_{t,v}(\nabla_{t,y} Y_v)$ for the left-hand side of (2.35) to emphasize the dependence on the tangent process $\nabla_{t,y} Y_v$ we also have

$$\Phi_{t,v} = \lim_{\tau\to 0} \Phi_{t,v}\big(\nabla_{t,y}^{\tau,\alpha} Y_v\big)$$

P-*a.s.*, leading to

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] = \lim_{\tau\to 0} \mathbf{E}_t\big[G_{t,T}\big(\Phi_{t,v}\big(\nabla_{t,y}^{\tau,\alpha} Y_v\big)\big)\big]. \tag{3.19}$$

Finite difference approximations of the hedging term are obtained by fixing τ and approximating the conditional expectation $\mathbf{E}_t[G_{t,T}(\Phi)]$ by

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] \simeq \mathbf{E}_t\big[G_{t,T}\big(\Phi_{t,v}\big(\nabla_{t,y}^{\tau,\alpha} Y_v\big)\big)\big], \tag{3.20}$$

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] \simeq \mathbf{E}_t\big[G_{t,T}\big(\Phi_{t,v}^{\tau,\alpha}\big)\big], \tag{3.21}$$

$$\mathbf{E}_t\big[G_{t,T}(\Phi)\big] \simeq \mathbf{E}_t\big[\nabla_{t,y}^{\tau,\alpha} F_{t,T}\big]. \tag{3.22}$$

The difference between these approximations is that (3.20) calculates explicitly the derivative of the inverse marginal utilities, the interest rate and the market

price of risk and approximates the tangent process of the state variables by a finite difference, (3.21) calculates explicitly the derivative of the inverse marginal utilities and approximates the tangent process of the logarithmic state price density by a finite difference, while (3.22) approximates the whole functional in the conditional expectation, including the marginal utilities, by a finite difference.

The numerical implementation of MCFD estimators, such as (3.20)–(3.22), is similar to the implementation of MCMD estimators. The procedure estimates conditional expectations by first simulating M replications of the random variable within the expectation and then averaging over these replications. In most parametric examples the conditional distribution of the random variable of interest is unknown. A numerical discretization scheme, such as the Euler or the Milshtein schemes, based on N discretization points can nevertheless be used to obtain a convergent approximation. The MCFD estimator is then calculated by averaging independent replications of these simulated random variables. The choice of α_j gives different types of finite difference approximations. The estimator obtained from forward differences ($\alpha_j = 1$) is the MCFFD estimator, the estimator obtained from backward differences ($\alpha_j = 0$) is the MCBFD estimator, and the estimator based on central differences ($\alpha_j = 1/2$) the MCCFD estimator. As in the case of deterministic finite difference methods (i.e. finite difference methods for ODEs or PDEs) the computational cost is greater for MCCFD than for MCFFD or MCBFD estimators. This stems from the need to simulate two auxiliary processes with forward and backward perturbed initial values for MCCFD estimators. In contrast, MCFFD and MCBFD estimators only require the simulation of one auxiliary process with either forward, or backward perturbed initial value. A subsequent section will show the effect on the convergence properties of the methods.

Like MCC estimators, MCFD estimators are based on approximations of the conditional expectation in the hedging terms. Note, in particular, that MCFD estimators can be viewed as approximate MCMD estimators where the tangent process and therefore the Malliavin derivative has been approximated by a finite difference. The quality of the approximation will therefore depend on the additional convergence parameter τ. This additional structure will also affect the asymptotic error distribution of an MCFD estimator.

The finite difference methods described above are used to compute portfolio hedging components that depend on tangent processes. Although the mean–variance portfolio component also takes the form of an expectation, it does not involve Malliavin derivatives or tangent processes. It can therefore be calculated in a standard manner by simulating the underlying processes (using some suitable discretization scheme) and computing the relevant expectation using an average over independent replications.

3.5 *Monte Carlo regression* (Brandt et al., 2005)

The last method surveyed is an approximation method developed to solve discrete time portfolio choice problems. This approach, proposed by Brandt et al. (2005), is based on the standard recursive dynamic programming algorithm. It combines Monte Carlo simulation with a Taylor series approximation of the value function and a regression-based computation of conditional expectations in order to calculate approximate "optimal" policies. The methodology applies to large-scale problems with path-dependent and nonstationary dynamics as well as arbitrary utility functions. We summarize the main steps in the context of a pure portfolio problem (without intermediate utility).

The procedure is recursive in nature. It is based on the (discrete time) Bellman equation for the value function V of the dynamic portfolio problem,

$$V_t(X_t, Z_t) = \max_{\pi_t} \mathbf{E}_t\big[V_{t+1}\big(X_t\big(\pi_t' R_{t+1}^e + R^f\big), Z_{t+1}\big)\big], \tag{3.23}$$

where X_t, is the endogenous wealth at time t, Z_t, is a vector of exogenous state variables at t, R_{t+1}^e the vector of risky assets' excess returns from t to $t+1$, R^f the return on the risk-free asset and π_t is the portfolio. To keep matters simple we follow Brandt et al. (2005) and assume a constant interest rate R^f. The first-order conditions (FOC) for the portfolio choice problem are

$$\mathbf{E}_t\big[\partial_1 V_{t+1}\big(X_t\big(\pi_t' R_{t+1}^e + R^f\big), Z_{t+1}\big) R_{t+1}^e\big] = 0, \tag{3.24}$$

where $\partial_1 V_{t+1}$ is the derivative of the value function with respect to future wealth.

There are three steps which are as follows:

Step 1: Simplify the initial problem (3.23) by expanding the value function in a Taylor series around $X_t R^f$, the value at $t+1$ of current wealth. To account for skewness and kurtosis effects Brandt et al. (2005) propose the fourth-order expansion[6]

$$\begin{aligned} V_t^a(X_t, Z_t) = \max_{\pi_t} {} & \mathbf{E}_t\big[V_{t+1}^a\big(X_t R^f, Z_{t+1}\big)\big] \\ & + \mathbf{E}_t\big[\partial_1 V_{t+1}^a\big(X_t R^f, Z_{t+1}\big)\big(X_t \pi_t' R_{t+1}^e\big)\big] \\ & + \frac{1}{2}\mathbf{E}_t\big[\partial_1^2 V_{t+1}^a\big(X_t R^f, Z_{t+1}\big)\big(X_t \pi_t' R_{t+1}^e\big)^2\big] \\ & + \frac{1}{6}\mathbf{E}_t\big[\partial_1^3 V_{t+1}^a\big(X_t R^f, Z_{t+1}\big)\big(X_t \pi_t' R_{t+1}^e\big)^3\big] \\ & + \frac{1}{24}\mathbf{E}_t\big[\partial_1^4 V_{t+1}^a\big(X_t R^f, Z_{t+1}\big)\big(X_t \pi_t' R_{t+1}^e\big)^4\big] \end{aligned}$$

[6] Brandt et al. (2005) report that a fourth-order expansion around $X_t R^f$ gives very accurate results for the particular problems that they considered.

where V^a is the value function for this new (approximate) problem. Let π_t^a be the solution of the approximate problem. The FOC leads to the following implicit expression for π^a,

$$\begin{aligned}\pi_t^a &= -\{\mathbf{E}_t[\partial_1^2 V_{t+1}^a(X_t R^f, Z_{t+1}) R_{t+1}^e (R_{t+1}^e)'] X_t^2\}^{-1} \\ &\quad \times \Big\{\mathbf{E}_t[\partial_1 V_{t+1}^a(X_t R^f, Z_{t+1}) R_{t+1}^e] X_t \\ &\quad + \frac{1}{2}\mathbf{E}_t[\partial_1^3 V_{t+1}^a(X_t R^f, Z_{t+1})((\pi_t^a)' R_{t+1}^e)^2 R_{t+1}^e] X_t^3 \\ &\quad + \frac{1}{6}\mathbf{E}_t[\partial_1^4 V_{t+1}^a(X_t R^f, Z_{t+1})((\pi_t^a)' R_{t+1}^e)^3 R_{t+1}^e] X_t^4\Big\} \\ &\equiv -\{\mathbf{E}_t[B_{t+1}] X_t\}^{-1}\{\mathbf{E}_t[A_{t+1}] + \mathbf{E}_t[C_{t+1}(\pi_t^a)] X_t^2 \\ &\quad + \mathbf{E}_t[D_{t+1}(\pi_t^a)] X_t^3\}. \end{aligned} \tag{3.25}$$

The structure of (3.25) shows that the solution depends on conditional moments involving the derivatives of the approximate value function and powers of the returns. Assume for now that these moments can be calculated by some procedure. The solution of (3.25) is then computed as follows:

(a) calculate the solution of the quadratic problem corresponding to the second-order expansion of the value function. This gives an explicit expression which can be used as an initial guess for solving (3.25),
(b) substitute this initial guess into the right-hand side of (3.25) to produce a new estimate of π^a on the left-hand side,
(c) iterate by repeating the previous step until consecutive estimates become close enough, i.e. the distance between consecutive estimates falls below some pre-selected tolerance level.

Step 2: Simulate a large number of sample paths of the vector $Y_t = [R_t^e, Z_t]$. This set of paths serves as the underlying tree for the application of a recursive procedure where the portfolio is approximated at each step, along each trajectory, by the solution of (3.25).

Step 3: Proceed recursively, along each trajectory, starting from the terminal date. To compute the approximate portfolio at date t proceed in the following manner. Suppose that approximate weights π_s^a for $s = t+1, \ldots, T-1$ have been found. Terminal wealth starting from $X_t^a R^f$ at $t+1$ is

$$X_T^a = X_t^a R^f \prod_{s=t+1}^{T-1} (\pi_s^a R_{s+1}^e + R^f). \tag{3.26}$$

The coefficients in (3.25) can then be approximated by

$$A_{t+1} \approx \mathbf{E}_{t+1}\left[\partial u\big(X_T^a\big)\prod_{s=t+1}^{T-1}\big(\pi_s^a R_{s+1}^e + R^f\big)\right]R_{t+1}^e \tag{3.27}$$

in the case of A_{t+1}, and similar expressions for B_{t+1}, C_{t+1} and D_{t+1}. Let $a_{t+1} \equiv \partial u(X_T^a)\prod_{s=t+1}^{T-1}(\pi_s^a R_{s+1}^e + R^f)R_{t+1}^e$ be the random variable inside the expectation in (3.27) and define b_{t+1}, c_{t+1} and d_{t+1} in a similar manner. Then

$$\begin{aligned}\pi_t^a \approx -\big\{\mathbf{E}_t[b_{t+1}]X_t^a\big\}^{-1}\big\{&\mathbf{E}_t[a_{t+1}] + \mathbf{E}_t\big[c_{t+1}\big(\pi_t^a\big)\big]\big(X_t^a\big)^2\\ &+ \mathbf{E}_t\big[d_{t+1}\big(\pi_t^a\big)\big]\big(X_t^a\big)^3\big\}.\end{aligned} \tag{3.28}$$

This approximation is treated as an exact equality to find π_t^a (in fact this construction produces an approximation of the approximate policy π_t^a). To calculate the conditional expectations of a, b, c, d the regression method of Longstaff and Schwartz (2001) is used. This simple approach uses regressions across the simulated paths to evaluate conditional expectations. Let y be a typical element of the vector $[a, b, c, d]$. The expectation of y_{t+1} is computed by regressing y_{t+1} on a vector of polynomial bases in the state variables Z_t so that,

$$\mathbf{E}_t[y_{t+1}] = \varphi(Z_t)'k_t,$$

where k_t is the vector of regression parameters, and the ith element of $\varphi(Z_t)$ corresponds to the ith term of a polynomial in Z_t of order K. The fitted values of this regression are used to construct estimates of the time t-conditional expectations of a_{t+1}, b_{t+1}, c_{t+1} and d_{t+1}, along each path m. Solving (3.28) produces an approximate portfolio $\pi_t^{a,m}$.

4 Asymptotic properties of portfolio estimators

This section describes the asymptotic error distributions of MCMD, MCC and MCFD portfolio estimators and discusses convergence issues for MCR. The results provided extend Detemple et al. (2005b, 2005c, 2005d) to settings with both running utility and utility of terminal wealth and to smooth utility functions outside the power (constant relative risk averse) class.

4.1 Notation and assumptions

Throughout the section utility functions are assumed to be smooth in the sense that $u, U \in \mathcal{C}^5$, the space of five-times continuously differentiable functions. In addition, marginal utilities satisfy the Inada conditions

$$\lim_{x\to 0} u'(x,t) = +\infty \quad \text{and} \quad \lim_{x\to 0} U'(x,T) = +\infty. \tag{4.1}$$

Let $\{t_n\colon n = 0, \ldots, N-1\}$ be an equidistant discretization of the time interval $[t, T]$, with $\Delta \equiv t_{n+1} - t_n = (T-t)/N$. To state some of the results it proves useful to introduce the notation $\eta_v^N \equiv [Nv]/N$ for $v \in [0, T]$ if $Nv \notin \mathbb{N}$ and $\eta_v^N = v - 1/N$ otherwise, where $[Nv]$ is the integer part of Nv. With this definition sums can be written as integrals, e.g.

$$\sum_{n=0}^{N-1} f_{t_n} \Delta \equiv \int_t^T f_{\eta_v^N} \, \mathrm{d}v.$$

For empirical means write $\mathbf{E}^M[U] \equiv (\sum_{i=1}^M U^i)/M$, where the random variables U^i are i.i.d. replications of U.

Given that analytic formulas for the distributions associated with diffusions are usually unknown, a numerical scheme is required in order to approximate the solutions of SDEs. Let X^i be a random variable associated with the solution of an SDE and $X^{i,N}$ an approximation based on N discretization points. The notion of weak convergence is employed to assess the behavior of the approximation: the sequence $X^{i,N}$ is said to converge *weakly* to X^i as the number of discretization points N goes to infinity if and only if $\mathbf{E}[f(X^{i,N})] \to \mathbf{E}[f(X^i)]$ for all continuous, bounded functions $f \in \mathcal{C}_b$.

For a parsimonious representation of portfolios it also proves useful to define the shadow price of optimal wealth. This is the function $y_t^* \equiv y^*(t, X_t^*, Y_t)$ that is the unique solution of the nonlinear equation

$$X_t^* = \mathbf{E}_t\left[\int_t^T \xi_{t,v} I(y_t^* \xi_{t,v}, v) \, \mathrm{d}v + \xi_{t,T} J(y_t^* \xi_{t,T}, T)\right]. \tag{4.2}$$

The right-hand side of this equation is the present value of optimal consumption, post date t. Decreasing marginal utility and the Inada conditions (4.1) ensure that the shadow price y_t^* exists and is unique for all $X_t^* > 0$. Further, note that $y_t^* = y^* \xi_t$ where y^* corresponds to the initial shadow price of wealth defined previously.

4.2 Expected approximation errors

Let us now record a general result for expected approximation errors. Suppose the d_z-dimensional process Z satisfies the SDE

$$\mathrm{d}Z_t = a(Z_t) \, \mathrm{d}t + \sum_{j=1}^d b_j(Z_t) \, \mathrm{d}W_t^j; \quad Z_0 \text{ given}, \tag{4.3}$$

whose coefficients a, b satisfy Lipschitz and growth conditions so as to guarantee the existence and uniqueness of a solution. Let Z^N be the numerical solution of (4.3) based on the Euler scheme with N discretization points.

To describe the expected approximation error it is convenient to define the tangent process of the diffusion Z (see Section A.8), as

$$\nabla_{t,z} Z_v = \mathcal{E}^R\left(\int_t^{\cdot} \partial a(Z_s)\,\mathrm{d}s + \sum_{j=1}^{d} \int_t^{\cdot} \partial b_j(Z_s)\,\mathrm{d}W_s^j \right)_v,$$

where $\mathcal{E}^R(\cdot)$ is the right stochastic exponential (i.e. the solution of $d\mathcal{E}^R(M)_v = dM_v \mathcal{E}^R(M)_v$). Also for a function $f \in \mathcal{C}^3(\mathbb{R}^{d_z})$ define the random variables

$$\begin{aligned} V_1(t,v) \equiv -\nabla_{t,z} Z_v \int_t^v (\nabla_{t,z} Z_s)^{-1} \Bigg(& \partial a(Z_s)\,\mathrm{d}Z_s \\ & + \sum_{j=1}^{d} \left[\partial b_j a - \sum_{i=1}^{d} (\partial b_j)(\partial b_j) b_i \right](Z_s)\,\mathrm{d}W_s^j \Bigg) \\ & + \nabla_{t,z} Z_v \int_t^v (\nabla_{t,z} Z_s)^{-1} \\ & \times \sum_{j=1}^{d} \left[\partial b_j \partial b_j a - \sum_{k,l=1}^{d} \partial_k (\partial_l a b_{l,j}) b_{k,j} \right](Z_s)\,\mathrm{d}s \\ & + \nabla_{t,z} Z_v \int_t^v (\nabla_{t,z} Z_s)^{-1} \\ & \times \sum_{i,j=1}^{d} \left[\partial(\partial b_j \partial b_j b_i) b_i - \partial b_i \partial b_j \partial b_j b_i \right](Z_s)\,\mathrm{d}s \end{aligned} \tag{4.4}$$

and

$$V_2(t,v) \equiv -\int_t^v \sum_{i,j=1}^{d} \nu_{i,j}(s,v)\,\mathrm{d}s, \tag{4.5}$$

where

$$\begin{aligned} & \nu_{i,j}(s,v) \equiv \left[h^{i,j} (\nabla_{t,z} Z)^{-1} \left[(\partial b_j) b_i \right](Z), W^i \right]_s \quad \text{with} \\ & h_t^{i,j} \equiv \mathbf{E}_t\left[\mathcal{D}_{jt}\left(\partial f(Z_T) \nabla_{t,z} Z_T e_i \right) \right] \end{aligned} \tag{4.6}$$

and e_i is the ith unit vector. A more explicit expression for $\nu_{i,j}(s,v)$ is given in Detemple et al. (2005c). Finally, for $v \in [t,T]$, define the conditional expectations

$$K_{t,v}(Z_t) \equiv \frac{1}{2} \mathbf{E}_t\left[\partial f(Z_v) V_1(t,v) + V_2(t,v) \right], \tag{4.7}$$

$$k_{t,v}(Z_t) \equiv -\mathbf{E}_t\left[\int_t^v \left(\partial f(Z_s)\left[a + \sum_{j=1}^d (\partial b_j) b_j\right](Z_s) + \sum_{j=1}^d \left[b_j' \partial^2 f b_j\right](Z_s)\right) \mathrm{d}s\right] \tag{4.8}$$

and set

$$\kappa_{t,v}(Z_t) \equiv K_{t,v}(Z_t) + k_{t,v}(Z_t). \tag{4.9}$$

With this notation we can state the following

Proposition 3. *Let $f \in \mathcal{C}^3(\mathbb{R}^{d_z})$ be such that the uniform integrability conditions*

$$\lim_{r\to\infty} \limsup_N \mathbf{E}_t\left[\mathbf{1}_{\{\|N(f(Z_T^N)-f(Z_T))\|>r\}} N \left\|f\left(Z_T^N\right) - f(Z_T)\right\|\right] = 0, \tag{4.10}$$

$$\lim_{r\to\infty} \limsup_N \mathbf{E}_t\left[\mathbf{1}_{\{\|N\int_t^T (f(Z^N_{\eta_v^N})-f(Z_v))\,\mathrm{d}v\|>r\}} N \times \left\|\int_t^T \left(f\left(Z^N_{\eta_v^N}\right) - f(Z_v)\right)\mathrm{d}v\right\|\right] = 0 \tag{4.11}$$

hold (**P**-*a.s.*). *Then, as $N \to \infty$,*

$$N\mathbf{E}_t\left[f\left(Z_T^N\right) - f(Z_T)\right] \to \frac{1}{2} K_{t,T}(Z_t), \tag{4.12}$$

$$N\mathbf{E}_t\left[\int_t^T f\left(Z^N_{\eta_v^N}\right)\mathrm{d}v - \int_t^T f(Z_v)\,\mathrm{d}v\right] \to \frac{1}{2}\int_t^T \kappa_{t,v}(Z_t)\,\mathrm{d}v \tag{4.13}$$

with $K_{t,\cdot}(Z_t)$, $\kappa_{t,\cdot}(Z_t)$ as defined in (4.7), (4.9).

This proposition provides formulas for the expected errors in the approximation of functions (or functionals) evaluated at solutions of stochastic differential equations. The expressions obtained can be viewed as probabilistic representations of the formulas in Talay and Tubaro (1990) that give the errors in terms of conditional expectations of functions solving PDEs. As will become clear below expected approximation errors appear in the second-order bias terms associated with the efficient Monte Carlo estimators of conditional expectations of functions of diffusions. The formulas in Proposition 3 can be used to estimate second-order biases and infer second-order bias corrected estimators.

4.3 Asymptotic error distribution of MCMD estimators

With y_t^* as the solution of (4.2), the MCMD portfolio estimator can be written as

$$\begin{aligned}\widehat{X_t^*\pi_t^*}^{N,M} &= -(\sigma_t')^{-1}\theta_t\Bigg(\mathbf{E}_t^M\big[g_1^{MV}\big(Z_{t,T}^N; y_t^*\big)\big] \\ &\quad + \mathbf{E}_t^M\Bigg[\int_t^T g_2^{MV}\big(Z_{t,\eta_v^N}^N; y_t^*\big)\,\mathrm{d}v\Bigg]\Bigg) \\ &\quad - (\sigma_t')^{-1}\Bigg(\mathbf{E}_t^M\big[g_1^H\big(Z_{t,T}^N; y_t^*\big)\big] \\ &\quad + \mathbf{E}_t^M\Bigg[\int_t^T g_2^H\big(Z_{t,\eta_v^N}^N; y_t^*\big)\,\mathrm{d}v\Bigg]\Bigg). \end{aligned} \tag{4.14}$$

In this expression $\{Z_{t,v}^N\colon\ v \in [t,T]\}$ is a numerical approximation of the d_z-dimensional process $\{Z_{t,v}' \equiv [\xi_{t,v}, H_{t,v}', \mathrm{vec}(\mathcal{D}_t Y_v)', Y_v', v]\colon\ v \in [t,T]\}$, with $d_z = 2 + d(k+1) + k$ and H as defined in (2.16). The process $Z_{t,v}$ solves

$$\mathrm{d}Z_{t,v} = a(Z_{t,v})\,\mathrm{d}v + \sum_{j=1}^d b_j(Z_{t,v})\,\mathrm{d}W_v^j; \quad Z_{t,t} \text{ given.}$$

The functions $g_1^{MV}, g_1^H, g_2^{MV}, g_2^H$ are $\mathcal{C}^3$-functions that determine various portfolio components and are defined by

$$\begin{aligned} g_1^{MV}(z;y) &\equiv z_1 J'(yz_1, z_5); \qquad g_1^H(z;y) \equiv z_1 J'(yz_1, z_5) z_2, \\ g_2^{MV}(z;y) &\equiv z_1 I'(yz_1, z_5); \qquad g_2^H(z;y) \equiv z_1 I'(yz_1, z_5) z_2. \end{aligned}$$

Close inspection reveals that g_1^{MV}, g_1^H are portfolio demand components associated with terminal wealth, while g_2^{MV}, g_2^H relate to intermediate consumption.

Each portfolio component gives rise to an error term. To study the convergence properties of the joint error define

$$e_{1,t,T}^{MV,M,N} \equiv -\big(\mathbf{E}_t^M\big[g_1^{MV}\big(Z_{t,T}^N; y_t^*\big)\big] - \mathbf{E}_t\big[g_1^{MV}(Z_{t,T}; y_t^*)\big]\big)(\sigma_t')^{-1}\theta_t, \tag{4.15}$$

$$e_{1,t,T}^{H,M,N} \equiv -(\sigma_t')^{-1}\big(\mathbf{E}_t^M\big[g_1^H\big(Z_{t,T}^N; y_t^*\big)\big] - \mathbf{E}_t\big[g_1^H(Z_{t,T}; y_t^*)\big]\big), \tag{4.16}$$

$$e_{2,t,T}^{MV,M,N} \equiv -\left(\mathbf{E}_t^M\left[\int_t^T g_2^{MV}\big(Z_{t,\eta_v^N}^N; y_t^*\big)\,\mathrm{d}v\right] - \mathbf{E}_t\left[\int_t^T g_2^{MV}(Z_{t,v}; y_t^*)\,\mathrm{d}v\right]\right)(\sigma_t')^{-1}\theta_t, \tag{4.17}$$

$$e_{2,t,T}^{H,M,N} \equiv -(\sigma_t')^{-1}\left(\mathbf{E}_t^M\left[\int_t^T g_2^{H}\big(Z_{t,\eta_v^N}^N; y_t^*\big)\,\mathrm{d}v\right] - \mathbf{E}_t\left[\int_t^T g_2^{H}(Z_{t,v}; y_t^*)\,\mathrm{d}v\right]\right). \tag{4.18}$$

For $j \in \{1,2\}$, let $(e_{j,t,T}^{M,N})' = [(e_{j,t,T}^{MV,M,N})', (e_{j,t,T}^{H,M,N})']$ be the $1 \times 2d$ random vector of approximation errors associated with the mean–variance and hedging demands for terminal wealth ($j = 1$) and intermediate consumption ($j = 2$). Finally, let $(e_{t,T}^{M,N})' = [(e_{1,t,T}^{M,N})', (e_{2,t,T}^{M,N})']$ be the $1 \times 4d$ vector that incorporates all the portfolio components. Similarly, define the $1 \times 4d$ random vector $C_{t,T}' \equiv [C_{1,t,T}', C_{2,t,T}']$ where

$$C_{1,t,T}' \equiv \big[-g_1^{MV}(Z_{t,T}; y_t^*)\theta_t'\sigma_t^{-1}, -g_1^{H}(Z_{t,T}; y_t^*)'\sigma_t^{-1}\big],$$

$$C_{2,t,T}' \equiv \left[-\int_t^T g_2^{MV}(Z_{t,v}; y_t^*)\,\mathrm{d}v\,\theta_t'\sigma_t^{-1}, -\int_t^T g_2^{H}(Z_{t,v}; y_t^*)'\,\mathrm{d}v\,\sigma_t^{-1}\right]$$

are random variables involved in the various portfolio components. The random variable $C_{t,T}$ plays a critical role for the joint variance of the asymptotic error distribution.

Let $\mathbb{D}^{1,2}$ be the space of random variables for which Malliavin derivatives are defined (see Section A.3). Our next proposition describes the asymptotic behavior of the estimation error.

Proposition 4. *Suppose $g \in \mathcal{C}^3(\mathbb{R}^{d_z})$ and $g(Z_{t,v}; y_t^*) \in \mathbb{D}^{1,2}$ for all $v \in [t,T]$. Also suppose that the assumptions of Proposition 3 hold, and that*

$$\lim_{r\to\infty} \mathbf{E}_t\big[\mathbf{1}_{\{\|g_j^\alpha(Z_{t,v};y_t^*)-\mathbf{E}_t[g_j^\alpha(Z_{t,v};y_t^*)]\|>r\}} \times \big\|g_j^\alpha(Z_{t,v}; y_t^*) - \mathbf{E}_t\big[g_j^\alpha(Z_{t,v}; y_t^*)\big]\big\|^2\big] = 0 \tag{4.19}$$

for all $j \in \{1, 2\}$ *and* $\alpha \in \{MV, H\}$. *Then, as* $M \to \infty$,

$$\sqrt{M} e_{t,T}^{M,N_M} \Rightarrow \epsilon^{md} \frac{1}{2} \begin{bmatrix} -K_{1,t,T}^{MV}(Y_t; y_t^*)(\sigma_t')^{-1}\theta_t \\ -(\sigma_t')^{-1}[K_{i,1,t,T}^{H}(Y_t; y_t^*)]_{i=1,\dots,d} \\ -\int_t^T \kappa_{2,t,v}^{MV}(Y_t; y_t^*)\,\mathrm{d}v(\sigma_t')^{-1}\theta_t \\ -(\sigma_t')^{-1}\int_t^T [\kappa_{i,2,t,v}^{H}(Y_t; y_t^*)]_{i=1,\dots,d}\,\mathrm{d}v \end{bmatrix} + \begin{bmatrix} L_{1,t,T}^{MV}(Y_t; y_t^*) \\ L_{1,t,T}^{H}(Y_t; y_t^*) \\ L_{2,t,T}^{MV}(Y_t; y_t^*) \\ L_{2,t,T}^{H}(Y_t; y_t^*) \end{bmatrix}, \tag{4.20}$$

where $N_M \to \infty$, *as* $M \to \infty$, $\epsilon^{md} = \lim_{M\to\infty} \sqrt{M}/N_M$ *and*

$$L_{t,T}(Y_t; y_t^*)' \equiv \big[L_{1,t,T}^{MV}(Y_t; y_t^*)', L_{1,t,T}^{H}(Y_t; y_t^*)', L_{2,t,T}^{MV}(Y_t; y_t^*)', L_{2,t,T}^{H}(Y_t; y_t^*)'\big] \tag{4.21}$$

is the terminal value of a Gaussian martingale with (*deterministic*) *quadratic variation and conditional variance given by*

$$[L, L]_{t,T}(Y_t; y_t^*) = \int_t^T \mathbf{E}_t\big[N_s(N_s)'\big]\,\mathrm{d}s = \mathbf{VAR}_t[C_{t,T}], \tag{4.22}$$

$$N_s = \mathbf{E}_s[\mathcal{D}_s C_{t,T}]. \tag{4.23}$$

The mean–variance component associated with terminal wealth $g_1^{MV}(z; y_t^*)$ *induces the second-order bias function* $K_{1,t,T}^{MV}(Y_t; y_t^*)$ *as defined in* (4.7). *The components of the d-dimensional vector of hedging terms for terminal wealth* $[g_1^H(z; y_t^*)]_i$ *induce the second-order bias functions* $K_{i,1,t,T}^{H}(Y_t; y_t^*)$ *as given in* (4.7). *In contrast, the mean–variance component for running consumption* $g_2^{MV}(z; y_t^*)$ *induces two second-order biases embedded in the function* $\kappa_{2,t,v}^{MV}(Y_t; y_t^*)$ *as given in* (4.9). *Similarly, the components of the d-dimensional vector of hedging terms for running consumption* $[g_2^H(z; y_t^*)]_i$ *induce the second-order bias functions* $\kappa_{i,2,t,v}^{H}(Y_t; y_t^*)$ *defined in* (4.9).

The expression for the asymptotic error distribution (4.20) has two components. The first one depends on the expected approximation error and corresponds to the second-order bias of the estimator. To illustrate the role of the parameter ϵ^{md}, and the second-order bias, note that for $i = 1, \dots, d$ confidence intervals with coverage probability $1 - \alpha$, calculated on the basis of the Gaussian process L, are

$$\big[\psi_i^-(M, N_M, \alpha), \psi_i^+(M, N_M, \alpha)\big]$$

where

$$\psi_i^{\pm}(M, N_M, \alpha) \equiv \sqrt{M}\widehat{\pi_{it}^* X_t^*}^{M,N_M} \pm \Phi^{-1}(\alpha/2)\frac{\sigma_{ii}^{M,N_M}}{\sqrt{M}}$$

with Φ the cumulative Gaussian distribution function and σ_{ii}^{M,N_M} a convergent estimator of the variance of the Gaussian martingale $[L_{t,T}]_i$ in (4.21). As $M \to \infty$ the true coverage probability of this interval converges to

$$\mathbf{P}\big(\pi_{it}^* X_t^* \in \big[\psi_i^-(M, N_M, \alpha), \psi_i^+(M, N_M, \alpha)\big]\big) \to \Psi\big(\alpha, \delta_i^{md}\big), \quad (4.24)$$

with

$$\Psi(\alpha, x) \equiv \Phi\big(\Phi^{-1}\big((1-\alpha)/2\big) - x\big) - \Phi\big(\Phi^{-1}(\alpha/2) - x\big), \quad (4.25)$$

$$\delta_i^{md} \equiv \frac{1}{2}\epsilon^{md}\big[\mathbf{VAR}_t[L_{t,T}]^{-\frac{1}{2}} K_{t,T}(Y_t; y_t^*)\big]_i \quad (4.26)$$

and

$$\begin{aligned} K_{t,T}(Y_t; y_t^*) \equiv & -K_{1,t,T}^{MV}(Y_t; y_t^*)(\sigma_t')^{-1}\theta_t - (\sigma_t')^{-1}K_{1,t,T}^{H}(Y_t; y_t^*) \\ & - \int_t^T \kappa_{2,t,v}^{MV}(Y_t; y_t^*)\,\mathrm{d}v(\sigma_t')^{-1}\theta_t \\ & - (\sigma_t')^{-1}\int_t^T \kappa_{2,t,v}^{H}(Y_t; y_t^*)\,\mathrm{d}v, \end{aligned} \quad (4.27)$$

where the $d \times 1$ vectors of second-order biases associated with the hedging demands for terminal wealth $K_{1,t,v}^H$ and running consumption $\kappa_{2,t,v}^H$ are given by $K_{1,t,v}^H(Y_t; y_t^*)' \equiv [K_{1,1,t,v}^H(Y_t; y_t^*), \ldots, K_{d,1,t,v}^H(Y_t; y_t^*)]$ and $\kappa_{2,t,v}^H(Y_t; y_t^*)' \equiv [\kappa_{1,2,t,v}^H(Y_t; y_t^*), \ldots, \kappa_{d,2,t,v}^H(Y_t; y_t^*)]$.

The limit (4.24) shows that a confidence interval of nominal size α, based on L, will suffer from size distortion as it will in fact cover the true value $\pi_{it}^* X_t^*$ only with probability $\Psi(\alpha, \delta_i^{md})$ and not $1-\alpha$, as initially prescribed. The degree of size distortion is measured by the distance

$$s\big(\delta_i^{md}\big) \equiv 1 - \alpha - \Psi\big(\alpha, \delta_i^{md}\big).$$

Given that $\Psi(\alpha, \cdot)$ is strictly monotone and $\Psi(\alpha, 0) = 1-\alpha$, a confidence interval has the requested nominal size if and only if there is no second-order bias, i.e. $\delta_i^{md} = 0$. In the univariate case $d = 1$, the degree of size distortion $s(\delta_1^{md})$ is negatively related to $\mathbf{VAR}_t[L_{t,T}]$, the asymptotic variance implied by the Monte Carlo averaging procedure and positively related to $\frac{\epsilon^{md}}{2}K_{t,T}(Y_t; y_t^*)$, the second-order bias implied by the discretization scheme used for the resolution of SDEs.

When $\delta_i^{md} \neq 0$ efficiency comparisons based on the length of asymptotic confidence intervals $\psi_i^+(M, N_M, \alpha) - \psi_i^-(M, N_M, \alpha)$ are invalid, because the asymptotic coverage probability is less than the requested nominal size. Conclusions pertaining to the effects of various parameters should also be drawn with care. For instance in the univariate case, note that a reduction in the variance of an estimator has two effects. On the one hand, it reduces the length of a confidence interval. On the other hand, it also, if a second-order bias exists, increases the size distortion and therefore reduces the effective coverage probability. This trade-off also appears when the numbers of replications M and discretization points N are modified. If the variance of an estimator is reduced by increasing M, leaving N unchanged, efficiency may appear to improve when in fact the effective coverage probability decreases. Alternatively, if for a fixed budget of computation time, the number of discretization points N becomes large (thus, the number of replications M goes to zero), the resulting confidence interval of the estimator becomes free of size distortion (as $\epsilon^{md} = \lim_{M\to\infty} \sqrt{M}/N_M = 0$) but its length explodes (as $\sigma_{ii}^{M,N_M}/\sqrt{M} \to \infty$).

The trade-off between the effects of M and N implies that the efficient scheme is such that the number of Monte Carlo replications must be quadrupled whenever the number of discretization points is doubled (because $\epsilon^{md} = \lim_{M\to\infty} \sqrt{M}/N_M$). In addition, the asymptotic second-order bias has to be taken into account in order to draw valid efficiency comparisons. Methods to correct for the second-order bias require the calculation of the function K. Expressions for bias corrected estimators are provided in Detemple et al. (2005c).

A similar result applies to MCMD estimators based on the Doss transformation (see Section 3.2). The use of the Doss transformation increases the rate of convergence of the Euler scheme, but not the rate of convergence of the expected approximation error. The associated portfolio estimator converges at the same speed as the estimator based on the Euler scheme without Doss transformation, has the same asymptotic covariance matrix but a different second-order bias. Likewise, using the Milshtein scheme does not improve the rate of convergence and produces an asymptotic error distribution with the same covariance matrix. The sole modification is the expression for the second-order bias (see Detemple et al., 2005c for details).

4.4 Asymptotic properties of MCC estimators

MCC estimators are described in (3.14)–(3.16). In what follows we examine the error behavior for (3.16). Similar convergence results hold for estimators based on (3.14) and (3.15). With the definitions

$$F_{t,T} \equiv f_1(Z_{t,T}; y_t^*) + \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v \tag{4.28}$$

where

$$f_1(z; y) = z_1 J(yz_1, z_5), \tag{4.29}$$

$$f_2(z; y) = z_1 I(yz_1, z_5) \tag{4.30}$$

the estimation error is $(e_{t,T}^{M,N,h})' = [(e_{1,t,T}^{M,N,h})', (e_{2,t,T}^{M,N,h})']$ with

$$\begin{aligned} e_{1,t,T}^{M,N,h} \equiv (\sigma_t')^{-1} \Bigg(& \frac{1}{h} \mathbf{E}_t^M \big[f_1\big(Z_{t,T}^N; y_t^*\big) \Delta_h W_t \big] \\ & - \big(\mathbf{E}_t \big[\mathcal{D}_t f_1(Z_{t,T}; y) \big]_{|y=y_t^*} \big)' \Bigg), \end{aligned} \tag{4.31}$$

$$\begin{aligned} e_{2,t,T}^{M,N,h} \equiv (\sigma_t')^{-1} \Bigg(& \frac{1}{h} \mathbf{E}_t^M \Bigg[\int_t^T f_2\big(Z_{t,\eta_v^N}^N; y_t^*\big) \, \mathrm{d}v \, \Delta_h W_t \Bigg] \\ & - \Bigg(\mathbf{E}_t \Bigg[\mathcal{D}_t \int_t^T f_2(Z_{t,v}; y) \, \mathrm{d}v \Bigg]_{|y=y_t^*} \Bigg)' \Bigg), \end{aligned} \tag{4.32}$$

and $\Delta_h W_t \equiv W_{t+h} - W_t$. The asymptotic behavior of the error is described in the next proposition.

Proposition 5. *Assume that $f_1, f_2 \in \mathcal{C}^3(\mathbb{R}^{d_z})$ and suppose that $f_i(Z_{t,v}; y_t^*) \in \mathbb{D}^{1,2}$ for $i = 1, 2$. Let $K_{1,t,v}(Y_t; y_t^*)$ be defined for f_1 as in (4.7) and $\kappa_{2,t,v}(Y_t; y_t^*)$ for f_2 as in (4.9). Define the events*

$$F_1(N, h, r) = \Bigg\{ \Bigg\| Q_{1,t,T}^{N,h}(y_t^*) - \frac{1}{2} \mathbf{E}_t \big[Q_{1,t,T}^{N,h}(y_t^*) \big] \Bigg\| > r \Bigg\},$$

$$F_2(N, h, r) = \Bigg\{ \Bigg\| \int_t^T \Bigg(Q_{2,t,\eta_v^N}^{N,h}(y_t^*) - \frac{1}{2} \mathbf{E}_t \big[Q_{2,t,\eta_v^N}^{N,h}(y_t^*) \big] \Bigg) \mathrm{d}v \Bigg\| > r \Bigg\},$$

$$G_1(h, r) = \Bigg\{ \Bigg\| f_1(Z_{t,T}; y_t^*) \frac{\Delta_h W_t}{h} - \mathbf{E}_t \Bigg[f_1(Z_{t,T}; y_t^*) \frac{\Delta_h W_t}{h} \Bigg] \Bigg\| > r \Bigg\},$$

$$\begin{aligned} G_2(h, r) = \Bigg\{ \Bigg\| & \int_t^T f_2(Z_{t,v}; y_t^*) \, \mathrm{d}v \frac{\Delta_h W_t}{h} \\ & - \mathbf{E}_t \Bigg[\int_t^T f_2(Z_{t,v}; y_t^*) \, \mathrm{d}v \frac{\Delta_h W_t}{h} \Bigg] \Bigg\| > r \Bigg\} \end{aligned}$$

where, for $j = 1, 2$, the processes $Q_{j,t,\cdot}^{N,h}$ are given by

$$Q_{j,t,v}^{N,h}(y_t^*) \equiv N\big(f_j\big(Z_{t,v}^N; y_t^*\big) - f_j(Z_{t,v}; y_t^*) \big) \frac{\Delta_h W_t}{h}. \tag{4.33}$$

Suppose that the conditions

$$\lim_{r\to\infty}\limsup_{h,N}\mathbf{E}_t\left[\mathbf{1}_{F_1(N,h,r)}\left\|Q^{N,h}_{1,t,T}(y^*_t)-\frac{1}{2}\mathbf{E}_t\big[Q^{N,h}_{1,t,T}(y^*_t)\big]\right\|\right]=0, \tag{4.34}$$

$$\lim_{r\to\infty}\limsup_{h,N}\mathbf{E}_t\left[\mathbf{1}_{F_2(N,h,r)}\left\|\int_t^T\left(Q^{N,h}_{2,t,\eta^N_v}(y^*_t)-\frac{1}{2}\mathbf{E}_t\big[Q^{N,h}_{2,t,\eta^N_v}(y^*_t)\big]\right)\mathrm{d}v\right\|\right]=0, \tag{4.35}$$

$$\lim_{r\to\infty}\limsup_{h}\mathbf{E}_t\left[\mathbf{1}_{G_1(h,r)}\left\|f_1(Z_{t,T};y^*_t)\frac{\Delta_h W_t}{h}-\mathbf{E}_t\left[f_1(Z_{t,T};y^*_t)\frac{\Delta_h W_t}{h}\right]\right\|^2\right]=0, \tag{4.36}$$

$$\lim_{r\to\infty}\limsup_{h}\mathbf{E}_t\left[\mathbf{1}_{G_2(h,r)}\left\|\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v\frac{\Delta_h W_t}{h}-\mathbf{E}_t\left[\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v\frac{\Delta_h W_t}{h}\right]\right\|^2\right]=0 \tag{4.37}$$

hold. Then, as $M\to\infty$,

$$\begin{aligned}M^{1/3}e^{M,N_M,h_M}_{t,T}\Rightarrow{}&\varepsilon_1^c(\sigma_t')^{-1}\left(\partial_s\mathbf{E}_s\begin{bmatrix}\mathcal{D}_s f_1(Z_{t,T};y^*_t)\\ \mathcal{D}_s\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v\end{bmatrix}\right)'_{|s=t}\\&+\varepsilon_2^c\frac{1}{2}(\sigma_t')^{-1}\left(\mathcal{D}_t\begin{bmatrix}K_{1,t,T}(Y_t;y)\\ \int_t^T\kappa_{2,t,v}(Y_t;y)\,\mathrm{d}v\end{bmatrix}\right)'_{|y=y^*_t}\\&+\big(I_2\otimes(\sigma_t')^{-1}\big)O_{t,T}(Y_t;y^*_t),\end{aligned} \tag{4.38}$$

where $\otimes$ *denotes the Kronecker product.*[7] *In* (4.38) *the convergence parameters satisfy* $N_M, 1/h_M\to\infty$ *when* $M\to\infty$, *and the constants are* $\varepsilon_1^c=\lim_{M\to\infty}M^{1/3}h_M$ *and* $\varepsilon_2^c=\lim_{M\to\infty}M^{1/3}/N_M$. *The* $2d$*-dimensional Gaussian martingale* $O_{t,T}$ *has* (*deterministic*) *quadratic variation*

$$\begin{aligned}&[O,O]_{t,T}(Y_t;y^*_t)\\&\quad=\mathbf{E}_t\begin{bmatrix}f_1(Z_{t,T};y^*_t)^2I_d & f_1(Z_{t,T};y^*_t)\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v\,I_d\\ f_1(Z_{t,T};y^*_t)\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v\,I_d & (\int_t^T f_2(Z_{t,v};y^*_t)\,\mathrm{d}v)^2I_d\end{bmatrix}.\end{aligned}$$

[7] The Kronecker product of an $m\times n$ matrix A and a $p\times q$ matrix B corresponds to the $mp\times nq$ matrix $A\otimes B\equiv[A_{ij}B]_{i=1,\dots,m}^{j=1,\dots,n}$.

The asymptotic error distribution has three components. The first two lines on the right-hand side of (4.38) are second-order bias terms: the first line is due to the approximation of the Brownian increment by the discrete difference $\Delta_h W_t$, the second line to the approximation of the diffusion by the solution of the discretized SDE. The last line on the right-hand side of (4.38) is a martingale component associated with the estimation of the mean by an average over independent replications.

4.5 Asymptotic properties of MCFD estimators

Recall that $Z_{4,t,v} \equiv Y_v$ and, for any functional $H_{t,v}$ of Y, let $\nabla_{t,z_{j,4}} H_{t,v}$ denote the jth element of the tangent process associated with an infinitesimal perturbation of the state variable $Y_{j,t}$. To simplify the notation define the $d \times k$ matrix process $\gamma'_t \equiv \gamma(t, Y_t)' \equiv [(\sigma')^{-1}(\sigma^Y)'](t, Z_{4,t,t})$. With these definitions, the approximation error for MCFD estimators is given by $(e^{M,N,\tau,\alpha}_{t,T})' = [(e^{M,N,\tau,\alpha}_{1,t,T})', (e^{M,N,\tau,\alpha}_{2,t,T})']$ with

$$e^{M,N,\tau,\alpha}_{1,t,T} \equiv e^{MV,M,N}_{1,t,T} + e^{H,M,N,\tau,\alpha}_{1,t,T}, \tag{4.39}$$

$$e^{M,N,\tau,\alpha}_{2,t,T} \equiv e^{MV,M,N}_{2,t,T} + e^{H,M,N,\tau,\alpha}_{2,t,T}, \tag{4.40}$$

where $e^{MV,M,N}_{1,t,T}$ is given by (4.15), $e^{MV,M,N}_{2,t,T}$ by (4.17), and the hedging term approximation errors by

$$\begin{aligned} e^{H,M,N,\tau,\alpha}_{1,t,T} \equiv{}& \gamma'_t \big[\mathbf{E}^M_t \big[\nabla^{\tau,\alpha}_{t,z_{j,4}} f_1\big(Z^N_{t,T}; y^*_t\big)\big] \\ &- \mathbf{E}_t\big[\nabla_{t,z_{j,4}} f_1(Z_{t,T}; y^*_t)\big]\big]_{j=1,\dots,k}, \end{aligned} \tag{4.41}$$

$$\begin{aligned} e^{H,M,N,\tau,\alpha}_{2,t,T} \equiv{}& \gamma'_t \left[\mathbf{E}^M_t \left[\nabla^{\tau,\alpha}_{t,z_{j,4}} \int_t^T f_2\big(Z^N_{t,\eta^N_v}; y^*_t\big)\,\mathrm{d}v\right]\right. \\ &\left. - \mathbf{E}_t\left[\nabla_{t,z_{j,4}} \int_t^T f_2(Z_{t,v}; y^*_t)\,\mathrm{d}v\right]\right]_{j=1,\dots,k}. \end{aligned} \tag{4.42}$$

For MCFD estimators, the estimators of the mean–variance components are identical to those of MCMD: their convergence properties are as described in Proposition 4. The asymptotic error behavior of the hedging component is as follows. To simplify matters, we assume that $(\tau_j, \alpha_j) = (\tau, \alpha)$ for all $j = 1, \dots, k$.

Proposition 6. *Assume that the functions $f_1, f_2 \in \mathcal{C}^3(\mathbb{R}^{d_z})$ and suppose that $f_i(Z_{t,T}; y^*_t) \in \mathbb{D}^{1,2}$ for $i = 1, 2$. Let $K_{1,t,v}(Y_t; y^*_t)$ be defined for f_1 as in (4.7) and $\kappa_{2,t,v}(Y_t; y^*_t)$ for f_2 as in (4.9). Define the events*

$$F^j_1(N, \tau, r) = \left\{\left|N\nabla^{\tau,\alpha}_{t,z_{j,4}} f_1\big(Z^N_{t,T}; y^*_t\big) - \frac{1}{2}\partial_{y_j} K_{1,t,T}(Y_t; y^*_t)\right| > r\right\},$$

$$F_2^j(N,\tau,r) = \left\{ \left| N\nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T f_2\big(Z_{t,\eta_v^N}^N; y_t^*\big)\,\mathrm{d}v - \frac{1}{2}\int_t^T \partial_{y_j}\kappa_{2,t,\eta_v^N}(Y_t; y_t^*)\,\mathrm{d}v \right| > r \right\},$$

$$G_1^j(\tau,r) = \big\{ \big|\nabla_{t,z_{j,4}}^{\tau,\alpha} f_1(Z_{t,T}; y_t^*) - \mathbf{E}_t\big[\nabla_{t,z_{j,4}}^{\tau,\alpha} f_1(Z_{t,T}; y_t^*)\big]\big| > r \big\},$$

$$G_2^j(\tau,r) = \left\{ \left| \nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v - \mathbf{E}_t\left[\nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v\right] \right| > r \right\}.$$

Suppose that the conditions

$$\lim_{r\to\infty} \limsup_{1/\tau,N} \mathbf{E}_t\left[\mathbf{1}_{F_1^j(N,\tau,r)} \left| N\nabla_{t,z_{j,4}}^{\tau,\alpha} f_1\big(Z_{t,T}^N; y_t^*\big) - \frac{1}{2}\partial_{y_j} K_{1,t,T}(Y_t; y_t^*) \right| \right] = 0, \tag{4.43}$$

$$\lim_{r\to\infty} \limsup_{1/\tau,N} \mathbf{E}_t\left[\mathbf{1}_{F_2^j(N,\tau,r)} \left| N\nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T \left(f_2\big(Z_{t,\eta_v^N}^N; y_t^*\big) - \frac{1}{2}\partial_{y_j}\kappa_{2,t,\eta_v^N}(Y_t; y_t^*) \right) \mathrm{d}v \right| \right] = 0, \tag{4.44}$$

$$\lim_{r\to\infty} \limsup_{1/\tau} \mathbf{E}_t\big[\mathbf{1}_{G_1^j(\tau,r)} \big|\nabla_{t,z_{j,4}}^{\tau,\alpha} f_1(Z_{t,T}; y_t^*) - \mathbf{E}_t\big[\nabla_{t,z_{j,4}}^{\tau,\alpha} f_1(Z_{t,T}; y_t^*)\big]\big|^2 \big] = 0, \tag{4.45}$$

$$\lim_{r\to\infty} \limsup_{1/\tau} \mathbf{E}_t\left[\mathbf{1}_{G_2^j(\tau,r)} \left| \nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v - \mathbf{E}_t\left[\nabla_{t,z_{j,4}}^{\tau,\alpha} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v\right] \right|^2 \right] = 0 \tag{4.46}$$

hold, for all $j = 1, \ldots, k$. *Then, as* $M \to \infty$,

(i) *if* $\alpha = 1/2$ (*MCCFD*) *we have*

$$
\begin{aligned}
&M^{1/2} e_{t,T}^{H,M,N_M,\tau_M,\alpha} \\
&\quad \Rightarrow \frac{\varepsilon_1^{fcd}}{24} \begin{bmatrix} \gamma_t' \mathbf{E}_t[\nabla^3_{t,z_{j,4}} f_1(Z_{t,T}; y_t^*)]_{j=1,\dots,k} \\ \gamma_t' \mathbf{E}_t[\nabla^3_{t,z_{j,4}} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v]_{j=1,\dots,k} \end{bmatrix} \\
&\qquad + \frac{\varepsilon_2^{fcd}}{2} \begin{bmatrix} \gamma_t'[\partial_{y_j} K_{1,t,T}(Y_t; y_t^*)]_{j=1,\dots,k} \\ \gamma_t'[\partial_{y_j} \int_t^T \kappa_{2,t,\eta_v^N}(Y_t; y_t^*)\,\mathrm{d}v]_{j=1,\dots,k} \end{bmatrix} \\
&\qquad + (I_2 \otimes \gamma_t') P_{t,T}(Y_t; y_t^*),
\end{aligned} \tag{4.47}
$$

where $N_M, 1/\tau_M \to \infty$ *when* $M \to \infty$, *and where* $\varepsilon_1^{fcd} = \lim_{M\to\infty} M^{1/4}\tau_M$ *and* $\varepsilon_2^{fcd} = \lim_{M\to\infty} M^{1/2}/N_M$,

(ii) *if* $\alpha \neq 1/2$ (*MCBFD and MCFFD*)

$$
\begin{aligned}
&M^{1/2} e_{t,T}^{H,M,N_M,\tau_M,\alpha} \\
&\quad \Rightarrow \varepsilon_1^{fd} \delta(\alpha) \begin{bmatrix} \gamma_t' \mathbf{E}_t[\nabla^2_{t,z_{j,4}} f_1(Z_{t,T}; y_t^*)]_{j=1,\dots,k} \\ \gamma_t' \mathbf{E}_t[\nabla^2_{t,z_{j,4}} \int_t^T f_2(Z_{t,v}; y_t^*)\,\mathrm{d}v]_{j=1,\dots,k} \end{bmatrix} \\
&\qquad + \frac{\varepsilon_2^{fd}}{2} \begin{bmatrix} \gamma_t'[\partial_{y_j} K_{1,t,T}(Y_t; y_t^*)]_{j=1,\dots,k} \\ \gamma_t'[\partial_{y_j} \int_t^T \kappa_{2,t,\eta_v^N}(Y_t; y_t^*)\,\mathrm{d}v]_{j=1,\dots,k} \end{bmatrix} \\
&\qquad + (I_2 \otimes \gamma_t') P_{t,T}(Y_t; y_t^*),
\end{aligned} \tag{4.48}
$$

where $N_M, 1/\tau_M \to \infty$ *when* $M \to \infty$, *with* $\delta(\alpha) = (2\alpha - 1)/2$, *and where* $\varepsilon_1^{fd} = \lim_{M\to\infty} M^{1/2}\tau_M$ *and* $\varepsilon_2^{fd} = \lim_{M\to\infty} M^{1/2}/N_M$. *The random variable* $P_{t,T}(Y_t; y_t^*)$ *is the terminal point of a* $2d \times 1$ *dimensional Gaussian martingale with quadratic variation*

$$
[P, P]_{t,T}(Y_t; y) = \mathbf{E}_{t,Y_t}\left[\int_t^T L(v, Z_{t,v}; y) L(v, Z_{t,v}; y)'\,\mathrm{d}v \right],
$$

where

$$
L(v, Z_{t,v}; y) = \mathbf{E}_{v,Z_{t,v}}\left[\mathcal{D}_v \begin{bmatrix} (\nabla_{v,Z_{t,v}} f_1(Z_{t,T}; y))' \\ (\nabla_{v,Z_{t,v}} \int_t^T f_2(Z_{t,s}; y)\,\mathrm{d}s)' \end{bmatrix} \right].
$$

As for MCC estimators the asymptotic error distribution has three components. The first two are second-order bias terms due to the finite difference approximation of the tangent process and to the numerical approximation of SDEs characterizing the underlying diffusions. As for MCMD and MCC estimators there is a trade-off between these error components. Also, it is apparent that the convergence rate of MCFD estimators is better than that of MCC estimators. But, the nature of the differencing scheme (i.e. forward,

backward or central) only affects the second-order bias. The last component, $P_{t,T}$, is the terminal point of a Gaussian martingale. It describes the asymptotic distribution of the normalized difference between the empirical mean based on random variables drawn from the true distribution of the state variables and the true conditional expectation. This component is present even if simulation from the true distribution of the state variables is feasible.

For continuous functions the speeds of convergence of MCFD and MCMD estimators are identical. But even if sampling from the true distribution is possible, MCFD estimators will suffer from an additional second-order bias term due to the finite difference approximation of the tangent process. For discontinuous functions MCFD estimators converge more slowly than MCMD estimators (see Detemple et al., 2005d): MCCFD converges at the rate $M^{-2/5}$, whereas MCFFD and MCBFD converge at the same rate as MCC, $M^{-1/3}$.

4.6 Remarks and interpretations

The asymptotic MCMD error distribution depends on the number of Monte Carlo replications M used to approximate the conditional expectation and the number of discretization points N used to approximate the random variables in the expectation. As shown by Duffie and Glynn (1995) efficient Monte Carlo estimators of conditional expectations are obtained if the parameters M, N are chosen along the diagonal $\sqrt{M}/N =$ constant of the space of convergence parameters. Efficient Monte Carlo estimators, unfortunately, have noncentered error distributions, therefore suffer from a second-order bias. As discussed in Section 4.3, second-order biases cannot be ignored when the relative efficiencies of different Monte Carlo estimators are compared. For MCMD estimators Detemple et al. (2005c) provide analytic formulas for the second-order bias and second-order bias corrected estimators. They show that second-order bias corrected estimators are asymptotically equivalent to (generally) infeasible estimators that sample from the unknown true distribution of the state variables. Propositions 5 and 6 reveal that second-order biases are even more important for MCC and MCFD, as they both depend on an additional perturbation parameter. This dependence implies additional second-order bias components that appear difficult to correct for.

It should also be noted that the analysis above treats the shadow price of wealth y_t^* as a known quantity. This is clearly not the case when preferences are nonhomothetic. In this situation a Monte Carlo method (with discretized diffusion) can be combined with a numerical fixed point scheme to estimate the shadow price. The results of Proposition 3 show that the error associated with the estimation of y_t^* is of order $1/\sqrt{M}$ as long as $\lim_{M\to\infty} \sqrt{M}/N_M = \epsilon$ for some $\epsilon \in (0, \infty)$. As MCC estimators converge at the slower rate $1/M^{1/3}$ (see Proposition 5) the asymptotic error distribution is the same for known and estimated y_t^*. In contrast, because MCMD and MCFD estimators with known y_t^* converge at the faster rate $1/\sqrt{M}$ (see Propositions 4 and 6), the approximation error due to the estimation of y_t^* will not be asymptotically negligible.

An additional second-order bias term due to the approximation error of the shadow price of wealth will appear and affect the lengths of asymptotic confidence intervals. A detailed analysis of the error distribution is beyond the scope of this review article.

4.7 Asymptotic properties of MCR estimators

The portfolio estimator of Brandt et al. (2005) induces three error terms: the remainder of the Taylor approximation of the value function, the error due to the projections of conditional expectations on a polynomial basis and the Monte Carlo error introduced by the need to simulate random variables in order to perform these projections.

The convergence behavior of the MCR portfolio estimator has yet to be studied. In contrast, convergence results for Monte Carlo methods involving projections on basis functions are available for optimal stopping problems arising in the valuation of American contingent claims (see Tsitsiklis and van Roy, 2001; Clément et al., 2002; Egloff, 2005; and Glasserman and Yu, 2004). Although related, these convergence studies do not apply directly to the setting of Brandt et al. (2005): the control in the portfolio choice problem is not a binary variable and therefore has a more complex structure than the control of an optimal stopping problem. In addition, the papers of Tsitsiklis and van Roy (2001) and Clément et al. (2002) prove convergence but do not provide a convergence rate. Like the trade-off between the number of discretization points and the number of Monte Carlo replications described in Proposition 3, there is an optimal trade-off between the number of independent replications and the number of elements in the projection basis for the polynomial estimators of Brandt et al. (2005) and Longstaff and Schwartz (2001). Glasserman and Yu (2004) provide results for optimal stopping problems involving Brownian motion and geometric Brownian motion processes. In that context they show that the number of basis functions has to grow surprisingly fast to obtain convergence. For Brownian motion the number of polynomials $K = K_M$ for which accurate estimation is possible from M replications is $O(\log M)$. For geometric Brownian motion this growth rate is $O(\sqrt{\log M})$: the number of paths has to grow (faster than) exponentially with the number of polynomials.

All these results are derived in the context of American option pricing models. There are no reasons to expect better convergence results for the more complicated asset allocation problems. Egloff (2005) shows that results can be improved when bounded basis functions are used.[8] He also shows that the approximation error scales exponentially with the number of time steps. This

[8] A similar result is obtained by Gobet et al. (2005) for regression-based Monte Carlo methods used to solve backward stochastic differential equations. They provide a full convergence analysis in terms of $\mathbf{L}^2$ errors and a central limit theorem.

suggests that the MCR error will be large even for a moderate number of rebalancing times. This conjecture appears to be supported by the simulation evidence in Detemple et al. (2005b).

5 Performance evaluation: a numerical study

5.1 Experimental setting

In order to compare the different methods we use a model with an explicit solution. In this model the investor has constant relative risk aversion R (hence Corollary 1 applies) and operates in a market with a single risky stock and the riskless asset. There is a single Brownian motion W. The interest rate r is constant and the market price of risk θ follows the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}\theta_t = A(\bar{\theta} - \theta_t)\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t; \quad \theta_0 \text{ given}, \tag{5.1}$$

where A, $\bar{\theta}$, Σ are positive constants. The stock return has constant volatility σ. The investor cares about the expected utility of terminal wealth (there is no intermediate consumption).

The closed form solution for the optimal portfolio policy can be found in Wachter (2002).[9] Assume that the determinant condition

$$\Sigma^{-2}A^2 + \rho\big(1 + 2\Sigma^{-1}A\big) \geqslant 0, \tag{5.2}$$

holds, where $\rho = 1 - 1/R$, and define the constants

$$G = -\Sigma^{-1}A - \sqrt{\Sigma^{-2}A^2 + \rho\big(1 + 2\Sigma^{-1}A\big)}$$

and $\alpha = 2(A + \Sigma G)$. The optimal stock demand is $\pi_t^* = \pi_{1t}^* + \pi_{2t}^*$ where the mean–variance demand is $\pi_{1t}^* = (1/R)(\sigma_t)^{-1}\theta_t$ and the intertemporal hedging demand is

$$\pi_{2t}^* = -\frac{\rho}{R}\big[B(t,T) + C(t,T)\theta_t\big]\Sigma\sigma^{-1},$$

with

$$B(t,T) = \frac{2(1 - \exp(-\frac{1}{2}\alpha(T-t)))^2}{\alpha(\alpha + (\rho - G)\Sigma(1 - \exp(-\alpha(T-t))))}A\bar{\theta}, \tag{5.3}$$

$$C(t,T) = \frac{1 - \exp(-\alpha(T-t))}{\alpha + (\rho - G)\Sigma(1 - \exp(-\alpha(T-t)))}. \tag{5.4}$$

[9] Wachter shows that the problem reduces to a system of Riccatti ordinary differential equations. Liu (1998) and Schroder and Skiadas (1999) show that the same reduction applies when state variables follow affine processes.

5.2 Numerical results

This section reports comparison results for MCMD, MCC, MCFD and MCR. Three versions of MCFD, with forward finite differences (MCFFD), backward finite differences (MCBFD) and central finite differences (MCCFD) are tested. Three versions of MCR are also evaluated. The first one regresses on the excess returns for the last period (MCR-lin-1), the second on the excess returns for the last two periods (MCR-lin-2), the last one on the excess returns for the last four periods (MCR-lin-3).

In order to provide conclusive evidence about the efficiency of the different Monte Carlo methods, we draw 10,000 configurations of the parameters (R, T, θ_0, r) from independent uniform distributions. For each draw and each method, relative errors and execution times are recorded. A measure of accuracy, root mean square relative error (RMSRE), and a measure of speed, inverse average time (IAT), are computed from this sample, again for each method.[10] This experiment is repeated for different discretization values N and different numbers of trajectories M. The speed-accuracy trade-off can then be graphed to evaluate the relative performances of the candidate methods.

In order to use an Euler scheme that guarantees positive state price density we discretize $\log \xi$ and calculate the SPD ξ as the exponential of the discretized logarithmic SPD. Given the difficulties encountered in implementations of higher order polynomial-regression methods (see Detemple et al., 2005b) we only focus on the linear approximations MCR-lin-1, MCR-lin-2, MCR-lin-3.

The simulation experiment is designed in the following manner. The risk aversion parameter R is drawn from a uniform distribution with domain $[0.5, 5]$, the investment horizon T from a uniform distribution over the discrete set $\{1, 2, \ldots, 5\}$, the initial MPR θ from a uniform distribution over $[0.30, 1.50]$ and the constant interest rate r from a uniform distribution over $[0.01, 0.10]$. These distributions are assumed to be independent. Each draw consists of a vector $[R, T, \theta, r]$. Errors and computation times are recorded, for each method, for the pairs $(M, N) = \{(1000, 10), (4000, 20), (9000, 30), (16000, 40)\}$. These combinations of M, N are chosen so as to quadruple M when N is doubled, leaving the ratio $\sqrt{M}/N$ constant.[11] For MCC and MCFD an auxiliary parameter has to be selected. For MCC the time step h for the initial increment of the Brownian motion is set equal to the time step $1/N$, as in Cvitanic et al. (2003). Initial MPRs for MCFD methods are perturbed by setting $\tau = 0.1$. As is the case for M and N these auxiliary parameters decrease along the efficient path, in the manner

[10] IAT is measured by the number of portfolios computed per second.

[11] The ratio $\sqrt{M}/N$ is the efficiency ratio for MCMD. Increasing M and N while maintaining this ratio constant ensures convergence to the true value without modifying the structure of the second-order bias (see Detemple et al., 2005c).

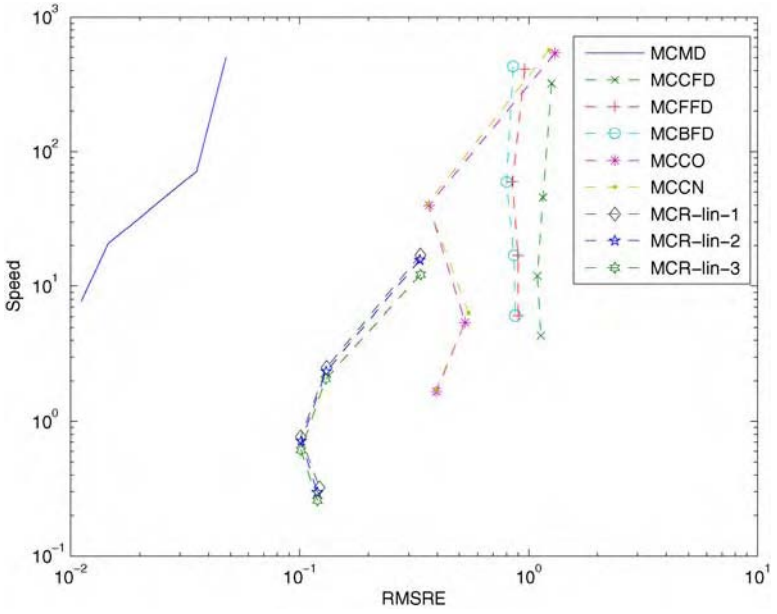


Fig. 1. This figure shows the speed-accuracy trade-off for MCMD, MCFD, MCC and MCR methods. MCCO corresponds to (3.15) and MCCN implements (3.16). All MCFD estimators are based on (3.22). Speed is measured by the inverse of the average computation time over the sample. Accuracy is measured by root mean square relative error. Four points, corresponding to the pairs $(M, N) = \{(1000, 10), (4000, 20), (9000, 30), (16000, 40)\}$, are graphed for each method. The auxiliary parameter for MCC is $h = 1/N$ and the initial auxiliary parameter for MCFD is $\tau = 0.1$. Both parameters decrease for efficient estimators as described in Propositions 5 and 6.

prescribed by the asymptotic convergence results in Propositions 5 and 6. For MCC the parameter h is cut in half if N doubles and M is multiplied by eight. For MCFD the parameter τ is cut in half if N doubles and M quadruples.

Sample statistics for RMSRE and IAT are based on 6415 "good" draws, i.e. draws for which all methods provide real results, out of the 10,000 replications. To provide perspective it is useful to note that all the "bad draws" are recorded when one of the three MCR methods fails to produce a result. Eliminating bad draws therefore advantages MCR.

Figure 1 displays the results from this experiment. The first observation is that MCMD dominates MCR, MCC and MCFD. At the same time MCR weakly dominates MCC, whereas MCC fares better than MCFD. MCMD improves on MCR by a factor in excess of 10. For a speed in the neighborhood of 10 the RMSRE of MCMD nears 10^{-2} while that of MCR is about 3×10^{-1}. Given the slopes of these trade-offs along MCMD and MCR this gap is expected to widen if M and N are further increased.

Next, we compare different versions of Monte Carlo estimators within each class.

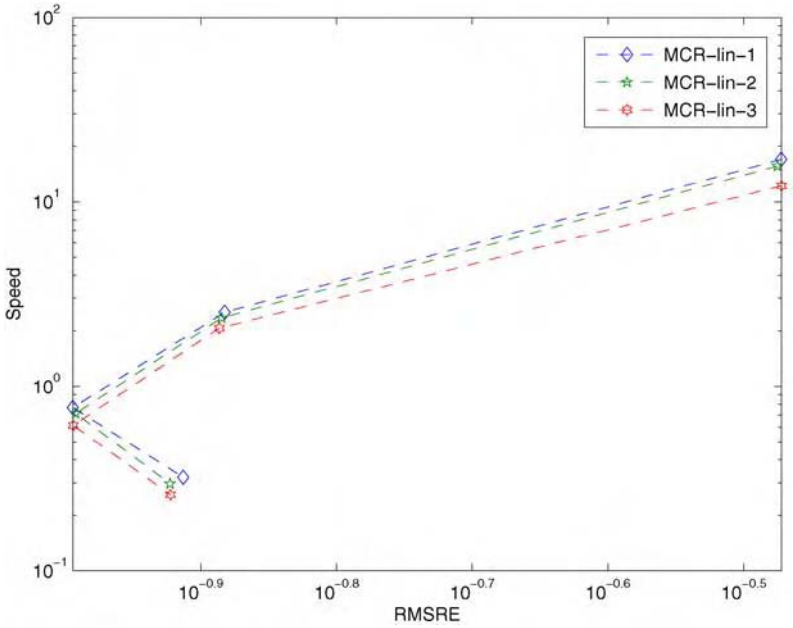


Fig. 2. This figure shows the speed-accuracy trade-off for three MCR-lin methods. Speed is measured by the inverse of the average computation time over the sample. Accuracy is measured by root mean square relative error. Four points, corresponding to the pairs $(M, N) = \{(1000, 10), (4000, 20), (9000, 30), (16000, 40)\}$, are graphed for each method.

Figure 2 illustrates that the three regression methods have a very similar performance: regressing on additional lagged returns does not improve performance. As a matter of fact it turns out that adding lagged regressors may cause the fixed point algorithm proposed by Brandt et al. (2005) to fail more frequently. Among the 10,000 configurations of $[R, T, \theta, r]$, MCR-lin-1 produced 7229 and MCR-lin-2 6984 good draws. But, in accordance with the results for American option pricing in Longstaff and Schwartz (2001), when MCR-lin provides results, the performance does not seem to depend on the choice of the orthonormal basis. This, however, is not a general property. In the present example this finding may simply be due to the fact that the true policy is linear in the MPR. MCR-lin-1 is therefore closest to the functional form of the true portfolio weight.

A similar comparison for MCC methods in Figure 3 reveals that the performance of both MCC methods is similar. Close inspection indicates that the MCC method based on (3.16) performs slightly better than the original method proposed by Cvitanic et al. (2003), based on (3.15). The RMSRE of the modified MCC method may be smaller because it estimates the hedging demand directly. In contrast, the original MCC method calculates the total portfolio weight but does not exploit the fact that for CRRA preferences the mean–variance component is known in closed form. The small size of the hedging

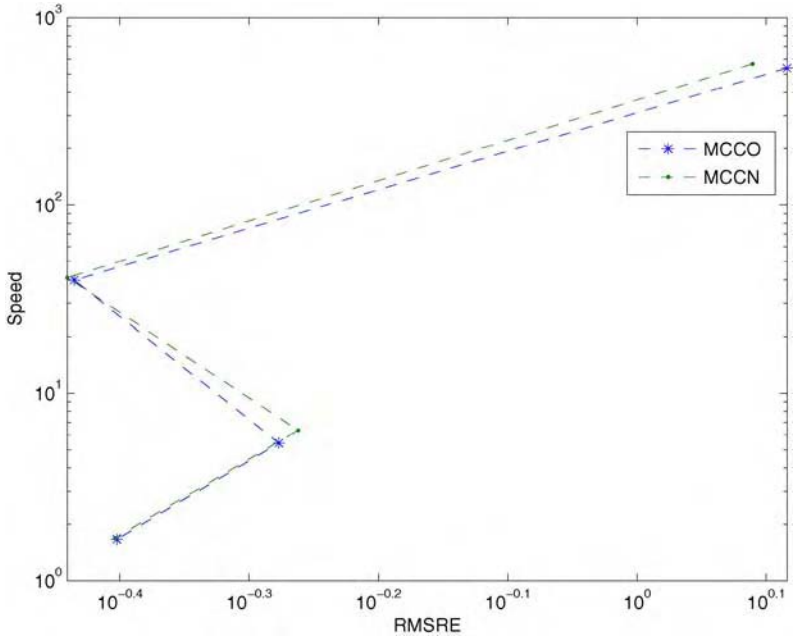


Fig. 3. This figure shows the speed-accuracy trade-off for two MCC methods. MCCO corresponds to (3.15) and MCCN implements (3.16). Speed is measured by the inverse of the average computation time over the sample. Accuracy is measured by root mean square relative error. Four points, corresponding to the triplets $(M, N, h) = \{(1000, 10, 1/10), (4000, 20, 1/20), (9000, 30, 1/30), (16000, 40, 1/40)\}$, are graphed for each method. The auxiliary time step for the initial Brownian increment h is chosen equal to the discretization step $1/N$.

demand for horizons between one and five years may be the source of the smaller relative error produced by the modified MCC method.

Finally we compare three different MCFD methods. The results in Figure 4 show that MCBFD estimators outperform both MCFFD and MCCFD estimators. MCCFD estimators are least efficient. This may be due to the fact that these estimators require the simulation of two additional perturbed processes, whereas MCFFD and MCBFD are based on a one-sided perturbation of the MPR diffusion. Hence, the computational effort to calculate MCCFD estimators is greater. At the same time Proposition 6 establishes that the speed of convergence for all methods is the same. The three MCFD estimators only differ in the second-order bias for which a ranking based on the theoretical results does not appear possible. The simulation in Figure 4 suggests that the second-order bias is larger for MCCFD than MCBFD and MCFFD.

6 Conclusion

Monte Carlo simulation is the approach of choice for high dimensional problems with large numbers of underlying variables. In contrast to alterna-

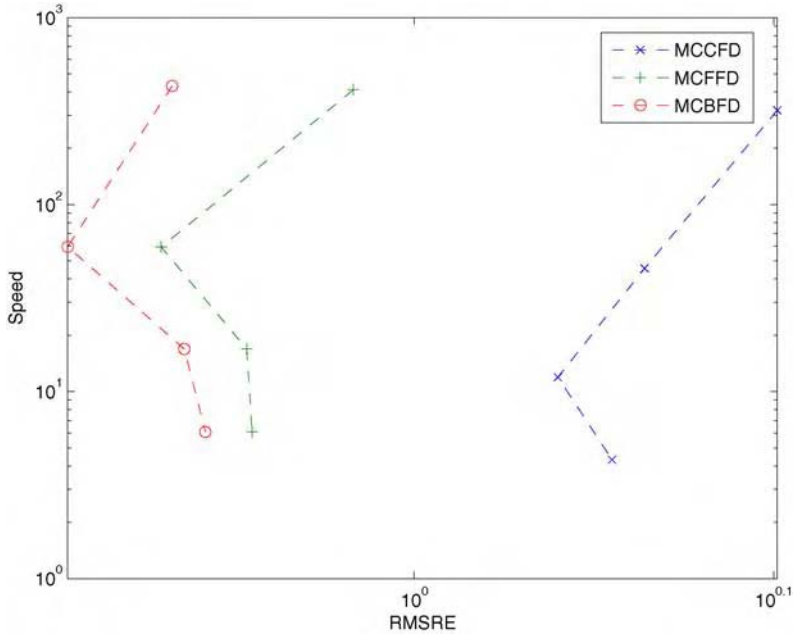


Fig. 4. This figure shows the speed-accuracy trade-off for three MCFD methods based on (3.22). Speed is measured by the inverse of the average computation time over the sample. Accuracy is measured by root mean square relative error. Four points, corresponding to the triples $(M, N, \tau) = \{(1000, 10, 1/10), (4000, 20, 1/20), (9000, 30, 1/30), (16000, 40, 1/40)\}$, are graphed for each method.

tives such as lattice methods (finite difference and finite element schemes, Markov chain approximations, quantization and quadrature schemes, etc.), simulation methods do not suffer from the well-known curse of dimensionality. As a result they emerge as natural candidates for the numerical implementation of optimal portfolio rules in high dimensional portfolio choice models. MCMD, MCC, MCFD and MCR, are various simulation schemes that have been proposed and studied during the past few years with this particular application in mind. Among these candidates, MCMD has shown a number of attractive features. One important consideration is that it is the only simulation method that attains the optimal convergence rate implied by the central limit theorem. In numerical experiments conducted it also showed superior efficiency, as measured by the trade-off between speed of computation and accuracy.

Asset allocation models with complete markets and diffusion state variables are natural candidates for the application of MCMD. In these settings the optimal portfolios can be expressed as conditional expectations of functionals of the state variables and their Malliavin derivatives, and these can be calculated numerically using Monte Carlo simulation. Settings with incomplete markets and more general forms of portfolio constraints prove more challeng-

ing. MCMD extends to these models as well, when the dual problem has an explicit solution. Constrained problems, embedding affine models, for which this can be achieved are described in Detemple and Rindisbacher (2005). Extensions of the method to more general settings, where an explicit solution to the dual is not available, remain to be carried out.

Acknowledgement

We thank MITACS for financial support.

Appendix A. An introduction to Malliavin calculus

The Malliavin calculus is a calculus of variations for stochastic processes. It applies to *Wiener* (*or Brownian*) *functionals*, i.e. random variables and stochastic processes that depend on the trajectories of Brownian motions. The Malliavin derivative, which is one element of this calculus of variations, measures the effect of a small variation in the trajectory of an underlying Brownian motion on the value of a Wiener functional.

A.1 Smooth Brownian functionals

To set the stage consider a Wiener space generated by the d-dimensional Brownian motion process $W = (W_1, \ldots, W_d)'$. As is well known we can associate each state of nature with a trajectory of the Brownian motion (the set of states of nature is the space of trajectories). Let $(t_1, \ldots, t_n)$ be a partition of the time interval $[0, T]$ and let $F(W)$ be a random variable of the form

$$F(W) \equiv f(W_{t_1}, \ldots, W_{t_n}),$$

where f is a continuously differentiable function. The random variable $F(W)$ depends (smoothly) on the d-dimensional Brownian motion W at a finite number of points along its trajectory; it is called a *smooth Brownian functional*.

A.2 The Malliavin derivative of a smooth Brownian functional

The Malliavin derivative of F is the change in F due to a change in the path of W. To simplify matters assume first that $d = 1$, i.e. there is a unique Brownian motion. Consider shifting the trajectory of W by ε starting at time t. Suppose $t_k \leqslant t < t_{k+1}$ for some $k = 1, \ldots, n-1$. The Malliavin derivative of F at t is defined by

$$\mathcal{D}_t F(W) \equiv \left.\frac{\partial f(W_{t_1} + \varepsilon \mathbf{1}_{[t,\infty[}(t_1), \ldots, W_{t_n} + \varepsilon \mathbf{1}_{[t,\infty[}(t_n))}{\partial \varepsilon}\right|_{\varepsilon=0}$$

$$= \lim_{\varepsilon \to 0} \frac{F(W + \varepsilon \mathbf{1}_{[t,\infty[}) - F(W)}{\varepsilon}, \tag{A.1}$$

where $\mathbf{1}_{[t,\infty[}$ is the indicator process of the set $[t, \infty)$ (that is $\mathbf{1}_{[t,\infty[}(s) = 1$ for $s \in [t, \infty)$; $= 0$ otherwise). In more compact form we can write

$$\mathcal{D}_t F(\omega) = \sum_{j=k}^{n} \partial_j f(W_{t_1}, \ldots, W_{t_k}, \ldots, W_{t_n}) \mathbf{1}_{[t,\infty[}(t_j), \tag{A.2}$$

where $\partial_j f$ is the derivative with respect to the jth argument of f.

A simple example will illustrate the notion. Consider the price of the stock in the Black–Scholes model. Its value at date T is given by

$$S_T = S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma W_T\right),$$

where W_T is the terminal value of the univariate Brownian motion process defining the uncertainty in this model. Since $S_T = f(W_T)$ with $f(x) = S_0 \exp((\mu - \frac{1}{2}\sigma^2)T + \sigma x)$ it is clear that S_T is a smooth Brownian functional. A direct application of the definition gives

$$\mathcal{D}_t S_T = \partial f(W_T) \mathbf{1}_{[t,\infty[}(T) = \sigma S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma W_T\right) = \sigma S_T.$$

In this example the stock price depends only on the Brownian motion at time T. The Malliavin derivative is then the derivative with respect to W_T. This reflects the fact that a perturbation of the path of the Brownian motion from t onward, affects S_T only through the terminal value W_T.

Suppose next that $d > 1$, i.e. the underlying Brownian motion is multi-dimensional. The Malliavin derivative of F at t is now a $1 \times d$-dimensional vector denoted by $\mathcal{D}_t F = (\mathcal{D}_{1t} F, \ldots, \mathcal{D}_{dt} F)$. The ith coordinate of this vector, $\mathcal{D}_{it} F$, captures the impact of a perturbation in W_i by ε starting at some time t. If $t_k \leqslant t < t_{k+1}$ we have

$$\mathcal{D}_{it} F = \sum_{j=k}^{n} \frac{\partial f}{\partial x_{ij}}(W_{t_1}, \ldots, W_{t_k}, \ldots, W_{t_n}) \mathbf{1}_{[t,\infty[}(t_j), \tag{A.3}$$

where $\partial f / \partial x_{ij}$ is the derivative with respect to the ith component of the jth argument of f (i.e. the derivative with respect to W_{it_j}).

A.3 The domain of the Malliavin derivative operator

The definition above can be extended to random variables that depend on the path of the Brownian motion over a continuous interval $[0, T]$. This extension uses the fact that a path-dependent functional can be approximated by a suitable sequence of smooth Brownian functionals. The Malliavin derivative of the path-dependent functional is then given by the limit of the Malliavin

derivatives of the smooth Brownian functionals in the approximating sequence. The space of random variables for which Malliavin derivatives are defined is called $\mathbb{D}^{1,2}$. This space is the completion of the set of smooth Brownian functional with respect to the norm $\|F\|_{1,2} = (E[F^2] + \mathbf{E}[\int_0^T \|\mathcal{D}_t F\|^2 \,\mathrm{d}t])^{\frac{1}{2}}$ where $\|\mathcal{D}_t F\|^2 = \sum_i (\mathcal{D}_{it} F)^2$.

A.4 Malliavin derivatives of Riemann, Wiener and Itô integrals

This extension enables us to handle stochastic integrals, which depend on the path of the Brownian motion over a continuous interval, in a very natural manner. Consider, for instance, the stochastic Wiener integral $F(W) = \int_0^T h(t)\,\mathrm{d}W_t$, where $h(t)$ is a function of time and W is one-dimensional. Integration by parts shows that $F(W) = h(T)W_T - \int_0^T W_s\,\mathrm{d}h(s)$. Straightforward calculations give

$$\begin{aligned} F(W + \varepsilon \mathbf{1}_{[t,\infty[}) - F(W) &= h(T)\big(W_T + \varepsilon \mathbf{1}_{[t,\infty[}(T)\big) \\ &\quad - \int_0^T \big(W_s + \varepsilon \mathbf{1}_{[t,\infty[}(s)\big)\,\mathrm{d}h(s) \\ &\quad - \left(h(T)W_T - \int_0^T W_s\,\mathrm{d}h(s)\right) \\ &= h(T)\varepsilon - \int_0^T \varepsilon \mathbf{1}_{[t,\infty[}(s)\,\mathrm{d}h(s) = \varepsilon h(t). \end{aligned}$$

It then follows, from the definition (A.1), that $\mathcal{D}_t F = h(t)$. The Malliavin derivative of F at t is the volatility $h(t)$ of the stochastic integral at t: this volatility measures the sensitivity of the random variable F to the Brownian innovation at t.

Next, let us consider a random Riemann integral with integrand that depends on the path of the Brownian motion. This Brownian functional takes the form $F(W) \equiv \int_0^T h_s\,\mathrm{d}s$ where h_s is a progressively measurable process (i.e. a process that depends on time and the past trajectory of the Brownian motion) such that the integral exists (i.e. $\int_0^T |h_s|\,\mathrm{d}s < \infty$ with probability one). We now have

$$F\big(W + \varepsilon \mathbf{1}[t,\infty[\big) - F(W) = \int_0^T \big(h_s\big(W + \varepsilon \mathbf{1}[t,\infty[\big) - h_s(W)\big)\,\mathrm{d}s.$$

Since $\lim_{\varepsilon\to 0}(h_s(W + \varepsilon \mathbf{1}_{[t,\infty[} - h_s(W))/\varepsilon = \mathcal{D}_t h_s(W)$ it follows that $\mathcal{D}_t F = \int_t^T \mathcal{D}_t h_s\,\mathrm{d}s$.

Finally, consider the Itô integral $F(\omega) = \int_0^T h_s(W)\,\mathrm{d}W_s$. To simplify the notation write $h^\varepsilon \equiv h(W + \varepsilon \mathbf{1}_{[t,\infty[})$ and $W^\varepsilon \equiv W + \varepsilon \mathbf{1}_{[t,\infty[}$. Integration by parts then gives

$$\begin{aligned}
F^\varepsilon - F &= \int_0^T \big(h_s^\varepsilon - h_s\big)\,\mathrm{d}W_s + \int_0^T h_s^\varepsilon\,\mathrm{d}\big(W_s^\varepsilon - W_s\big) \\
&= \int_t^T \big(h_s^\varepsilon - h_s\big)\,\mathrm{d}W_s + h_T^\varepsilon\big(W_T^\varepsilon - W_T\big) - \int_0^T \big(W_s^\varepsilon - W_s\big)\,\mathrm{d}h_s^\varepsilon \\
&\quad - \int_0^T \mathrm{d}\big[W^\varepsilon - W, h^\varepsilon\big]_s \\
&= \int_t^T \big(h_s^\varepsilon - h_s\big)\,\mathrm{d}W_s + h_T^\varepsilon \varepsilon - \varepsilon \int_t^T \mathrm{d}h_s^\varepsilon \\
&= \int_t^T \big(h_s^\varepsilon - h_s\big)\,\mathrm{d}W_s + \varepsilon h_t^\varepsilon.
\end{aligned}$$

The second equality above uses $h_s^\varepsilon = h_s$ for $s < t$ to simplify the first integral and the integration by parts formula to expand the second integral. The third equality is based on the fact that the cross-variation is null (i.e. $[W^\varepsilon - W, h^\varepsilon]_T = 0$) because $W_s^\varepsilon - W_s = \varepsilon \mathbf{1}_{[t,\infty[}(s)$ and $\mathbf{1}_{[t,\infty[}(s)$ is of bounded total variation.[12] The last equality uses, again, the integration by parts formula to simplify the last two terms. As $\lim_{\varepsilon\to 0}(h_s^\varepsilon - h_s)/\varepsilon = \mathcal{D}_t h_s$ we obtain $\mathcal{D}_t F = h_t + \int_t^T \mathcal{D}_t h_s\,\mathrm{d}W_s$.

Malliavin derivatives of Wiener, Riemann and Itô integrals depending on multidimensional Brownian motions can be defined in a similar manner. As in Section A.2 the Malliavin derivative is a d-dimensional process which can be defined component-by-component, by the operations described above.

A.5 Martingale representation and the Clark–Ocone formula

In Wiener spaces martingales with finite variances can be written as sums of Brownian increments.[13] That is, $M_t = M_0 + \int_0^t \phi_s\,\mathrm{d}W_s$ for some progressively

[12] The *total variation* of a function f is $\lim_{N\to\infty} \sum_{t_n \in \Pi^N([0,t])} |f(t_{n+1}) - f(t_n)|$ where $\Pi^N([0,t])$ is a partition with N points of the interval $[0,t]$.

[13] A *Wiener space* is the canonical probability space $(\mathcal{C}_0(\mathbb{R}_+;\mathbb{R}^d), \mathcal{B}(\mathcal{C}_0(\mathbb{R}_+;\mathbb{R}^d)), \mathbf{P})$ of nowhere differentiable functions $\mathcal{C}_0$, endowed with its Borel sigma field and the Wiener measure. The *Wiener*

measurable process ϕ, which represents the volatility coefficient of the martingale. This result is known as the martingale representation theorem. One of the most important benefits of Malliavin calculus is to identify the integrand ϕ in this representation. This is the content of the Clark–Ocone formula.

The Clark–Ocone formula states that any random variable $F \in \mathbb{D}^{1,2}$ can be decomposed as

$$F = \mathbf{E}[F] + \int_0^T \mathbf{E}_t[\mathcal{D}_t F]\,\mathrm{d}W_t, \tag{A.4}$$

where $\mathbf{E}_t[\cdot]$ is the conditional expectation at t given the information generated by the Brownian motion W. For a martingale closed by $F \in \mathbb{D}^{1,2}$ (i.e. $M_t = \mathbf{E}_t[F]$) conditional expectations can be applied to (A.4) to obtain $M_t = \mathbf{E}[F] + \int_0^t \mathbf{E}_s[\mathcal{D}_s F]\,\mathrm{d}W_s$.

An intuitive derivation of this formula can be provided along the following lines. Assume that $F \in \mathbb{D}^{1,2}$. From the martingale representation theorem we have $F = \mathbf{E}[F] + \int_0^T \phi_s\,\mathrm{d}W_s$. Taking the Malliavin derivative on each side, and applying the rules of Malliavin calculus described above, gives $\mathcal{D}_t F = \phi_t + \int_t^T \mathcal{D}_t \phi_s\,\mathrm{d}W_s$. Taking conditional expectations on each side now produces $\mathbf{E}_t[\mathcal{D}_t F] = \phi_t$ (given that $\mathbf{E}_t[\int_t^T \mathcal{D}_t \phi_s\,\mathrm{d}W_s] = 0$ and ϕ_t is known at t). Substituting this expression in the representation of F leads to (A.4).

The results above also show that the Malliavin derivative and the conditional expectation operator commute. Indeed, let $v \geqslant t$ and consider the martingale M closed by $F \in \mathbb{D}^{1,2}$. From the representations for M and F above we obtain $\mathcal{D}_t M_v = \int_t^v \mathcal{D}_t \mathbf{E}_s[\mathcal{D}_s F]\,\mathrm{d}W_s + \mathbf{E}_t[\mathcal{D}_t F]$ and $\mathcal{D}_t F = \int_t^T \mathcal{D}_t \mathbf{E}_s[\mathcal{D}_s F]\,\mathrm{d}W_s + \mathbf{E}_t[\mathcal{D}_t F]$. Taking the conditional expectation at time v of the second expression gives $\mathbf{E}_v[\mathcal{D}_t F] = \int_t^v \mathcal{D}_t \mathbf{E}_s[\mathcal{D}_s F]\,\mathrm{d}W_s + \mathbf{E}_t[\mathcal{D}_t F]$. As the formulas on the right-hand sides of these two equalities are the same we conclude that $\mathcal{D}_t M_v = \mathbf{E}_v[\mathcal{D}_t F]$. Using the definition of M_v we can also write $\mathcal{D}_t \mathbf{E}_v[F] = \mathbf{E}_v[\mathcal{D}_t F]$: the Malliavin derivative operator and the conditional expectation operator commute.

A.6 The chain rule of Malliavin calculus

In applications one often needs to compute the Malliavin derivative of a function of a path-dependent random variable. As in ordinary calculus, a chain rule also applies in the Malliavin calculus. Let $F = (F_1, \ldots, F_n)$ be a vector of random variables in $\mathbb{D}^{1,2}$ and suppose that ϕ is a differentiable function of F

measure is the measure such that the d-dimensional coordinate mapping process is a Brownian motion.

with bounded derivatives. The Malliavin derivative of $\phi(F)$ is then,

$$\mathcal{D}_t\phi(F) = \sum_{i=1}^{n} \frac{\partial \phi}{\partial x_i}(F)\mathcal{D}_t F_i$$

where $\frac{\partial \phi}{\partial x_i}(F)$ represents the derivative relative to the ith argument of ϕ.

A.7 Malliavin derivatives of stochastic differential equations

For applications to portfolio allocation it is essential to be able to calculate the Malliavin derivative of the solution of a stochastic differential equation (SDE) (i.e. the Malliavin derivative of a diffusion process). The rules of Malliavin calculus presented above can be used to that effect.

Suppose that a state variable Y_t follows the diffusion process $\mathrm{d}Y_t = \mu^Y(Y_t)\,\mathrm{d}t + \sigma^Y(Y_t)\,\mathrm{d}W_t$ where Y_0 is given and $\sigma^Y(Y_t)$ is a scalar (W is single dimensional). Equivalently, we can write the process Y in integral form as

$$Y_t = Y_0 + \int_0^t \mu^Y(Y_s)\,\mathrm{d}s + \int_0^t \sigma^Y(Y_s)\,\mathrm{d}W_s.$$

Using the results presented above, it is easy to verify that the Malliavin derivative $\mathcal{D}_t Y_s$ satisfies

$$\mathcal{D}_t Y_s = D_t Y_0 + \int_t^s \partial\mu^Y(Y_v)\mathcal{D}_t Y_v\,\mathrm{d}v + \int_t^s \partial\sigma^Y(Y_v)\mathcal{D}_t Y_v\,\mathrm{d}W_v + \sigma(Y_t).$$

As $\mathcal{D}_t Y_0 = 0$, the Malliavin derivative obeys the following linear SDE

$$\mathrm{d}(\mathcal{D}_t Y_s) = \big[\partial\mu^Y(Y_s)\,\mathrm{d}s + \partial\sigma^Y(Y_s)\,\mathrm{d}W_s\big](\mathcal{D}_t Y_s) \tag{A.5}$$

subject to the initial condition $\lim_{s\to t}\mathcal{D}_t Y_s = \sigma^Y(Y_t)$.

If $\sigma^Y(Y_t)$ is a $1\times d$ vector (W is a d-dimensional Brownian motion) the same arguments apply to yield (A.5) subject to the initial condition $\lim_{s\to t}\mathcal{D}_t Y_s = \sigma(Y_t)$. In this multi-dimensional setting $\partial\sigma^Y(Y_s) \equiv (\partial\sigma_1^Y(Y_s), \ldots, \partial\sigma_d^Y(Y_s))$ is the row vector composed of the derivatives of the components of $\sigma^Y(Y_s)$. The Malliavin derivative $\mathcal{D}_t Y_s$ is the $1 \times d$ row vector $\mathcal{D}_t Y_s = (\mathcal{D}_{1t} Y_s, \ldots, \mathcal{D}_{dt} Y_s)$.

A.8 Stochastic flows and tangent processes

For implementation purposes it is useful to relate the Malliavin derivative to the notion of stochastic flow of a stochastic differential equation and the

associated concept of a tangent process. These notions have been explored by various authors including Kunita (1986) and Malliavin (1997).

A stochastic flow of homeomorphisms (or stochastic flow for short) is an $\mathbb{R}^d$-valued random field $\{\psi_{t,v}(y,\omega)\colon 0 \leqslant t \leqslant v \leqslant T,\ y \in \mathbb{R}^d\}$ such that for almost all ω

(a) $\psi_{t,v}(y)$ is continuous in t, v, y,
(b) $\psi_{v,u}(\psi_{t,v}(y)) = \psi_{t,u}(y)$ for all $t \leqslant v \leqslant u$ and $y \in \mathbb{R}^d$,
(c) $\psi_{t,t}(y) = y$ for any $t \leqslant T$,
(d) the map: $\psi_{t,v}\colon \mathbb{R}^d \mapsto \mathbb{R}^d$ is a homeomorphism for any t, v.[14]

An important class of stochastic flows is given by the solutions of SDEs of the form

$$\mathrm{d}Y_v = \mu^Y(Y_v)\,\mathrm{d}v + \sigma^Y(Y_v)\,\mathrm{d}W_v, \quad v \in [t,T]; \quad Y_t = y.$$

The stochastic flow $\psi_{t,v}(y,\omega)$ is the position of the diffusion Y at time v, in state ω, given an initial position $Y_t = y$ at time t. A subclass of stochastic flows of homeomorphisms is obtained if $\psi_{t,v}\colon \mathbb{R}^d \mapsto \mathbb{R}^d$ is also required to be a diffeomorphism.[15] A element of this subclass is called a stochastic flow of diffeomorphism. For a stochastic flow of diffeomorphism determined by the solutions of an SDE, the derivative $\nabla_{t,y}\psi_{t,\cdot}(y)$ with respect to the initial condition satisfies

$$d\big(\nabla_{t,y}\psi_{t,v}(y)\big) = \left(\partial\mu^Y(Y_v)\,\mathrm{d}v + \sum_{j=1}^{d}\partial\sigma_j^Y(Y_v)\,\mathrm{d}W_v^j\right)\nabla_{t,y}\psi_{t,v}, \quad v \in [t,T], \tag{A.6}$$

subject to the initial condition $\nabla_{t,y}\psi_{t,t}(y) = I_d$. The process $\nabla_{t,y}\psi_{t,\cdot}(y)$ is called the first variation process or the tangent process.

A comparison of (A.5) and (A.6) shows that

$$\mathcal{D}_t Y_t = \mathcal{D}_t\psi_{t,v}(y) = \nabla_{t,y}\psi_{t,v}(y)\sigma^Y(y).$$

The Malliavin derivative is therefore a linear transformation of the tangent process.

Appendix B. Proofs

Proof of Proposition 1. Recall that the deflated optimal wealth process is given by $\xi_t X_t^* = \mathbf{E}_t[\int_t^T \xi_v I(y^*\xi_v, v)^+\,\mathrm{d}v + \xi_T J(y^*\xi_T, T)^+]$. Applying Itô's lemma

[14] A function is a homeomorphism if it is bijective, continuous and its inverse is also continuous.
[15] A diffeomorphism is a map between manifolds that is differentiable and has a differentiable inverse.

and the Clark–Ocone formula to this expression shows that

$$\begin{aligned}
&\xi_t X_t^* \pi_t^{*\prime} \sigma_t - \xi_t X_t^* \theta_t' \\
&\quad = -\mathbf{E}_t\left[\int_t^T \xi_v Z_1(y^*\xi_v, v)\,\mathrm{d}v + \xi_T Z_2(y^*\xi_T, T)\right]\theta_t' \\
&\qquad - \mathbf{E}_t\left[\int_t^T \xi_v Z_1(y^*\xi_v, v) H_{t,v}'\,\mathrm{d}v + \xi_T Z_2(y^*\xi_T, T) H_{t,T}'\right]
\end{aligned}$$

where

$$\begin{aligned}
Z_1(y^*\xi_v, v) &= I(y^*\xi_v, v)^+ + y^*\xi_v I'(y^*\xi_v, v) 1_{\{I(y^*\xi_v, v)\geqslant 0\}}, \\
Z_2(y^*\xi_T, T) &= J(y^*\xi_T, T)^+ + y^*\xi_T J'(y^*\xi_T, T) 1_{\{J(y^*\xi_T, T)\geqslant 0\}}, \\
H_{t,v}' &= \int_t^v (\mathcal{D}_t r_s + \theta_s' \mathcal{D}_t \theta_s)\,\mathrm{d}s + \int_t^v \mathrm{d}W_s' \cdot \mathcal{D}_t \theta_s
\end{aligned}$$

and $\mathcal{D}_t r_s$, $\mathcal{D}_t \theta_s$ are the Malliavin derivatives of the interest rate and the market price of risk. The chain rule of Malliavin calculus (Section A.6), along with the results for Malliavin derivatives of SDEs (Section A.7) now lead to (2.16) and (2.17).

From the definition of optimal wealth X^*, it also follows that

$$X_t^* - \mathbf{E}_t\left[\int_t^T \xi_{t,v} Z_1(y^*\xi_v, v)\,\mathrm{d}v + \xi_{t,T} Z_2(y^*\xi_T, T)\right] = -\mathbf{E}_t[D_{t,T}]$$

where

$$\begin{aligned}
D_{t,T} = &\int_t^T \xi_{t,v}(y^*\xi_v) I'(y^*\xi_v, v) 1_{\{I(y^*\xi_v, v)\geqslant 0\}}\,\mathrm{d}v \\
&+ \xi_{t,T}(y^*\xi_T) J'(y^*\xi_T, T) 1_{\{J(y^*\xi_T, T)\geqslant 0\}}
\end{aligned}$$

so that,

$$\begin{aligned}
X_t^* \pi_t^{*\prime} \sigma_t = &-\mathbf{E}_t[D_{t,T}]\theta_t' \\
&- \mathbf{E}_t\left[\int_t^T \xi_{t,v} Z_1(y^*\xi_v, v) H_{t,v}'\,\mathrm{d}v + \xi_{t,T} Z_2(y^*\xi_T, T) H_{t,T}'\right].
\end{aligned}$$

Transposing this formula and identifying the first term with π_1^* and the second with π_2^* leads to the formulae in the proposition. □

Proof of Corollary 1. For constant relative risk aversion $u(c,t) = \eta_t c^{1-R}/(1-R)$ and $U(X,T) = \eta_T X^{1-R}/(1-R)$, with $\eta_t \equiv \exp(-\beta t)$, we obtain the functions,

$$
\begin{aligned}
&I(y\xi_v, v) = (y\xi_v/\eta_v)^{-1/R}, \qquad J(y\xi_T, T) = (y\xi_T/\eta_T)^{-1/R},\\
&y\xi_v I'(y\xi_v, v) = -(1/R)(y\xi_v/\eta_v)^{-1/R} = -(1/R)I(y\xi_v, v),\\
&y\xi_T J'(y\xi_T, T) = -(1/R)(y\xi_T/\eta_T)^{-1/R} = -(1/R)J(y\xi_T, T),\\
&Z_1(y\xi_v, v) = (1-1/R)I(y\xi_v, v),\\
&Z_2(y\xi_T, T) = (1-1/R)J(y\xi_T, T).
\end{aligned}
$$

The formulas in the corollary follow by substituting these expressions in the policies of Proposition 1. □

Proof of Proposition 2. Note that the optimal consumption policy (2.18) satisfies the budget constraint

$$
\mathbf{E}_t\left[\int_t^T \xi_{t,v} I(y^*\xi_t\xi_{t,v}, v)^+ \,\mathrm{d}v + \xi_{t,T} J(y^*\xi_t\xi_{t,T}, T)^+\right] = X_t^*. \tag{B.1}
$$

Given the regularity conditions on preferences, the function $\mathcal{I}(t, y, Y_t)$ defined for $y > 0$ as

$$
\mathcal{I}(t, y, Y_t) \equiv \mathbf{E}_t\left[\int_t^T \xi_{t,v} I(y\xi_{t,v}, v)^+ \,\mathrm{d}v + \xi_{t,T} J(y\xi_{t,T}, T)^+\right] \tag{B.2}
$$

has an inverse $y^*(t, X_t, Y_t)$ that is unique and satisfies

$$
\begin{aligned}
&\mathbf{E}_t\left[\int_t^T \xi_{t,v} I\big(y^*(t, X_t^*, Y_t)\xi_{t,v}, v\big)^+ \,\mathrm{d}v + \xi_{t,T} J\big(y^*(t, X_t^*, Y_t)\xi_{t,T}, T\big)^+\right]\\
&\quad = X_t^*.
\end{aligned} \tag{B.3}
$$

We conclude that $y^*\xi_t = y^*(t, X_t^*, Y_t)$ **P**-a.s. Substituting the shadow price of wealth at time t, i.e. $y^*(t, X_t^*, Y_t)$, and the optimal consumption policy (2.18) in the objective function yields

$$
\begin{aligned}
V(t, X_t^*, Y_t) = \mathbf{E}_t\Bigg[&\int_t^T [u \circ I^+]\big(y^*(t, X_t^*, Y_t)\xi_{t,v}, v\big)\,\mathrm{d}v\\
&+ [U \circ J^+]\big(y^*(t, X_t^*, Y_t)\xi_{t,T}, T\big)\Bigg].
\end{aligned} \tag{B.4}
$$

Taking derivatives with respect to X_t^* in (B.4) and using $y^*\xi_t = y^*(t, X_t^*, Y_t)$ gives

$$V_x(t, X_t^*, Y_t) = \mathbf{E}_t[D_{t,T}]\partial_x y^*(t, X_t^*, Y_t), \tag{B.5}$$

where $D_{t,T}$ is defined in (2.21). Differentiating both sides of (B.3) with respect to wealth produces

$$\mathbf{E}_t[D_{t,T}]\frac{\partial_x y^*(t, X_t^*, Y_t)}{y^*(t, X_t^*, Y_t)} = 1, \tag{B.6}$$

so that

$$V_x(t, X_t^*, Y_t) = y^*(t, X_t^*, Y_t). \tag{B.7}$$

This establishes (2.28). Furthermore, taking logarithmic derivatives on both sides of (B.7) and using (B.6), establishes (2.29).

Finally, differentiating (B.3) with respect to the state variables and using $y^*\xi_t = y^*(t, X_t^*, Y_t)$ gives

$$\begin{aligned}&\mathbf{E}_t\left[\int_t^T Z_1(y^*\xi_v, v)\nabla_{t,y}\xi_{t,v}\,\mathrm{d}v + Z_2(y^*\xi_T, T)\nabla_{t,y}\xi_{t,T}\right]\\ &\quad + \mathbf{E}_t[D_{t,T}]\frac{\partial_y y^*(t, X_t^*, Y_t)}{y^*(t, X_t^*, Y_t)} = 0,\end{aligned}$$

and, as $[V_{xy}/V_{xx}](t, X_t^*, Y_t) = [\partial_y y^*/\partial_x y^*](t, X_t^*, Y_t)$, with the aid of (2.29) and (B.7), we obtain

$$\begin{aligned}\frac{V_{xy}(t, X_t^*, Y_t)}{-V_{xx}(t, X_t^*, Y_t)} = \mathbf{E}_t\Bigg[&\int_t^T \xi_{t,v}Z_1(y^*\xi_v, v)\nabla_{t,y}\log\xi_{t,v}\,\mathrm{d}v\\ &+ \xi_{t,T}Z_2(y^*\xi_T, T)\nabla_{t,y}\log\xi_{t,T}\Bigg],\end{aligned} \tag{B.8}$$

where $\nabla_{t,y}\log\xi_{t,\cdot}$ is the tangent process of $\log\xi_{t,\cdot}$ (see Appendix A). In a Markovian setting the first variation process and the Malliavin derivative are linked by $\nabla_{t,y}\log\xi_{t,v}\sigma^Y(t, Y_t) = \mathcal{D}_t\log\xi_{t,v}$ and $\mathcal{D}_t\log\xi_{t,\cdot} = -H'_{t,\cdot}$. The relation (2.30) follows. □

Proof of (3.12)–(3.13). The limits of interest are found as follows. The definition of the optimal wealth process

$$X^*_{t+h} - X^*_t + \int_t^{t+h} c^*_v \,\mathrm{d}v = \int_t^{t+h} r_v X^*_v \,\mathrm{d}v + \int_t^{t+h} X^*_v (\pi^*_v)' \big[(\mu_v - r_v 1_d)\,\mathrm{d}v + \sigma_v \,\mathrm{d}W_v\big]$$

and the Itô formula

$$\begin{aligned}
&\left(X^*_{t+h} - X^*_t + \int_t^{t+h} c^*_v \,\mathrm{d}v\right)(W_{t+h} - W_t)' \\
&\quad = \int_t^{t+h} (W_v - W_t)'(\mathrm{d}X^*_v + c^*_v \,\mathrm{d}v) \\
&\qquad + \int_t^{t+h} \left(X^*_v - X^*_t + \int_t^v c^*_s \,\mathrm{d}s\right) \mathrm{d}W'_v \\
&\qquad + \int_t^{t+h} X^*_v (\pi^*_v)' \sigma_v \,\mathrm{d}v
\end{aligned}$$

lead to

$$\begin{aligned}
&\mathbf{E}_t\left[\left(X^*_{t+h} - X^*_t + \int_t^{t+h} c^*_v \,\mathrm{d}v\right)(W_{t+h} - W_t)'\right] \\
&\quad = \mathbf{E}_t\left[\int_t^{t+h} (W_v - W_t)'(\mathrm{d}X^*_v + c^*_v \,\mathrm{d}v) + \int_t^{t+h} X^*_v (\pi^*_v)' \sigma_v \,\mathrm{d}v\right] \\
&\quad = \mathbf{E}_t\left[\int_t^{t+h} \big((W_v - W_t)'\big(r_v X^*_v + X^*_v (\pi^*_v)'(\mu_v - r_v 1_d)\big)\right. \\
&\qquad \left. + X^*_v (\pi^*_v)' \sigma_v\big) \,\mathrm{d}v\right]
\end{aligned}$$

and the limit

$$\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\left[\left(X^*_{t+h}-X^*_t+\int_t^{t+h} c^*_v\,\mathrm{d}v\right)(W_{t+h}-W_t)'\right]=X^*_t(\pi^*_t)'\sigma_t. \tag{B.9}$$

Using $\mathbf{E}_t[X^*_t(W_{t+h}-W_t)']=0$,

$$\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\left[\left(\int_t^{t+h} c^*_v\,\mathrm{d}v\right)(W_{t+h}-W_t)'\right]=0$$

and

$$\begin{aligned}\mathbf{E}_t\big[X^*_{t+h}(W_{t+h}-W_t)'\big]&=\mathbf{E}_t\big[\mathbf{E}_{t+h}[F_{t+h,T}](W_{t+h}-W_t)'\big]\\&=\mathbf{E}_t\big[\mathbf{E}_{t+h}\big[F_{t+h,T}(W_{t+h}-W_t)'\big]\big]\\&=\mathbf{E}_t\big[F_{t+h,T}(W_{t+h}-W_t)'\big],\end{aligned}$$

with $F_{t+h,T}\equiv\int_{t+h}^T \xi_{t+h,v}c^*_v\,\mathrm{d}v+\xi_{t+h,T}X^*_T$, enables us to rewrite (B.9) as

$$X_t(\pi^*_t)'\sigma_t=\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\big[F_{t+h,T}(W_{t+h}-W_t)'\big]. \tag{B.10}$$

This establishes (3.14). To get (3.15) expand the coefficient $F_{t+h,T}$ as

$$\begin{aligned}F_{t+h,T}&\equiv\int_{t+h}^T \xi_{t+h,v}c^*_v\,\mathrm{d}v+\xi_{t+h,T}X^*_T\\&=\left(\int_{t+h}^T \xi_{t,v}c^*_v\,\mathrm{d}v+\xi_{t,T}X^*_T\right)\xi_{t+h,t}\\&=\left(F_{t,T}-\int_t^{t+h}\xi_{t,v}c^*_v\,\mathrm{d}v\right)\xi_{t+h,t}\end{aligned}$$

and substitute in (B.10) to write

$$\begin{aligned}X_t(\pi^*_t)'\sigma_t&=\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\big[F_{t+h,T}(W_{t+h}-W_t)'\big]\\&=\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\left[\left(F_{t,T}-\int_t^{t+h}\xi_{t,v}c^*_v\,\mathrm{d}v\right)\xi_{t+h,t}(W_{t+h}-W_t)'\right]\\&=\lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\big[F_{t,T}\xi_{t+h,t}(W_{t+h}-W_t)'\big]\end{aligned}$$

$$- \lim_{h\to 0} \frac{1}{h}\mathbf{E}_t\left[\left(\int_t^{t+h} \xi_{t,v}c_v^*\,\mathrm{d}v\right)\xi_{t+h,t}(W_{t+h}-W_t)'\right]. \quad \text{(B.11)}$$

Another application of the integration by parts formula

$$\begin{aligned}
&\mathbf{E}_t\left[\left(\int_t^{t+h} \xi_{t,v}c_v^*\,\mathrm{d}v\right)\xi_{t+h,t}(W_{t+h}-W_t)'\right]\\
&\quad = \mathbf{E}_t\left[\left(\int_t^{t+h} \xi_{t,v}c_v^*\,\mathrm{d}v\right)\left(\int_t^{t+h} \xi_{v,t}\,\mathrm{d}W_v + \int_t^{t+h} (W_v-W_t)'\,\mathrm{d}\xi_{v,t}\right.\right.\\
&\qquad \left.\left. + \int_t^{t+h} \mathrm{d}[W_v', \xi_{v,t}]\right)\right]
\end{aligned}$$

shows that

$$\lim_{h\to 0} \frac{1}{h}\mathbf{E}_t\left[\left(\int_t^{t+h} \xi_{t,v}c_v^*\,\mathrm{d}v\right)\xi_{t+h,t}(W_{t+h}-W_t)'\right] = 0$$

and (B.11), therefore, becomes

$$X_t(\pi_t^*)'\sigma_t = \lim_{h\to 0} \frac{1}{h}\mathbf{E}_t\big[F_{t,T}\xi_{t+h,t}(W_{t+h}-W_t)'\big]$$

which corresponds to (3.15).

For the third expression (3.16) use the forward and backward representations of optimal wealth

$$\xi_t X_t^* + \int_0^t \xi_v c_v^*\,\mathrm{d}v = X_0 + \int_0^t \xi_v X_v^*\big((\pi_v^*)'\sigma_v - \theta_v'\big)\,\mathrm{d}W_v = \mathbf{E}_t[F_{0,T}]$$

to derive

$$\begin{aligned}
&\mathbf{E}_t\big[\mathbf{E}_{t+h}[F_{0,T}](W_{t+h}-W_t)'\big]\\
&\quad = \mathbf{E}_t\left[\left(X_0 + \int_0^{t+h} \xi_v X_v^*\big((\pi_v^*)'\sigma_v - \theta_v'\big)\,\mathrm{d}W_v\right)(W_{t+h}-W_t)'\right]\\
&\quad = \mathbf{E}_t\left[\left(\int_0^{t+h} \xi_v X_v^*\big((\pi_v^*)'\sigma_v - \theta_v'\big)\,\mathrm{d}W_v\right)(W_{t+h}-W_t)'\right]\\
&\quad = \mathbf{E}_t\left[\int_t^{t+h} \xi_v X_v^*\big((\pi_v^*)'\sigma_v - \theta_v'\big)'\,\mathrm{d}v\right].
\end{aligned}$$

From this equality and $\mathbf{E}_t[\mathbf{E}_{t+h}[F_{0,T}](W_{t+h} - W_t)] = \mathbf{E}_t[F_{0,T}(W_{t+h} - W_t)]$, it follows that

$$\lim_{h\to 0} \frac{1}{h}\mathbf{E}_t\big[F_{0,T}(W_{t+h} - W_t)\big] = \xi_t X_t^*\big(\sigma_t'(\pi_t^*) - \theta_t\big).$$

Substituting $\mathbf{E}_t[F_{0,T}(W_{t+h} - W_t)] = \xi_t \mathbf{E}_t[F_{t,T}(W_{t+h} - W_t)]$ on the left-hand side we conclude that

$$X_t^*\pi_t^* = (\sigma_t')^{-1}\left(X_t^*\theta_t + \lim_{h\to 0}\frac{1}{h}\mathbf{E}_t\big[F_{t,T}(W_{t+h} - W_t)\big]\right) \tag{B.12}$$

thereby establishing (3.16). □

Proof of Proposition 3. See Theorem 4 and Corollary 2 in Detemple et al. (2005c). □

Proof of Proposition 4. The functions g_i^α where $(i, \alpha) \in \{1, 2\} \times \{MV, H\}$ satisfy the conditions of Theorem 1 in Detemple et al. (2005d). The result follows. □

Proof of Proposition 5. The introduction of functions f_i, $i \in \{1, 2\}$, puts the problem into the setting of Theorem 2 in Detemple et al. (2005d). The proposition follows from their result. □

Proof of Proposition 6. The portfolio allocation problem is formulated so as to permit the application of Theorem 3 in Detemple et al. (2005d). The result of the proposition follows immediately. □

References

Brandt, M.W., Goyal, A., Santa-Clara, P., Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *Review of Financial Studies* 18, 831–873.

Breeden, D. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7, 265–296.

Brennan, M., Schwartz, E., Lagnado, R. (1997). Strategic asset allocation. *Journal of Economic Dynamics and Control* 21, 1377–1403.

Clément, E., Lamberton, D., Protter, P. (2002). An analysis of a least squares regression method for American option pricing. *Finance and Stochastics* 5, 449–471.

Cox, J.C., Huang, C.-f. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory* 49, 33–83.

Cvitanic, J., Goukasian, L., Zapatero, F. (2003). Monte Carlo computation of optimal portfolio in complete markets. *Journal of Economic Dynamics and Control* 27, 971–986.

Detemple, J.B., Rindisbacher, M. (2005). Explicit solutions for a portfolio problem with incomplete markets and investment constraints. *Mathematical Finance* 15, 539–568.

Detemple, J.B., Garcia, R., Rindisbacher, M. (2003). A Monte-Carlo method for optimal portfolios. *Journal of Finance* 58, 401–446.

Detemple, J.B., Garcia, R., Rindisbacher, M. (2005a). Representation formulas for Malliavin derivatives of diffusion processes. *Finance and Stochastics* 9, 349–367.

Detemple, J.B., Garcia, R., Rindisbacher, M. (2005b). Intertemporal asset allocation: A comparison of methods. *Journal of Banking and Finance* 29, 2821–2848.
Detemple, J.B., Garcia, R., Rindisbacher, M. (2005c). Asymptotic properties of Monte Carlo estimators of diffusion processes. *Journal of Econometrics*, in press.
Detemple, J.B., Garcia, R., Rindisbacher, M. (2005d). Asymptotic properties of Monte Carlo estimators of derivatives of diffusion processes. *Management Science* 51, 1657–1675.
Duffie, D., Glynn, P. (1995). Efficient Monte Carlo simulation of security prices. *Annals of Applied Probability* 5, 897–905.
Egloff, D. (2005). Monte Carlo algorithms for optimal stopping and statistical learning. *Annals of Applied Probability* 15, 1396–1432.
Glasserman, P., Yu, B. (2004). Number of paths versus number of basis functions in American option pricing. *Annals of Applied Probability* 14, 2090–2119.
Gobet, E., Lemor, J.-P., Warin, X. (2005). A regression-based Monte-Carlo method to solve backward stochastic differential equations. *Annals of Applied Probability* 15, 2002–2172.
Karatzas, I., Shreve, S.E. (1998). *Methods of Mathematical Finance*. Springer-Verlag, New York.
Karatzas, I., Lehoczky, J.P., Shreve, S.E. (1987). Optimal portfolio and consumption decisions for a "small investor" on a finite horizon. *SIAM Journal of Control and Optimization* 25, 1557–1586.
Kunita, H. (1986). *Lectures on Stochastic Flows and Applications*. Springer-Verlag, Berlin/Heidelberg/New York/Tokyo.
Liu, J. (1998). Portfolio selection in stochastic environments. *Working paper*, Stanford University.
Longstaff, F., Schwartz, E. (2001). Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies* 14, 113–147.
Malliavin, P. (1997). *Stochastic Analysis*. Springer-Verlag, Berlin/Heidelberg/New York.
Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7, 77–91.
Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous time model. *Journal of Economic Theory* 3, 273–413.
Ocone, D., Karatzas, I. (1991). A generalized Clark representation formula, with application to optimal portfolios. *Stochastics and Stochastics Reports* 34, 187–220.
Press, W., Teukolsky, S., Vetterling, W., Flannery, B. (1992). Numerical Recipes in C: The Art of Scientific Computing, second ed. Cambridge University Press, Cambridge.
Pliska, S. (1986). A stochastic calculus model of continuous trading: Optimal portfolios. *Mathematics of Operations Research* 11, 371–382.
Schroder, M., Skiadas, C. (1999). Optimal consumption and portfolio selection with stochastic differential utility. *Journal of Economic Theory* 89, 68–126.
Talay, D., Tubaro, L. (1990). Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and its Application* 8, 483–509.
Tsitsiklis, J., Van Roy, B. (2001). Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks* 12, 694–703.
Wachter, J. (2002). Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets. *Journal of Financial and Quantitative Analysis* 37, 63–91.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15022-2

Chapter 22

Duality Theory and Approximate Dynamic Programming for Pricing American Options and Portfolio Optimization[1]

Martin B. Haugh

Department of IE and OR, Columbia University, New York, NY 10027, USA
E-mail: martin.haugh@columbia.edu

Leonid Kogan

Sloan School of Management, MIT, Cambridge, MA 02139, USA
E-mail: lkogan@mit.edu

Abstract

This chapter describes how duality and approximate dynamic programming (ADP) methods can be used in financial engineering. It focuses on American option pricing and portfolio optimization problems when the underlying state space is high-dimensional. In general, it is not possible to solve these problems exactly due to the so-called "curse of dimensionality" and as a result, approximate solution techniques are required. ADP and dual-based methods have been proposed for constructing and evaluating good approximate solutions to these problems. In this chapter we describe these methods. Some directions for future research are also outlined.

1 Introduction

Portfolio optimization and American option pricing problems are among the most important problems in financial engineering. Portfolio optimization problems occur throughout the financial services as pension funds, mutual funds, insurance companies, endowments and other financial entities all face the fundamental problem of dynamically allocating their resources across different securities in order to achieve a particular goal. These problems are often

[1] This chapter is a revised and extended version of Haugh [Haugh, M.B. (2003). Duality theory and simulation in financial engineering. In: Chick, S., Sánchez, P.J., Ferrin, D., Morrice, D.J. (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, IEEE Press, Piscataway, NJ, pp. 327–334].

very complex owing to their dynamic and stochastic nature, their high dimensionality and the complexity of real-world constraints. While researchers have developed a number of sophisticated models for addressing these problems, the current state-of-the-art is such that explicit solutions are available only in very special circumstances. (See, for example, Merton, 1990; Cox and Huang 1989; Karatzas and Shreve, 1997; Liu, 1998.)

American option pricing has also presented several challenges to the field of financial engineering. Even in the simple Black–Scholes framework (Black and Scholes, 1973), a closed form expression for the price of an American put option is not available and so it must therefore be computed numerically. This does not present a challenge when there is just one or two underlying securities. However, as pricing an American option amounts to solving an optimal stopping problem, Bellman's curse of dimensionality implies that pricing high-dimensional American options using standard numerical techniques is not practically feasible. Unfortunately, the same conclusion also applies to solving general high-dimensional portfolio optimization problems.

Because these high-dimensional problems occur frequently in practice, they are of considerable interest to both researchers and practitioners. In recent years there has been some success in tackling these problems using approximate dynamic programming (ADP) and dual-based methods. ADP methods (see, for example, Bertsekas and Tsitsiklis, 1996) have had considerable success in tackling large-scale complex problems and have recently been applied successfully to problems in financial engineering (Brandt et al., 2005; Longstaff and Schwartz, 2001; Tsitsiklis and Van Roy, 2001). One difficulty with ADP, however, is in establishing how far the sub-optimal ADP solution to a given problem is from optimality. In the context of optimal stopping problems and pricing American options, Haugh and Kogan (2004) and Rogers (2002) developed dual formulations[2] which allows one to evaluate sub-optimal strategies, including those obtained from ADP methods (see, for example, Haugh and Kogan, 2004; Anderson and Broadie, 2004; Glasserman and Yu, 2004; Chen and Glasserman, 2007). A stochastic duality theory also exists for portfolio optimization problems and this has been developed by many researchers in recent years (see, for example, Shreve and Xu, 1992a, 1992b; He and Pearson, 1991; Cvitanic and Karatzas, 1992; Karatzas and Shreve, 1997). While this theory has had considerable success in characterizing optimal solutions, explicit solutions are still rare (see Rogers, 2003). Recently Haugh et al. (2003) have shown how some of these dual formulations can be used to evaluate suboptimal policies by constructing lower and upper bounds on the true optimal value function. These suboptimal policies could be simple heuristic policies or policies resulting from some approximation techniques such as ADP.

[2] The dual formulation in these papers relies on an alternative characterization of an optimal stopping problem, which can be traced back to Davis and Karatzas (1994).

Another promising approach to approximate solution of dynamic portfolio choice problems is based on the equivalent linear programming formulation of the dynamic program (see, for example, de Farias and Van Roy, 2003, 2004). The approximate linear programming formulation provides an approximation to the portfolio policy as well as an upper bound on the value function. This approach is computationally intensive and its ability to handle large-scale practical problems still needs to be evaluated. Some encouraging results in this direction are obtained by Han (2005).

Simulation techniques play a key role in both the ADP and dual-based evaluation methods that have been used to construct and evaluate solutions to these problems. While it has long been recognized that simulation is an indispensable tool for financial engineering (see the surveys of Boyle et al., 1997; Staum, 2002), it is only recently that simulation has begun to play an important role in solving *control* problems in financial engineering. These control problems include portfolio optimization and the pricing of American options, and they are the focus of this paper.

The remainder of the paper is outlined as follows. Section 2 describes the American option pricing problem. We briefly describe the ADP methods in Section 2.1 after which we will focus on the duality theory for optimal stopping in Section 2.2. Section 2.3 discusses extensions and other applications of these dual-based ideas. We then conclude Section 2 by outlining some directions for future research in Section 2.4. Section 3 concentrates on portfolio optimization. In Sections 3.1 and 3.2, respectively, we describe the portfolio optimization framework and review the corresponding duality theory. In Section 3.3 we show how an upper bound on the portfolio optimization problem can be obtained and we summarize the algorithm for obtaining lower and upper bounds in Section 3.4. We conclude by outlining directions for future research in Section 3.5. Results will not be presented in their full generality, and technical details will often be omitted as we choose to focus instead on the underlying concepts and intuition.

2 Pricing American options

The financial market. We assume there exists a dynamically complete financial market that is driven by a vector-valued Markov process, $X_t = (X_t^1, \ldots, X_t^n)$. In words, we say a financial market is dynamically complete if any random variable, W_T, representing a terminal cash-flow can be attained by using a *self-financing trading strategy*. (A self-financing trading strategy is a strategy where changes in the value of the portfolio are only due to accumulation of dividends and capital gains or losses. In particular, no net addition or withdrawal of funds is allowed after date $t = 0$ and any new purchases of securities must be financed by the sale of other securities.) X_t represents the time t vector of risky asset prices as well as the values of any relevant state variables in the market. We also assume there exists a risk-free security whose

time t price is $B_t = e^{rt}$, where r is the continuously compounded risk-free rate of interest.[3] Finally, since markets are assumed to be dynamically complete, there exists (see Duffie, 1996) a unique risk-neutral valuation measure, $\mathcal{Q}$.

Option payoff. Let $h_t = h(X_t)$ be a nonnegative adapted process representing the payoff of the option so that if it is exercised at time t the holder of the option will then receive h_t.

Exercise dates. The American feature of the option allows the holder of the option to *exercise* it at any of the pre-specified exercise dates in $\mathcal{T} = \{0, 1, \ldots, T\}$.[4]

Option price. The value process of the American option, V_t, is the price process of the option conditional on it not having been exercised before t. It satisfies

$$V_t = \sup_{\tau \geqslant t} \mathrm{E}_t^{\mathcal{Q}}\left[\frac{B_t h_\tau}{B_\tau}\right], \tag{1}$$

where τ is any stopping time with values in the set $\mathcal{T} \cap [t, T]$.

If X_t is high-dimensional, then standard solution techniques such as dynamic programming become impractical and we cannot hope to solve the optimal stopping problem (1) exactly. Fortunately, efficient ADP algorithms for addressing this problem have recently been developed independently by Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001).[5] We now briefly describe the main ideas behind these algorithms, both of which rely on the ability to simulate paths of the underlying state vectors.

2.1 ADP for pricing American options

Once again, the pricing problem at time $t = 0$ is to compute

$$V_0 = \sup_{\tau \in \mathcal{T}} \mathrm{E}_0^{\mathcal{Q}}\left[\frac{h_\tau}{B_\tau}\right]$$

[3] Note that we can easily handle the case where $r_t = r(X_t)$ is stochastic.

[4] Strictly speaking, we consider Bermudan options that may only be exercised at one of a finite number of possible dates. While American options can be exercised at any time in a continuum of dates, in practice it is necessary to discretize time when pricing them numerically. As a result, we do not distinguish between Bermudan and American options in this chapter.

[5] See also Carriére (1996) for the original introduction of regression-based ideas for pricing American options.

and in theory this problem is easily solved using value iteration. In particular, a standard recursive argument implies

$$V_T = h(X_T) \quad \text{and} \quad V_t = \max\left(h(X_t), \mathrm{E}_t^{\mathcal{Q}}\left[\frac{B_t}{B_{t+1}} V_{t+1}(X_{t+1})\right]\right).$$

The price of the option is then given by $V_0(X_0)$ where X_0 is the initial state of the economy. As an alternative to value iteration we could use *Q-value iteration*. If the Q-value function is defined to be the value of the option conditional on it not being exercised today, i.e. the *continuation value* of the option, then we also have

$$Q_t(X_t) = \mathrm{E}_t^{\mathcal{Q}}\left[\frac{B_t}{B_{t+1}} V_{t+1}(X_{t+1})\right].$$

The value of the option at time $t+1$ is then

$$V_{t+1}(X_{t+1}) = \max\big(h(X_{t+1}), Q_{t+1}(X_{t+1})\big)$$

so that we can also write

$$Q_t(X_t) = \mathrm{E}_t^{\mathcal{Q}}\left[\frac{B_t}{B_{t+1}} \max\big(h(X_{t+1}), Q_{t+1}(X_{t+1})\big)\right]. \tag{2}$$

Equation (2) clearly gives a natural analog to value iteration, namely Q-value iteration. As stated earlier, if n is large so X_t is high-dimensional, then both value iteration and Q-value iteration are not feasible in practice. However, we could perform an *approximate* and efficient version of Q-value iteration, and this is precisely what the ADP algorithms of Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001) do. We now describe their main contribution, omitting some of the more specific details that can nevertheless have a significant impact on performance.

The first step is to choose a set of *basis functions*, $\phi_1(\cdot), \ldots, \phi_m(\cdot)$. These basis functions define the *linear architecture* that will be used to approximate the Q-value functions. In particular, we will approximate $Q_t(X_t)$ with

$$\tilde{Q}_t(X_t) = r_t^1 \phi_1(X_t) + \cdots + r_t^m \phi_m(X_t),$$

where $r_t := (r_t^1, \ldots, r_t^m)$ is a vector of time t parameters that is determined by the algorithm which proceeds as follows:

Approximate Q-value iteration

generate N paths of state vector, X, conditional on initial state, X_0
set $\tilde{Q}_T(X_T^i) = 0$ for all $i = 1$ to N
for $t = T-1$ *downto* 1
  regress $B_t \tilde{V}_{t+1}(X_{t+1}^i)/B_{t+1}$ on $(\phi_1(X_t^i), \ldots, \phi_m(X_t^i))$ where
    $\tilde{V}_{t+1}(X_{t+1}^i) := \max\big(h(X_{t+1}^i), \tilde{Q}(X_{t+1}^i)\big)$

set $\tilde{Q}_t(X_t^i) = \sum_k r_t^k \phi_k(X_t^i)$ where the r_t^ks are the estimated regression coefficients
end for
generate M samples of state vector, X_1, conditional on initial state, X_0
set $\tilde{V}_0(X_0) = (\sum_{j=1}^M \max(h(X_1^j), \tilde{Q}_1(X_1^j)))/MB_1$.

The key idea in this algorithm is the use of regression methods to estimate $\tilde{Q}_t(X_t^i)$. In practice, standard least squares regression is used and because this technique is so fast the resulting Q-value iteration algorithm is also very fast. For typical problems that arise in practice, N is often taken to be on the order of 10,000 to 50,000. Obviously, many more details are required to fully specify the algorithm. In particular, parameter values and basis functions need to be chosen. It is generally a good idea to use problem-specific information when choosing the basis functions. For example, if the value of the corresponding European option is available in closed form then that would typically be an ideal candidate for a basis function. Other commonly used basis functions are the intrinsic value of the option and the prices of other related derivative securities that are available in closed from.

Specific implementation details can also vary. While the algorithm described above is that of Tsitsiklis and Van Roy (2001), Longstaff and Schwartz (2001) omit states X_t^i where $h(X_t^i) = 0$ when estimating the regression coefficients, r_t^k for $k = 1, \ldots, m$. They also define $\tilde{V}_{t+1}(X_{t+1}^i)$ so that

$$\tilde{V}_{t+1}(X_{t+1}^i) = \begin{cases} h(X_{t+1}^i), & h(X_{t+1}^i) \geqslant \tilde{Q}(X_{t+1}^i), \\ \tilde{V}_{t+2}(X_{t+2}^i)B_{t+1}/B_{t+2}, & h(X_{t+1}^i) < \tilde{Q}(X_{t+1}^i). \end{cases}$$

In particular, they take $\tilde{V}_{t+1}(X_{t+1}^i)$ to be the *realized* discounted payoff on the ith path as determined by the exercise policy, $\tilde{\tau}$, implicitly defined by $\tilde{Q}_l(\cdot)$, for $l = t+1, \ldots, T$.

In practice, it is quite common for an alternative estimate, $\underline{V}_0$, of V_0 to be obtained by simulating the exercise strategy that is defined by $\tilde{\tau}$. Formally, we define $\tilde{\tau} = \min\{t \in \mathcal{T}\colon \tilde{Q}_t \leqslant h_t\}$ and

$$\underline{V}_0 = \mathrm{E}_0^{\mathcal{Q}}\left[\frac{h_{\tilde{\tau}}}{B_{\tilde{\tau}}}\right].$$

$\underline{V}_0$ is then an unbiased lower bound on the true value of the option. That the estimator is a lower bound follows from the fact $\tilde{\tau}$ is a feasible adapted exercise strategy. Typically, $\underline{V}_0$ is a much better estimator of the true price than $\tilde{V}_0(X_0)$ as the latter often displays a significant upwards bias. Glasserman (2004, Section 8.7) provides a very nice intuition for determining when $\tilde{V}_0(X_0)$ performs poorly as an estimator and, using the duality ideas of Section 2.2, he relates the quality of $\tilde{V}_0(X_0)$ to the quality of the chosen basis functions.

These ADP algorithms have performed surprisingly well on realistic high-dimensional problems (see Longstaff and Schwartz, 2001 for numerical examples) and there has also been considerable theoretical work (e.g. Tsitsiklis and Van Roy, 2001; Clemént et al., 2002) justifying this. Clemént et al. (2002), for example, show that the Longstaff and Schwartz algorithm converges as the number of paths, N, goes to infinity and that the limiting price, $\tilde{V}_0(X_0)$, equals the true price, V_0, if Q_t can be written as a linear combination of the chosen basis functions.

Haugh and Kogan (2004) also show that for any approximation, $\tilde{Q}_t$, the quality of the lower bound, $\underline{V}_0$, satisfies

$$V_0 - \underline{V}_0 \leqslant \mathrm{E}_0^{\mathcal{Q}}\left[\sum_{t=0}^{T} \frac{|\tilde{Q}_t - Q_t|}{B_t}\right].$$

While this may suggest that the quality of the lower bound deteriorates linearly in the number of exercise periods, in practice this is not the case. The quality of $\underline{V}_0$, for example, can be explained in part by noting that exercise errors are never made as long as $Q_t(\cdot)$ and $\tilde{Q}_t(\cdot)$ lie on the *same* side of the optimal exercise boundary. This means in particular, that it is possible to have large errors in $\tilde{Q}_t(\cdot)$ that do not impact the quality of $\underline{V}_0$.

More recently, it has been shown (see Glasserman, 2004; Glasserman and Yu, 2004) how the ADP – regression methods relate to the *stochastic mesh* method of Broadie and Glasserman (1997). In addition, Glasserman and Yu (2004) also study the trade-off between the number of paths, N, and the number of basis functions, m, when there is a finite computational budget available.

To complete this section, it is worth mentioning that there is an alternative to approximating the value function or Q-value function when using ADP methods (or indeed any other approximation methods). That is, we could choose instead to approximate the optimal exercise frontier. The exercise frontier is the boundary in X-space whereby it is optimal to exercise on one side of the boundary and to continue on the other side. It is possible to construct ADP methods that directly approximate this exercise boundary without directly approximating the value function. These methods often require work that is quadratic in the number of exercise periods. That said, in general it is very difficult to conduct a formal comparison between methods that approximate the exercise frontier and methods that approximate the value function.

2.2 *Duality theory for American options*

While ADP methods have been very successful, a notable weakness is their inability to determine how far the ADP solution is from optimality in any given problem. Haugh and Kogan (2004) and Rogers (2002) independently developed duality-based methods that can be used for constructing upper bounds on the true value function. Haugh and Kogan showed that any approximate

solution, arising from ADP or other[6] methods, could be evaluated by using it to construct an upper[7] bound on the true value function. We also remark that Broadie and Glasserman (1997) were the first to demonstrate that tight lower and upper bounds could be constructed using simulation techniques. Their method, however, does not work with arbitrary approximations to the value function and does not appear to be as efficient as the dual-ADP techniques. We now describe these duality-based methods.

For an arbitrary adapted supermartingale, π_t, the value of an American option, V_0, satisfies

$$\begin{aligned} V_0 &= \sup_{\tau\in\mathcal{T}} \mathrm{E}_0^{\mathcal{Q}}\left[\frac{h_\tau}{B_\tau}\right] = \sup_{\tau\in\mathcal{T}} \mathrm{E}_0^{\mathcal{Q}}\left[\frac{h_\tau}{B_\tau} - \pi_\tau + \pi_\tau\right] \\ &\leqslant \sup_{\tau\in\mathcal{T}} \mathrm{E}_0^{\mathcal{Q}}\left[\frac{h_\tau}{B_\tau} - \pi_\tau\right] + \pi_0 \leqslant \mathrm{E}_0^{\mathcal{Q}}\left[\max_{t\in\mathcal{T}}\left(\frac{h_t}{B_t} - \pi_t\right)\right] + \pi_0, \end{aligned} \tag{3}$$

where the first inequality follows from the optional sampling theorem for supermartingales. Taking the infimum over all supermartingales, π_t, on the right-hand side of (3) implies

$$V_0 \leqslant U_0 := \inf_{\pi} \mathrm{E}_0^{\mathcal{Q}}\left[\max_{t\in\mathcal{T}}\left(\frac{h_t}{B_t} - \pi_t\right)\right] + \pi_0. \tag{4}$$

On the other hand, it is known (see e.g. Duffie, 1996) that the process V_t/B_t is itself a supermartingale, which implies

$$U_0 \leqslant \mathrm{E}_0^{\mathcal{Q}}\left[\max_{t\in\mathcal{T}}\left(\frac{h_t}{B_t} - \frac{V_t}{B_t}\right)\right] + V_0.$$

Since $V_t \geqslant h_t$ for all t, we conclude that $U_0 \leqslant V_0$. Therefore, $V_0 = U_0$, and equality is attained when $\pi_t = V_t/B_t$.

It is of interest to note that we could have restricted ourselves to the case where π_t is a strict martingale, as was the case with Rogers (2002). In that case, the Doob–Meyer decomposition theorem and the supermartingale property of V_t/B_t imply the existence of a martingale, M_t, and an increasing, predictable process, A_t, satisfying $A_0 = 0$ and

$$\frac{V_t}{B_t} = M_t - A_t.$$

Taking $\pi_t = M_t$ in (4) we again obtain $U_0 \leqslant V_0$ implying once again that $V_0 = U_0$. These results demonstrate that an upper bound on the price of the American option can be constructed simply by evaluating the right-hand side

[6] See, for example, the iterative technique of Kolodko and Schoenmakers (2006) who construct upper as well as lower bounds on the true option price using the dual formulations we describe in this section.
[7] As we saw in Section 2.1, a lower bound is easy to compute given an approximation to the value function.

of (3) for a given supermartingale, π_t. In particular, if such a supermartingale satisfies $\pi_t \geqslant h_t/B_t$, the option price V_0 is bounded above by π_0.

When the supermartingale π_t in (3) coincides with the discounted option value process, V_t/B_t, the upper bound on the right-hand side of (3) equals the true price of the American option. This suggests that a tight upper bound can be obtained by using an accurate approximation, $\tilde{V}_t$, to define π_t. One possibility[8] is to define π_t as the following martingale[9]

$$\pi_0 = \tilde{V}_0, \tag{5}$$

$$\pi_{t+1} = \pi_t + \frac{\tilde{V}_{t+1}}{B_{t+1}} - \frac{\tilde{V}_t}{B_t} - \mathrm{E}_t\left[\frac{\tilde{V}_{t+1}}{B_{t+1}} - \frac{\tilde{V}_t}{B_t}\right]. \tag{6}$$

Let $\overline{V}_0$ denote the upper bound we obtain from (3) corresponding to our choice of supermartingale in (5) and (6). Then it is easy to see that the upper bound is explicitly given by

$$\overline{V}_0 = \tilde{V}_0 + \mathrm{E}_0^{\mathcal{Q}}\left[\max_{t\in\mathcal{T}}\left(\frac{h_t}{B_t} - \frac{\tilde{V}_t}{B_t} + \sum_{j=1}^{t}\mathrm{E}_{j-1}^{\mathcal{Q}}\left[\frac{\tilde{V}_j}{B_j} - \frac{\tilde{V}_{j-1}}{B_{j-1}}\right]\right)\right]. \tag{7}$$

As may be seen from (7), obtaining an accurate estimate of $\overline{V}_0$ can be computationally demanding. First, a number of sample paths must be simulated to estimate the outermost expectation on the right-hand side of (7). While this number can be quite small in practice, we also need to accurately estimate a conditional expectation at each time period along each simulated path. This requires some effort and clearly variance reduction methods would be useful in this context. Alternatively, if the initial value function approximation comes from ADP methods then, as suggested by Glasserman and Yu (2004), it might be possible to choose the basis functions in such a way that the conditional expectations in (7) can be computed analytically. In that case the need for conducting nested simulations would not arise.

2.3 Extensions

A number of variations and extensions of these algorithms have also been developed recently and are a subject of ongoing research. Andersen and Broadie (2004), for example, construct upper bounds by using an approximation to the optimal exercise frontier instead of an approximation to the Q-value function, while Meinshausen and Hambly (2004) use similar ideas to price options that may be exercised on multiple occasions. Jamshidan (2003)

[8] See Haugh and Kogan (2004) and Andersen and Broadie (2004) for further comments relating to the superiority of taking π_t to be a strict martingale.

[9] Haugh and Kogan (2004) also propose an alternative where π_t is constructed from $\tilde{V}_t$ in a multiplicative manner.

developed a multiplicative dual approach for constructing upper bounds and Chen and Glasserman (2007) compare this multiplicative dual approach with the additive approach of Haugh and Kogan (2004) and Rogers (2002). In this section we briefly describe these extensions. All of them, with the exception of Meinshausen and Hambly (2004), deal with pricing American options, or equivalently, optimal stopping problems. Brandt et al. (2005) extended the ADP ideas for optimal stopping to portfolio optimization problems. Haugh et al. (2003) showed how a duality theory that had already existed for portfolio optimization problems could be used in practice to create upper bounds on the solutions to these problems. These extensions to portfolio optimization problems will be described in Section 3.

2.3.1 Upper bounds from stopping rules

Approximations to the optimal exercise frontier can also be used to construct upper bounds. For example, suppose τ_i for $i = 1, \ldots, T$ is a sequence of stopping times with the property $\tau_i \geqslant i$ for all i. We interpret τ_i as the time at which the American option should be exercised (under some policy) given that it has not already been exercised before time i. These stopping times might, for example, be constructed from an approximation, $\tilde{Q}_t$, to the Q-value function so that $\tau_i := \min\{t \in \mathcal{T},\ t \geqslant i\colon\ \tilde{Q}_t \leqslant h_t\}$. Alternatively, τ_i may derive from a direct approximation to the exercise frontier. In this case,

$$\tau_i := \min\{t \in \mathcal{T},\ t \geqslant i\colon\ g_t = 1\}, \tag{8}$$

where $g_t = 1$ if the policy says "exercise" and $g_t = 0$ if the policy says "continue." Note that it is not necessary to have an approximation to the value function available when τ_i is defined in this manner.

Regardless of how τ_i is defined we can use it to construct martingales by setting $\tilde{M}_t := \sum_j^t \Delta_j$ where

$$\Delta_j := \mathrm{E}_j^{\mathcal{Q}}[h_{\tau_j}] - \mathrm{E}_{j-1}^{\mathcal{Q}}[h_{\tau_j}] = V_j - Q_{j-1}. \tag{9}$$

We can then take $\pi_t := \tilde{M}_t$ in (4) to construct upper bounds as before. It is necessary to simulate the stopping time τ_i as defined by (8) to estimate the Δ_js. This additional or *nested* simulation is required at each point along each simulated path when estimating the upper bound. This therefore suggests that the computational effort required to compute $\overline{V}_0$ when a stopping time is used is quadratic in the number of time periods. In contrast, it appears that the computational effort is linear when $\tilde{Q}_t$ is used to construct the upper bound of Section 2.2. However, an approximation to the optimal exercise frontier is likely to be more 'accurate' than an approximation to the Q-value function and so a more thorough analysis would be required before the superiority of one approach over the other could be established.

The stopping rule approach was proposed by Andersen and Broadie (2004) but see also Glasserman (2004) for further details. It is also worth mentioning that it is straightforward to combine the two approaches. In particular, an

explicit approximation, $\tilde{Q}_t$, could be used in some regions of the state space to estimate $\tilde{M}_t$ while a nested simulation to estimate the Δ_js could be used in other regions.

2.3.2 The multiplicative dual approach

An alternative dual formulation, the *multiplicative dual*, was recently formulated by Jamshidian (2003) for pricing American options. Using the *multiplicative* Doob–Meyer decomposition for supermartingales, Jamshidan showed that the American option price, V_0, could be represented as

$$V_0 = \inf_{M \in \mathcal{M}^+} \mathrm{E}_0^M \left[\max_{0 \leqslant t \leqslant T} \frac{h_t}{M_t} \right] := \inf_{M \in \mathcal{M}^+} \mathrm{E}_0^{\mathcal{Q}} \left[\max_{0 \leqslant t \leqslant T} \frac{h_t}{M_t} M_T \right], \tag{10}$$

where $\mathcal{M}^+$ is the set of all positive martingales, M_t, with $M_0 = 1$. Equation (10) suggests that if we choose a 'good' martingale, $\tilde{M}_t \in \mathcal{M}^+$ with $\tilde{M}_0 = 1$, then

$$\overline{V}_0 := \mathrm{E}_0^{\mathcal{Q}} \left[\max_{0 \leqslant t \leqslant T} \frac{h_t}{\tilde{M}_t} \tilde{M}_T \right]$$

should provide a good upper bound on V_0. As was the case with the *additive* approaches of Section 2.2, it is possible to construct a candidate martingale, $\tilde{M}_t$, using an approximation, $\tilde{V}_t$, to the true value function, V_t. As usual, this upper bound can be estimated using Monte Carlo methods.

Chen and Glasserman (2007) compare this multiplicative dual formulation with the additive-dual formulations of Rogers (2002) and Haugh and Kogan (2004). They show that neither formulation dominates the other in the sense that any multiplicative dual can be improved by an additive dual and that any additive dual can be improved by a multiplicative dual. They also compare the bias and variance of the two formulations and show that either method may have a smaller bias. The multiplicative method, however, typically has a variance that grows much faster than the additive method. While the multiplicative formulation is certainly of theoretical interest, in practice it is likely to be dominated by the additive approach.

Bolia et al. (2004) show that the dual formulation of Jamshidian may be interpreted using an importance sampling formulation of the problem. They then use importance sampling techniques and nonnegative least square methods for function approximation in order to estimate the upper bound associated with a given approximation to the value function. Results are given for pricing an American put option on a single stock that follows a geometric Brownian motion. While they report some success, considerable work remains before these ideas can be applied successfully to high-dimensional problems. In addition, since the multiplicative dual formulation tends to have a much higher variance than the additive formulation, importance sampling methods might have more of an impact if they could be successfully applied to additive formulations. More generally, importance sampling methods should also be of interest when constructing value function approximations in the first place.

2.3.3 Multiple exercise opportunities

In an interesting extension, Meinshausen and Hambly (2004), extend the ADP-dual techniques to *multiple exercise* options. If $\mathcal{T} = \{0, 1, \ldots, T\}$ are the possible exercise dates, then a multiple exercise option with n exercise opportunities may be exercised at any of n dates in $\mathcal{T}$. Clearly $n \leqslant T+1$ and the case where $n = 1$ corresponds to a standard American option. The standard examples of a multiple exercise option is a *swing* option that is traded in the energy derivatives industry. A swing option gives the holder a fixed number of exercise opportunities when electricity may be purchased at a given price. Meinshausen and Hambly (2004) apply ADP methods to construct an approximation to (and lower bound on) the price of swing options.[10] They then use this approximation and a dual formulation to construct an upper bound on the true price of the option. This is completely analogous to the methods for pricing high-dimensional American options though the computational requirements appear to be much greater.

2.4 Directions for future research

There are many possible directions for future research. First, it should be possible to employ ADP and duality ideas to other classes of problems. There has of course already been some success in this direction. As described in Section 2.3.2, Meinshausen and Hambly (2004) have extended these results to option pricing problems where multiple exercises are permitted. Haugh et al. (2003) also developed analogous results for dynamic portfolio optimization problems.[11] It should therefore be possible to extend and develop these techniques for solving other classes of control problems. Of particular interest are *real options* problems, which typically embed American-style or multiple-exercise features.

Because ADP-dual techniques require simulation methods and therefore often demand considerable computational effort, it is clear that variance reduction techniques should be of value, particularly as ever more complex problems are solved. We expect importance sampling to be especially useful in this regard. First, it may be used for estimating the value associated with a given approximately optimal policy.[12] Second, and perhaps more interesting, it should prove especially valuable in actually constructing the approximately optimal policy itself. This is because importance sampling ideas can be used to focus the computational effort on the more 'important' regions of the state space when approximating the value function. While these ideas are not new, they have certainly not been fully explored within the ADP literature.

[10] They also price *chooser flexible caps*, fixed income derivative securities that give the holder the right to exercise a given number of caplets over a given horizon.

[11] See Section 3.

[12] See Bolia et al. (2004) and Section 2.3.2.

Estimating the so-called 'greeks' is also a particularly interesting problem. While ADP-dual ideas have now been very successful for pricing high-dimensional American options, efficiently computing the greeks for these problems remains a difficult[13] task.

Another future research direction is to use the dual formulation, possibly in conjunction with the primal formulation, to construct approximate value functions or exercise frontiers. The resulting 'dual' or 'primal–dual' algorithms would be of theoretical interest and we expect they would, in some instances, be superior to 'primal' algorithms that have already been developed. While it is straightforward to design admittedly simple primal–dual style algorithms,[14] there appears to have been little work done on this topic. This, presumably, is due to the great success that ADP-regression algorithms have had in quickly generating good approximations to the true option price. It is possible, however, that this will become a more active research area as more challenging classes of problems are tackled with ADP techniques.

3 Portfolio optimization

Motivated by the success of ADP methods for pricing American options, Brandt et al. (2005) apply similar ideas to approximately solve a class of high-dimensional portfolio optimization problems. In particular, they simulate a large number of sample paths of the underlying state variables and then working backwards in time, they use *cross path regressions* (as we described in the approximate Q-value iteration algorithm) to efficiently compute an approximately optimal strategy. Propagation of errors is largely avoided, and though the price for this is an algorithm that is quadratic in the number of time periods, their methodology can comfortably handle problems with a large number of time periods. Their specific algorithm does not handle portfolio constraints and certain other complicating features, but it should be possible to tackle these extensions using the ADP methods that they and others have developed.

As was the case with ADP solutions to optimal stopping problems, a principal weakness of ADP solutions to portfolio optimization problems is the difficulty in determining how far a given solution to a given problem is from optimality. This issue has motivated in part the research of Haugh et al. (2005) who use portfolio duality theory to evaluate the quality of suboptimal solutions to portfolio optimization problems by constructing lower and upper bounds on the optimal value function. These bounds are evaluated by simulating the stochastic differential equations (see Kloeden and Platen, 1992) that describe the evolution of the state variables in the model in question. In Section 3.2 we

[13] See Kaniel et al. (2006) for an application where dual methods are employed to estimate the greeks of Bermudan-style options.

[14] See, for example, Haugh and Kogan (2004).

describe the particular portfolio duality theory that was used in Haugh et al. (2005) and that was developed by Xu (1990), Shreve and Xu (1992a, 1992b), Karatzas et al. (1991), and Cvitanic and Karatzas (1992).

Before doing so, we remark that the duality theory of Section 3.2 applies mainly to problems in continuous time. ADP techniques, on the other hand, are generally more suited to a discrete time framework. This inconsistency can be overcome by extrapolating discrete-time ADP solutions to construct continuous-time solutions.

3.1 The model

We now state a portfolio choice problem under incomplete markets and portfolio constraints. The problem is formulated in continuous time and stock prices follow diffusion processes.

The investment opportunity set. There are N stocks and an instantaneously riskfree bond. The vector of stock prices is denoted by $P_t = (P_{1t}, \ldots, P_{Nt})$ and the instantaneously riskfree rate of return on the bond is denoted by r_t. Without loss of generality, stocks are assumed to pay no dividends. The instantaneous moments of asset returns depend on the M-dimensional vector of state variables X_t:

$$r_t = r(X_t), \tag{11a}$$

$$\mathrm{d}P_t = P_t\big[\mu_P(X_t)\,\mathrm{d}t + \Sigma_P(X_t)\,\mathrm{d}B_t\big], \tag{11b}$$

$$\mathrm{d}X_t = \mu_X(X_t)\,\mathrm{d}t + \Sigma_X(X_t)\,\mathrm{d}B_t, \tag{11c}$$

where $P_0 = 1$, $X_0 = 0$, $B_t = (B_{1t}, \ldots, B_{Nt})$ is a vector of N independent Brownian motions, μ_P and μ_X are N- and M-dimensional drift vectors, and Σ_P and Σ_X are diffusion matrices of dimension N by N and M by N, respectively. The diffusion matrix of the stock return process Σ_P is lower-triangular and nondegenerate: $x^\top \Sigma_P \Sigma_P^\top x \geqslant \epsilon \|x\|^2$ for all x and some $\epsilon > 0$. Then, one can define a process η_t, given by

$$\eta_t = \Sigma_{Pt}^{-1}(\mu_{Pt} - r_t).$$

In a market without portfolio constraints, η_t corresponds to the vector of instantaneous market prices of risk of the N stocks (see, e.g., Duffie, 1996, Section 6.G). The process η_t is assumed to be square-integrable so that

$$\mathrm{E}_0\left[\int_0^T \|\eta_t\|^2\,\mathrm{d}t\right] < \infty.$$

Portfolio constraints. A portfolio consists of positions in the N stocks and the riskfree bond. The proportional holdings of risky assets in the total portfolio value are denoted by $\theta_t = (\theta_{1t}, \ldots, \theta_{Nt})$. The portfolio policy is assumed to

satisfy a square integrability condition: $\int_0^T \|\theta_t\|^2 \,\mathrm{d}t < \infty$ almost surely. The value of the portfolio changes according to

$$\mathrm{d}W_t = W_t\big\{\big[r_t + \theta_t^\top(\mu_{Pt} - r_t)\big]\,\mathrm{d}t + \theta_t^\top \Sigma_{Pt}\,\mathrm{d}B_t\big\}. \tag{12}$$

The portfolio weights are restricted to lie in a closed convex set, $\mathbf{K}$, containing the zero vector:

$$\theta_t \in \mathbf{K}. \tag{13}$$

For example, if short sales are not allowed, then the constraint set takes the form $\mathbf{K} = \{\theta\colon\ \theta \geqslant 0\}$. If in addition to prohibiting short sales, borrowing is not allowed, then $\mathbf{K} = \{\theta\colon\ \theta \geqslant 0,\ 1^\top\theta \leqslant 1\}$ where $1^\top = (1, \ldots, 1)$. The set $\mathbf{K}$ can be constant, or it can depend on time and the values of the exogenous state variables.

The objective function. For simplicity, the portfolio policy is chosen to maximize the expected utility of wealth at the terminal date T, $\mathrm{E}_0[U(W_T)]$. Preferences over intermediate consumption would be easy to incorporate. The function $U(W)$ is assumed to be strictly monotone with positive slope, concave, and smooth. Moreover, it is assumed to satisfy the Inada conditions at zero and infinity: $\lim_{W\to 0} U'(W) = \infty$ and $\lim_{W\to\infty} U'(W) = 0$. For instance, a common choice is a constant relative risk aversion (CRRA) utility function $U(W) = W^{1-\gamma}/(1-\gamma)$.

In summary, the portfolio choice problem is to solve for

$$V_0 := \sup_{\{\theta_t\}} \mathrm{E}_0\big[U(W_T)\big] \quad \text{subject to} \quad \text{(11), (12) and (13)}, \tag{$\mathcal{P}$}$$

where V_0 denotes the value function at 0.

3.2 *Review of the duality theory*

In this section we review the duality theory for the constrained portfolio optimization problem. In particular, the version of duality used in Haugh et al. (2005) is based on the work of Cvitanic and Karatzas (1992).

Starting with the portfolio choice problem ($\mathcal{P}$), one can define a fictitious problem ($\mathcal{P}^{(\nu)}$), based on a different financial market and without the portfolio constraints. First, define the *support function* of $\mathbf{K}$, $\delta(\cdot)\colon \mathbb{R}^N \to \mathbb{R}\cup\infty$, by

$$\delta(\nu) := \sup_{x\in\mathbf{K}}\big(-\nu^\top x\big). \tag{14}$$

The effective domain of the support function is given by

$$\tilde{\mathbf{K}} := \big\{\nu\colon\ \delta(\nu) < \infty\big\}. \tag{15}$$

Because the constraint set $\mathbf{K}$ is convex and contains zero, the support function is continuous and bounded from below on its effective domain $\tilde{\mathbf{K}}$. Then, one

can define the set $\mathbf{D}$ of $\mathcal{F}_t$-adapted $\mathbb{R}^N$ valued processes to be

$$\mathbf{D} := \left\{ \nu_t,\ 0 \leqslant t \leqslant T\colon\ \nu_t \in \tilde{\mathbf{K}},\ \mathrm{E}_0\left[\int_0^T \delta(\nu_t)\,\mathrm{d}t\right] + \mathrm{E}_0\left[\int_0^T \|\nu_t\|^2\,\mathrm{d}t\right] < \infty \right\}. \tag{16}$$

For each process, $\nu \in \mathbf{D}$, consider a fictitious market, $M^{(\nu)}$, in which the N stocks and the riskfree bond are traded. There are no constraints in the fictitious market. The diffusion matrix of stock returns in $M^{(\nu)}$ is the same as in the original market. However, the riskfree rate and the vector of expected stock returns are different. In particular, the riskfree rate process and the market price of risk in the fictitious market are defined respectively by

$$r_t^{(\nu)} = r_t + \delta(\nu_t), \tag{17a}$$

$$\eta_t^{(\nu)} = \eta_t + \Sigma_{Pt}^{-1}\nu_t, \tag{17b}$$

where $\delta(\nu)$ is the support function defined in (14). Assume that $\eta_t^{(\nu)}$ is square-integrable.

Because the number of Brownian motions, N, is equal to the number of stocks in the financial market described by (11) and the diffusion matrix is nondegenerate, it can be shown that the unconstrained fictitious market is *dynamically complete*. Dynamic completeness would imply the existence of a unique market-price-of-risk process, η_t and a unique *state-price-density* (SPD) process, π_t. $\pi_t(\omega)$ may be interpreted as the price per-unit-probability of \$1 at time t in the event ω occurs. (See Duffie, 1996 for further details.) It so happens that a portfolio optimization problem in complete markets is particularly easy to solve using martingale methods. Following Cox and Huang (1989), the state-price density process $\pi_t^{(\nu)}$ in the fictitious market is given by

$$\pi_t^{(\nu)} = \exp\left(-\int_0^t r_s^{(\nu)}\,\mathrm{d}s - \frac{1}{2}\int_0^t \eta_s^{(\nu)\top}\eta_s^{(\nu)}\,\mathrm{d}s - \int_0^t \eta_s^{(\nu)\top}\,\mathrm{d}B_s\right), \tag{18}$$

and the vector of expected returns is given by

$$\mu_{Pt}^{(\nu)} = r_t^{(\nu)} + \Sigma_{Pt}\,\eta_t^{(\nu)}.$$

The dynamic portfolio choice problem in the fictitious market without position constraints can be equivalently formulated in a static form (e.g., Cox and Huang, 1989; Karatzas and Shreve, 1997, Section 3):

$$V^{(\nu)} := \sup_{\{W_T\}} \mathrm{E}_0\big[U(W_T)\big] \quad \text{subject to} \quad \mathrm{E}_0\big[\pi_T^{(\nu)} W_T\big] \leqslant W_0. \tag{$\mathcal{P}^{(\nu)}$}$$

Once the optimal terminal wealth is computed, it can then be supported by a dynamic trading strategy. Due to its static nature, the problem $(\mathcal{P}^{(\nu)})$ is easy to solve. For example, when the utility function is of the CRRA type with relative risk aversion γ so that $U(W) = W^{1-\gamma}/(1-\gamma)$, the corresponding value function in the fictitious market is given explicitly by

$$V_0^{(\nu)} = \frac{W_0^{1-\gamma}}{1-\gamma} \mathrm{E}_0\left[\pi_T^{(\nu)\,\frac{\gamma-1}{\gamma}}\right]^{\gamma}. \tag{19}$$

It is easy to see that for any admissible choice of $\nu \in \mathbf{D}$, the value function in (19) gives an upper bound for the optimal value function of the original problem. In the fictitious market, the wealth dynamics of the portfolio are given by

$$\mathrm{d}W_t^{(\nu)} = W_t^{(\nu)}\big[\big(r_t^{(\nu)} + \theta_t^{\top}\Sigma_{Pt}\eta_t^{(\nu)}\big)\,\mathrm{d}t + \theta_t^{\top}\Sigma_{Pt}\,\mathrm{d}B_t\big], \tag{20}$$

so that

$$\begin{aligned}\frac{\mathrm{d}W_t^{(\nu)}}{W_t^{(\nu)}} - \frac{\mathrm{d}W_t}{W_t} &= \big[\big(r_t^{(\nu)} - r_t\big) + \theta_t^{\top}\Sigma_{Pt}\big(\eta_t^{(\nu)} - \eta_t\big)\big]\,\mathrm{d}t \\ &= \big[\delta(\nu_t) + \theta_t^{\top}\nu_t\big]\,\mathrm{d}t.\end{aligned}$$

The last expression is nonnegative according to (14) since $\theta_t \in \mathbf{K}$. Thus, $W_t^{(\nu)} \geqslant W_t\ \forall t \in [0,T]$ and

$$V_0^{(\nu)} \geqslant V_0. \tag{21}$$

Results in Cvitanic and Karatzas (1992) and Schroder and Skiadas (2003) imply that if the original optimization problem has a solution, then the upper bound is "tight," i.e., the value function of the fictitious problem $(\mathcal{P}^{(\nu)})$ coincides with the value function of the original problem $(\mathcal{P})$ at an optimally chosen ν^*:

$$V_0^{(\nu^*)} := \inf_{\{\nu\}} V^{(\nu)} = V_0 \tag{22}$$

(see Schroder and Skiadas, 2003, Proposition 3(b) and Theorems 7 and 9). The above equality holds for all times, and not just at time 0, i.e., $V_t^{(\nu^*)} = V_t$. Cvitanic and Karatzas (1992) have shown that the solution to the original problem exists under additional restrictions on the utility function, most importantly that the relative risk aversion does not exceed one. Cuoco (1997) proves a more general existence result, imposing minimal restrictions on the utility function.

3.3 The performance bound

The theoretical duality results of Section 3.2 suggest that one can construct an upper bound on the value function of the portfolio choice problem $(\mathcal{P})$ by computing the value function of any fictitious problem $(\mathcal{P}^{(\nu)})$. The fictitious

market is defined by the process ν_t as in (17). Of course, one can pick any fictitious market from the admissible set $\mathbf{D}$ to compute an upper bound. Such a bound is uninformative if it is too loose. Since the objective is to evaluate a particular candidate policy, one should construct a process $\hat{\nu}_t$ based on such a policy to obtain tighter bounds. The solution to the portfolio choice problem under the fictitious market defined by $\hat{\nu}_t$ then provides a performance bound on the candidate policy.

In order to construct the fictitious market as defined by $\hat{\nu}_t$, Haugh et al. (2005) first use the solution to the dual problem to establish the link between the optimal policy, θ^*, and value function, $V_0 = V_0^{(\nu^*)}$, and the corresponding fictitious asset price processes, as defined by ν^*. Not knowing the optimal portfolio policy and value function, they instead use approximations to obtain the candidate process for ν^*, which is denoted by $\tilde{\nu}$. This candidate process in general does not belong to $\mathbf{D}$ and cannot be used to define a fictitious problem. Instead, one must search for a qualified process $\hat{\nu} \in \mathbf{D}$, which is close to $\tilde{\nu}$. Haugh et al. (2005) then use $\hat{\nu}$ as an approximation to ν^* to define the fictitious problem $(\mathcal{P}^{(\hat{\nu})})$. Since $\hat{\nu} \in \mathbf{D}$, the solution to the corresponding unconstrained problem in $M^{(\hat{\nu})}$ provides a valid performance bound for the candidate policy.

The state-price density process is related via an envelope theorem to the value function by

$$d \ln \pi_t^{(\nu^*)} = d \ln \frac{\partial V_t}{\partial W_t}. \tag{23}$$

In particular, the stochastic part of $d \ln \pi_t^{(\nu^*)}$ is equal to the stochastic part of $d \ln \partial V_t / \partial W_t$. If V_t is smooth, Itô's lemma and Equations (18) and (12) imply that

$$\eta_t^* := \eta_t^{(\nu^*)} = -W_t \left(\frac{\partial^2 V_t / \partial W_t^2}{\partial V_t / \partial W_t} \right) \Sigma_{Pt}^\top \theta_t^* - \left(\frac{\partial V_t}{\partial W_t} \right)^{-1} \Sigma_{Xt}^\top \left(\frac{\partial^2 V_t}{\partial W_t \partial X_t} \right), \tag{24}$$

where θ_t^* denotes the optimal portfolio policy for the original problem. In the special but important case of a CRRA utility function the expression for $\eta_t^{(\nu^*)}$ simplifies. In particular, the first term in (24) simplifies to $\gamma \Sigma_{Pt}^\top \theta_t^*$, where γ is the relative risk aversion coefficient of the utility function, and one only needs to compute the first derivative of the value function with respect to the state variables X_t to evaluate the second term in (24). This simplifies the numerical implementation, since it is generally easier to estimate first-order than second-order partial derivatives of the value function.

Given an approximation to the optimal portfolio policy $\tilde{\theta}_t$, one can compute the corresponding approximation to the value function, $\tilde{V}_t$, defined as the conditional expectation of the utility of terminal wealth, under the portfolio policy $\tilde{\theta}_t$. One can then construct a process $\hat{\nu}$ as an approximation to ν^*, using (24). Approximations to the portfolio policy and value function can be

obtained using a variety of methods, e.g., the ADP method (see Brandt et al., 2005). Haugh et al. (2005) take $\tilde{\theta}_t$ as given and use it to construct an upper bound on the unknown true value function V_0.

Assuming that the approximate value function $\tilde{V}$ is sufficiently smooth, one can replace V_t and θ_t^* in (24) with $\tilde{V}_t$ and $\tilde{\theta}_t$ and obtain

$$\tilde{\eta}_t := \eta_t^{(\tilde{\nu})} = -W_t\left(\frac{\partial^2 \tilde{V}_t/\partial W_t^2}{\partial \tilde{V}_t/\partial W_t}\right)\Sigma_{Pt}^{\top}\tilde{\theta}_t - \left(\frac{\partial \tilde{V}_t}{\partial W_t}\right)^{-1}\Sigma_{Xt}^{\top}\left(\frac{\partial^2 \tilde{V}_t}{\partial W_t \partial X_t}\right). \tag{25}$$

$\tilde{\nu}_t$ is then defined as a solution to (17b).

Obviously, $\tilde{\eta}_t$ is a candidate for the market price of risk in the fictitious market. However, there is no guarantee that $\tilde{\eta}_t$ and the corresponding process $\tilde{\nu}_t$ belong to the feasible set **D** defined by (16). In fact, for many important classes of problems the support function $\delta(\nu_t)$ may be infinite for some values of its argument. Haugh et al. (2005) look for a price-of-risk process $\hat{\eta}_t \in \mathbf{D}$ that is close to $\tilde{\eta}_t$. They choose a Euclidian norm as the measure of distance between the two processes to make the resulting optimization problem tractable.

The requirement that $\hat{\eta}_t \in \mathbf{D}$ is not straightforward to implement computationally. Instead, Haugh et al. (2005) impose a set of tighter uniform bounds,

$$\|\hat{\eta} - \eta\| \leqslant A_1, \tag{26a}$$

$$\delta(\hat{\nu}) \leqslant A_2, \tag{26b}$$

where A_1 and A_2 are positive constants that can be taken to be arbitrarily large. The condition (26a) implies that the process $\hat{\nu}_t$ is square-integrable, since η_t is square integrable and $\|\hat{\eta} - \eta\|^2 = \hat{\nu}^{\top}(\Sigma_P{}^{-1})^{\top}\Sigma_P{}^{-1}\hat{\nu} \geqslant A\|\hat{\nu}\|^2$ for some $A > 0$. Haugh et al. (2005) provide a discussion on the choice of constants A_1 and A_2.

In summary, $\hat{\eta}_t$ and $\hat{\nu}_t$ are defined as a solution of the following problem:

$$\min_{\hat{\nu},\hat{\eta}} \|\hat{\eta} - \tilde{\eta}\|^2, \tag{27}$$

subject to

$$\hat{\eta} = \eta + \Sigma_P^{-1}\hat{\nu}, \tag{28a}$$

$$\delta(\nu) < \infty, \tag{28b}$$

$$\|\hat{\eta} - \eta\| \leqslant A_1, \tag{28c}$$

$$\delta(\hat{\nu}) \leqslant A_2. \tag{28d}$$

The value of $\hat{\eta}_t$ and $\hat{\nu}$ can be computed quite easily for many important classes of portfolio choice problems. The following two examples are taken from Haugh et al. (2005).

Incomplete markets. Assume that only the first L stocks are traded so that the positions in the remaining $N - L$ stocks are restricted to the zero level. In this case the set of feasible portfolio policies is given by

$$\mathbf{K} = \{\theta \mid \theta_i = 0 \text{ for } L < i \leqslant N\} \tag{29}$$

and hence the support function $\delta(\nu)$ is equal to zero if $\nu_i = 0, 1 \leqslant i \leqslant L$ and is infinite otherwise. Thus, as long as $\nu_i = 0, 1 \leqslant i \leqslant L$, the constraint (26b) does not need to be imposed explicitly. To find $\hat{\eta}$ and $\hat{\nu}$, one must solve

$$\min_{\hat{\eta},\hat{\nu}} \|\hat{\eta} - \tilde{\eta}\|^2, \tag{30}$$

subject to

$$\begin{aligned}
&\hat{\eta} = \eta + {\Sigma_P}^{-1}\hat{\nu},\\
&\hat{\nu}_i = 0, \quad 1 \leqslant i \leqslant L,\\
&\|\hat{\eta} - \eta\|^2 \leqslant A_1^2.
\end{aligned}$$

The diffusion matrix Σ_P is lower triangular and so is its inverse. Using this, the solution can be expressed explicitly as

$$\begin{aligned}
&\hat{\eta}_i = \eta_i, \quad 1 \leqslant i \leqslant L,\\
&\hat{\eta}_j = \eta_j + a(\tilde{\eta}_j - \eta_j), \quad L < j \leqslant N,\\
&\hat{\nu} = \Sigma_P(\hat{\eta} - \eta),
\end{aligned}$$

where

$$a = \min\left[1, \left(\frac{A_1^2 - \|\tilde{\eta} - \eta\|_{(L)}^2}{\|\tilde{\eta} - \eta\|^2 - \|\tilde{\eta} - \eta\|_{(L)}^2}\right)^{1/2}\right], \quad \|\tilde{\eta}\|_{(L)}^2 = \sum_{i=1}^{L} \tilde{\eta}_i^2.$$

Incomplete markets, no short sales, and no borrowing. The market is the same as in the previous case, but no short sales and borrowing are allowed. Then the set of admissible portfolios is given by

$$\mathbf{K} = \left\{\theta \mid \theta \geqslant 0,\ 1^\top\theta \leqslant 1,\ \theta_i = 0 \text{ for } L < i \leqslant N\right\}. \tag{31}$$

The support function is given by $\delta(\nu) = \max(0, -\nu_1, \ldots, -\nu_L)$, which is finite for any vector ν. Because in this case $\delta(\nu) \leqslant \|\nu\|$, the relation $\|\nu\| = \|\Sigma_P(\hat{\eta} - \eta)\| \leqslant \|\Sigma_P\|A_1$ implies that as long as A_2 is sufficiently large compared to A_1, one only needs to impose (26a) and (26b) is redundant. We therefore need to solve the following problem:

$$\min_{\hat{\eta},\hat{\nu}} \|\hat{\eta} - \tilde{\eta}\|^2, \tag{32}$$

subject to

$$\begin{aligned}
&\hat{\eta} = \eta + {\Sigma_P}^{-1}\hat{\nu},\\
&\|\hat{\eta} - \eta\|^2 \leqslant A_1^2.
\end{aligned}$$

Then the fictitious market is described by

$$\hat{\eta} = \eta + \min\left(1, \frac{A_1}{\|\tilde{\eta} - \eta\|}\right)(\tilde{\eta} - \eta),$$
$$\hat{\nu} = \Sigma_P(\hat{\eta} - \eta).$$

3.4 Summary of the algorithm

Bounds on the optimal value function can be computed by simulation in several steps.

1. Start with an approximation to the optimal portfolio policy of the original problem and the corresponding approximation to the value function. Both can be obtained using an ADP algorithm, as in Brandt et al. (2005) or Haugh et al. (2005). Alternatively, one may start with an approximate portfolio policy and then estimate the corresponding value function by simulation.
2. Use the approximate portfolio policy and partial derivatives of the approximate value function to construct a process $\tilde{\eta}_t$ according to the explicit formula (25). The process $\tilde{\eta}_t$ is a candidate for the market price of risk in the fictitious market.
3. Construct a process $\hat{\eta}_t$ that is close to $\tilde{\eta}_t$ and satisfies the conditions for the market price risk of a fictitious market in the dual problem. This involves solving the quadratic optimization problem (27).
4. Compute the value function from the static problem $(\mathcal{P}^{(\nu)})$ in the resulting fictitious market defined by the market price of risk process $\hat{\eta}_t$. This can be accomplished efficiently using Monte Carlo simulation. This results in an upper bound on the value function of the original problem.

The lower bound on the value function is obtained by simulating the terminal wealth distribution under the approximate portfolio strategy.

Successful practical implementation of the above algorithm depends on efficient use of simulation methods. For instance, the expectation in (19) cannot be evaluated explicitly and so it has to be estimated by simulating the underlying SDE's. This is a computationally intensive task, particularly when $\tilde{\nu}_t$ cannot be guaranteed in advance to be well-behaved. In such circumstances it is necessary to solve a quadratic optimization problem at each discretization point on each simulated path in order to convert $\eta_t^{(\tilde{\nu})}$ and $\tilde{\nu}_t$ into well-behaved versions that can then be used to construct an upper bound on V_0. (See Haugh et al., 2005 for further details.)

Besides the actual ADP implementation that constructs the initial approximate solution, simulation is also often necessary to approximate the value function and its partial derivatives in (25). This occurs when we wish to evaluate a given portfolio policy, $\tilde{\theta}_t$, but do not know the corresponding value

function, $\tilde{V}_t$. In such circumstances, it seems that it is necessary to simulate the policy, $\tilde{\theta}_t$, in order to approximate the required functions. Once again, this is computationally demanding and seeking efficient simulation techniques for all of these tasks will be an important challenge as we seek to solve ever more complex problems.

Haugh, Kogan and Wu (2005) present several numerical examples illustrating how approximate portfolio policies can be evaluated using duality techniques. While relatively simple, these examples illustrate the potential usefulness of the approach. Haugh et al. (2005) apply the above algorithm to evaluate an ADP solution of the portfolio choice problem in incomplete markets with no-borrowing constraints. Haugh and Jain (2006) use these duality techniques to evaluate other classes of portfolio strategies and to study in further detail the strategies studied by Haugh et al. (2005). Finally, Haugh and Jain (2007) show how path-wise Monte Carlo estimators can be used with the cross-path regression approach to estimate a given portfolio policy's value function as well as its partial derivatives.

3.5 Directions for further research

There are several remaining problems related to the use of duality-based methods in portfolio choice. On the theoretical side, there is room for developing new algorithms designed to tackle more complex and realistic problems. For example, the results summarized above apply to portfolio choice with constraints on proportions of risky assets. However, some important finance problems, such as asset allocation with illiquid/nontradable assets do not fit in this framework. Additional work is required to tackle such problems.

Another important class of problems involve a different kind of market frictions: transaction costs and taxes. Problems of this type are inherently difficult, since the nature of frictions often makes the problem path-dependent and leads to a high-dimensional state space. Some duality results are known for problems with transaction costs (see Rogers, 2003 for a review), while problems with capital gains taxes still pose a challenge. However, note that it is not sufficient to have a dual formulation in order to derive a useful algorithm for computing solution bounds. It is necessary that a high-quality approximation to the optimal dual solution can be recovered easily from an approximate value function. Not all existing dual formulations have such a property and further theoretical developments are necessary to address a broader class of problems.

On the computational side, there is a need for efficient simulation algorithms at various stages of practical implementation of the duality-based algorithms. For instance, motivated by the promising application of importance sampling methods to pricing American options, one could conceivably develop similar techniques to improve performance of portfolio choice algorithms. In particular, in a problem with dynamically complete financial markets and position constraints, approximation errors would tend to accumulate when

portfolio constraints are binding. By sampling more frequently from the "problematic" areas in the state space, one could achieve superior approximation quality.

The above discussion has been centered around the problem of evaluating the quality of approximate solutions. A major open problem both on the theoretical and computational fronts is how to use dual formulations to direct the search for an approximate solution. While a few particularly tractable problems have been tackled by duality methods, there are no efficient general algorithms that could handle multi-dimensional problems with nontrivial dynamics and constraints or frictions. Progress on this front may be challenging, but would significantly expand our ability to address outstanding problems in financial engineering theory and practice.

References

Andersen, L., Broadie, M. (2004). A primal–dual simulation algorithm for pricing multi-dimensional American options. *Management Science* 50 (9), 1222–1234.

Bertsekas, D., Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Black, F., Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–654.

Bolia, N., Glasserman, P., Juneja, S. (2004). Function-approximation-based importance sampling for pricing American options. In: Ingalls, R.G., Rossetti, M.D., Smith, J.S., Peters, B.A. (Eds.), *Proceedings of the 2004 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 604–611.

Boyle, P., Broadie, M., Glasserman, P. (1997). Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control* 21, 1267–1321.

Brandt, M.W., Goyal, A., Santa-Clara, P., Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *Review of Financial Studies* 18, 831–873.

Broadie, M., Glasserman, P. (1997). A stochastic mesh method for pricing high-dimensional American options. *Working paper*, Columbia Business School, Columbia University, New York.

Carriére, J. (1996). Valuation of early-exercise price of options using simulations and non-parametric regressions. *Insurance: Mathematics and Economics* 19, 19–30.

Chen, N., Glasserman, P. (2007). Additive and multiplicative duals for American option pricing. *Finance and Stochastics* 11 (2), 153–179.

Clemént, E., Lamberton, D., Protter, P. (2002). An analysis of a least-squares regression algorithm for American option pricing. *Finance and Stochastics* 6, 449–471.

Cox, J., Huang, C.-F. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory* 49, 33–83.

Cuoco, D. (1997). Optimal consumption and equilibrium prices with portfolio constraints and stochastic income. *Journal of Economic Theory* 72, 33–73.

Cvitanić, J., Karatzas, I. (1992). Convex duality in constrained portfolio optimization. *Annals of Applied Probability* 2, 767–818.

de Farias, D., Van Roy, B. (2003). The linear programming approach to approximate dynamic programming. *Operations Research* 51, 850–865.

de Farias, D., Van Roy, B. (2004). On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research* 29, 462–478.

Davis, M.H.A., Karatzas, I. (1994). A deterministic approach to optimal stopping, with applications. In: Kelly, F. (Ed.), *Probability, Statistics and Optimization: A Tribute to Peter Whittle*. John Wiley & Sons, New York/Chichester, pp. 455–466.

Duffie, D. (1996). *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, NJ.

Glasserman, P. (2004). *Monte-Carlo Methods in Financial Engineering*. Springer, New York.

Glasserman, P., Yu, B. (2004). Pricing American options by simulation: Regression now or regression later. In: Niederreiter, H. (Ed.), *Monte Carlo and Quasi-Monte Carlo Methods*. Springer-Verlag, Berlin.

Han, J. (2005). Dynamic portfolio management: An approximate linear programming approach. *Ph.D. thesis*, Stanford University.

Haugh, M.B., Jain, A. (2006). On the dual approach to portfolio evaluation. *Working paper*, Columbia University.

Haugh, M.B., Jain, A. (2007). Pathwise estimators and cross-path regressions: An application to evaluating portfolio strategies. In: Henderson, S.G., Biller, B., Hsieh, M.H., Shortle, J., Tew, J.D., Barton, R.R. (Eds.), *Proceedings of the 2007 Winter Simulation Conference*. IEEE Press, in press.

Haugh, M.B., Kogan, L. (2004). Pricing American options: A duality approach. *Operations Research* 52 (2), 258–270.

Haugh, M.B., Kogan, L., Wang, J. (2003). Portfolio evaluation: A duality approach. *Operations Research*, in press, available at: http://www.columbia.edu/∼mh2078/Research.html.

Haugh, M.B., Kogan, L., Wu, Z. (2005). Approximate dynamic programming and duality for portfolio optimization. *Working paper*, Columbia University.

He, H., Pearson, N. (1991). Consumption and portfolio policies with incomplete markets and short-sale constraints: The infinite dimensional case. *Journal of Economic Theory* 54 (2), 259–304.

Jamshidian, F. (2003). Minimax optimality of Bermudan and American claims and their Monte-Carlo upper bound approximation. *Working paper*.

Kaniel, R., Tompaidis, S., Zemlianov, A. (2006). Efficient computation of hedging parameters for discretely exercisable options. *Working paper*.

Karatzas, I., Shreve, S.E. (1997). *Methods of Mathematical Finance*. Springer-Verlag, New York.

Karatzas, I., Lehocky, J.P., Shreve, S.E., Xu, G.L. (1991). Martingale and duality methods for utility maximization in an incomplete market. *SIAM Journal Control Optimization* 259, 702–730.

Kloeden, P., Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin.

Kolodko, A., Schoenmakers, J. (2006). Iterative construction of the optimal Bermudan stopping time. *Finance and Stochastics* 10, 27–49.

Liu, J. (1998). Dynamic portfolio choice and risk aversion. *Working paper*, Stanford University, Palo Alto.

Longstaff, F., Schwartz, E. (2001). Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies* 14, 113–147.

Meinshausen, N., Hambly, B.M. (2004). Monte Carlo methods for the valuation of multiple exercise options. *Mathematical Finance* 14, 557–583.

Merton, R.C. (1990). *Continuous-Time Finance*. Basil Blackwell, New York.

Rogers, L.C.G. (2002). Monte-Carlo valuation of American options. *Mathematical Finance* 12 (3), 271–286.

Rogers, L.C.G. (2003). Duality in constrained optimal investment and consumption problems: A synthesis. *Working paper*, Statistical Laboratory, Cambridge University. http://www.statslab.cam.ac.uk/∼chris/.

Schroder, M., Skiadas, C. (2003). Optimal lifetime consumption-portfolio strategies under trading constraints and generalized recursive preferences. *Stochastic Processes and Their Applications* 108, 155–202.

Shreve, S.E., Xu, G.L. (1992a). A duality method for optimal consumption and investment under short-selling prohibition, Part I: General market coefficients. *Annals of Applied Probability* 2, 8–112.

Shreve, S.E., Xu, G.L. (1992b). A duality method for optimal consumption and investment under short-selling prohibition, Part I: Constant market coefficients. *Annals of Applied Probability* 2, 314–328.

Staum, J. (2002). Simulation in financial engineering. In: Yücesan, E., Chen, C.-H., Snowdon, J.L., Charnes, J.M. (Eds.), *Proceedings of the 2002 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1481–1492.

Tsitsiklis, J., Van Roy, B. (2001). Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks* 12 (4), 694–703.

Xu, G.L. (1990). A duality method for optimal consumption and investment under short-selling prohibition. *Ph.D. dissertation*, Carnegie Mellon University.

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15023-4

Chapter 23

Asset Allocation with Multivariate Non-Gaussian Returns

Dilip B. Madan

Department of Finance, University of Maryland, College Park, MD 20742, USA
E-mail: dmadan@rhsmith.umd.edu

Ju-Yi J. Yen

Department of Mathematics, University of Maryland, College Park, MD 20742, USA
E-mail: jjy@math.umd.edu

Abstract

We apply a signal processing technique known as independent component analysis (ICA) to multivariate financial time series. The main idea of ICA is to decompose the observed time series into statistically independent components (ICs). We further assume that the ICs follow the variance gamma (VG) process. The VG process is Brownian motion with drift evaluated at a random time given by a gamma process. We build a portfolio, using closed form expressions, that maximizes expected exponential utility when returns are driven by a mixture of independent factors with VG returns. The performance of this investment is compared with the Markowitz model as a benchmark.

1 Introduction

The relevance of higher moments for investment design has long been recognized in the finance literature and we cite Rubinstein (1973), and Krauss and Litzenberger (1976) from the earlier literature investigating the asset pricing implications of higher moments. More recently we refer to Harvey and Siddique (2000) for the investigation of coskewness in asset pricing. Additionally we note that there appear to be many investment opportunities yielding non-Gaussian returns in the shorter term. This is evidenced, as we will observe, by the ability to construct portfolios with return distributions that in fact display very high levels of kurtosis, a typical measure of non-Gaussianity (Cover and Thomas, 1991). The shorter term perspective is also appropriate for professional investors who can rebalance positions with a greater frequency. Furthermore, we also recognize that there are many ways to construct return

possibilities with the same mean and variance but differing levels of skewness and kurtosis. Investment analysis based on traditional mean–variance preferences (Markowitz, 1952) will not vary the investment across these alternatives, but the presence of skewness preference and kurtosis aversion suggests that the optimal levels of investment should vary across these alternatives.

Apart from these considerations from the economic side, there are implications for asset returns from an information theoretic or signal processing point of view. In this regard we note, importantly, that after centering and scaling a random variable, the shape of its density reflects information content. The Gaussian density among zero mean unit variance densities has the highest uncertainty and least information in the sense of entropy. This leads us to suspect that an inappropriate use of the Gaussian density for designing investment portfolios should result in suboptimal investments. We therefore wish to consider accounting for higher moments in the design of investment portfolios.

The theoretical advantages of accounting for higher moments notwithstanding, we note that the computational burdens of portfolio theory using multivariate distributions with long-tailed marginals are extensive in both the dimensions of model estimation and the subsequent portfolio design. This lies in sharp contrast to the relative ease with which mean–variance analysis may be executed for large portfolios.

Our primary contribution here is to enhance the computational aspects of such a higher moment exposure portfolio theory. In doing this we begin by recognizing that multivariate Gaussian returns with a nonsingular covariance matrix may be viewed as a linear combination of an equal number of independent standard Gaussian random variables. We generalize this perspective and view the vector of asset returns now as a linear combination of an equal number of independent but possibly non-Gaussian random variables.

One may alternatively adopt a factor structure whereby one would allow for fewer non-Gaussian factors and then permit Gaussian idiosyncratic disturbances as well. However, if one permits correlation among the Gaussian idiosyncratic components then the specification is more involved with the number of random variables describing returns now exceeding the number of assets. We leave generalizations in these directions to future research, focusing here on the initial case, that is in keeping with the original Markowitz formulation of keeping the number of independent component random variables equal to the number of assets. Our reformulation relative to Markowitz is therefore, simply, just to permit the independent components to be non-Gaussian.

For the identification of the independent components we recognize from the recent signal processing literature that procedures like principal components are ill-suited in recovering signals when they are present. We have already commented that the presence of signals or informative components presumes the existence of non-Gaussian distributions that when mixed approach Gaussianity. The procedures of independent components analysis (ICA) developed in the signal processing literature seek to maximize metrics of non-Gaussianity

with a view to detecting the original signal or independent random variables. To construct the data on the components we therefore adopt the methods of independent components analysis (ICA) (Hyvärinen et al., 2001) and in particular the fast ICA algorithm (Hyvärinen, 1999).

To describe the probability law of the independent components we use distributions that have proved successful in the derivative pricing literature and work in particular with the variance gamma family introduced by Madan and Seneta (1990) and developed further in Madan et al. (1998). For the estimation we employ the fast Fourier transform methods of Carr et al. (2002). The combined use of fast ICA and the fast Fourier transform renders a fast and efficiently estimable characterization of the multivariate distribution of asset returns. It remains to construct the optimal portfolio.

Given analytically tractable characteristic functions for our independent components makes the use of exponential utility particularly efficient. In this regard we note that with multivariate Gaussian returns it is exponential utility that supports mean variance analysis. Hence, staying close to the support structure of the Markowitz theory suggests that a first analysis should naturally proceed under exponential utility. For exponential utility the tractability of the final portfolio design is accomplished by reduction to univariate investment problems for the optimal investment in each of the independent components. These univariate investment problems are solved for in closed form. The final structure, though computationally more extensive than Markowitz investment, is nonetheless capable of a reasonably rapid execution and we illustrate our methods by a back test comparison with Markowitz investment.

The outline of the paper is as follows. Section 2 briefly presents results for skewness preference and kurtosis aversion in investment design. The distributional models used for the independent components are described in Section 3. In Section 4, we first solve the univariate component investment problem in closed form and then we reduce the multiasset allocation problem to these univariate problems. Section 5 briefly describes the procedures of ICA for identifying the independent components. The results of our illustrative back test are provided in Section 6. Section 7 concludes.

2 Non-Gaussian investment

We begin with fourth-order approximations to a general utility function. The reason for going up to the fourth order is that for investments which agree on mean and variance, the third order recognizes a higher order reward statistic for utilities displaying skewness preference, but no risk statistic is then accounted for once we have conditioned on common variances. The next higher order risk statistic is kurtosis, and to account for both reward and risk we consider fourth-order approximations to utility. Furthermore, we also note that it is the fourth moment or approximations to it that constitute the fundamental measures of non-Gaussianity in the presence of signals or information.

We further emphasize that the primary focus of our study is on shorter horizon returns for which fourth-order approximations are even more appropriate, though analytical tractability is enhanced in our exponential utility by simultaneously considering all the reward (odd) and risk (even) moments. For a more intuitive understanding of the role of higher moments at the risk and reward level we briefly consider just the fourth-order approximation here. The main purpose of this section is to develop a better understanding of the role of skewness preference and kurtosis aversion on investment.

We therefore write for utility $U(x)$ the approximation

$$U(x) \approx U(\mu) + U'(\mu)(x-\mu) + \frac{1}{2}U''(\mu)(x-\mu)^2 + \frac{1}{6}U'''(\mu)(x-\mu)^3 + \frac{1}{24}U''''(\mu)(x-\mu)^4.$$

Define the skewness s and the kurtosis k (Karr, 1993) by

$$s = \frac{E[(x-\mu)^3]}{\sigma^3}, \qquad k = \frac{E[(x-\mu)^4]}{\sigma^4}$$

and write an approximation for the expected utility as

$$E\big[U(x)\big] \approx U(\mu) + \frac{1}{2}U''(\mu)\sigma^2 + \frac{1}{6}U'''(\mu)s\sigma^3 + \frac{1}{24}U''''(\mu)k\sigma^4.$$

We also further approximate

$$U(\mu) \approx U(0) + U'(0)\mu$$

and assume that $U(0) = 0$ and $U'(0) = 1$. We may therefore write

$$E\big[U(x)\big] \approx \mu + \frac{1}{2}U''(\mu)\sigma^2 + \frac{1}{6}U'''(\mu)s\sigma^3 + \frac{1}{24}U''''(\mu)k\sigma^4.$$

For an approximation to exponential utility with risk aversion parameter η, we get

$$E\big[U(x)\big] \approx \mu - \frac{\eta}{2}\sigma^2 + \frac{\eta^2}{6}s\sigma^3 - \frac{\eta^3}{24}k\sigma^4.$$

Now consider the question of investing y dollars in a non-Gaussian return with mean μ, variance σ^2, skewness s, and kurtosis k. The expected utility from this investment on a financed basis with interest rate r is approximately

$$(\mu - r)y - \frac{\eta}{2}\sigma^2 y^2 + \frac{\eta^2}{6}s\sigma^3 y^3 - \frac{\eta^3}{24}k\sigma^4 y^4.$$

The first-order condition for the optimal level of investment is

$$\mu - r - \eta\sigma^2 y + \frac{\eta^2}{2}s\sigma^3 y^2 - \frac{\eta^3}{6}k\sigma^4 y^3 = 0. \tag{2.1}$$

We may rewrite Equation (2.1) as

$$\frac{\mu - r}{y^*} = \eta\sigma^2 - \frac{\eta^2}{2}s\sigma^3 y^* + \frac{\eta^3}{6}k\sigma^4(y^*)^2. \tag{2.2}$$

We can now see that for a positive excess return the optimal y is given by the intersection of a parabola and a hyperbola. This will occur at some positive level for y^*. We also observe the Gaussian or Markowitz result to first order in y^* when skewness is zero.

We may observe that increased excess returns raise the hyperbola and so raise the level of y^*. Also an increase in σ raises the parabola and so leads to a decrease in y^*. An increase in skewness decreases the slope of the parabola at 0 and shifts the intersection with the hyperbola out further thus increasing y^*, while an increase in kurtosis has the opposite effect. For a formal comparative static analysis of these effects we refer the reader to Appendix A.

By way of a comparison to Gaussian investment we note that for distributions with zero skewness, the parabola in y on the right-hand side of (2.2) has a zero slope at zero but then increases in y at a rate dependent of the level of $\eta^3 k\sigma^4$. As a result the optimal investment is consistently below the Gaussian level. With a negative skewness the slope of the parabola is positive at zero and this reduces the investment even further below the Gaussian level. It is only when skewness is positive and the slope of the parabola at zero is negative that investment rises above the Gaussian level. These observations reflect the levels of misinvestment made by Gaussian methods when taking positions in portfolios whose returns have a signal or information theoretic component.

3 Modeling distributions

We need to select models of the distribution of our non-Gaussian component random variables. For this we turn to distributions that have been successfully employed in recent years in modeling risk neutral and statistical asset returns. These distributions are associated with the unit time densities of Lévy processes and we mention the variance gamma (Madan and Seneta, 1990; Madan et al., 1998), the normal inverse Gaussian (Barndorff-Nielsen, 1998), and the generalized hyperbolic model (Eberlein et al., 1998; Eberlein and Prause, 1998). The resulting densities all have analytical characteristic functions and sufficient parametric flexibility to describe varying levels of skewness and kurtosis in addition to the mean and variance. We shall focus attention here on the variance gamma that is particularly tractable in both its characteristic function and the associated Lévy measure.

First we briefly define the variance gamma Lévy process and its use in modeling the stock price distribution at various horizons. The variance gamma process $(X_{VG}(t),\ t \geqslant 0)$ evaluates Brownian motion with drift at a random time change given by a gamma process $(G(t),\ t \geqslant 0)$. Let the Brownian mo-

tion with drift θ and volatility σ be $(Y(t;\sigma,\theta),\ t>0)$, where

$$Y(t;\sigma,\theta)=\theta t+\sigma W(t)$$

and $(W(t), t>0)$ is a standard Brownian motion.

The time change gamma process $(G(t;\nu),\ t>0)$ is a Lévy process whose increments $G(t+h;\nu)-G(t;\nu)=g$ have the gamma density with mean h and variance νh (Rohatgi, 2003) and density $f_h(g)$:

$$f_h(g)=\frac{g^{h/\nu-1}\exp(-g/\nu)}{\nu^{h/\nu}\Gamma(h/\nu)}.$$

Its characteristic function is (Billingsley, 1955):

$$\phi_G(u)=E\big[e^{iug}\big]=\left(\frac{1}{1-iu\nu}\right)^{h/\nu},$$

and for $x>0$, its Lévy density is

$$k_G(x)=\frac{\exp(-x/\nu)}{\nu x}.$$

The variance gamma process $X_{VG}(t;\sigma,\nu,\theta)$ is defined by

$$X_{VG}(t;\sigma,\nu,\theta)=Y\big(G(t;\nu);\sigma,\theta\big)=\theta G(t;\nu)+\sigma W\big(G(t;\nu)\big).$$

The characteristic function of the VG process may be evaluated by conditioning on the gamma process. This is because, given $G(t;\nu)$, $X_{VG}(t)$ is Gaussian. A simple calculation shows that the characteristic function of the variance gamma is

$$\phi_{X_{VG}}(t;u)=E\big[\exp(iuX_{VG})\big]=\left(\frac{1}{1-iu\theta\nu+\sigma^2\nu u^2/2}\right)^{t/\nu}. \tag{3.1}$$

The variance gamma process is a Lévy process with infinitely divisible distributions. Thus the characteristic function of the process may be written in the Lévy–Khintchine form (Sato, 1999), and the Lévy measure K_{VG} is given by (Carr et al., 2002)

$$K_{VG}(x)=\frac{C}{|x|}\exp\left(\frac{G-M}{2}x-\frac{G+M}{2}|x|\right), \tag{3.2}$$

where

$$C=\frac{1}{\nu},$$

$$G=\sqrt{\frac{2}{\sigma^2\nu}+\frac{\theta^2}{\sigma^4}}+\frac{\theta}{\sigma^2},$$

$$M=\sqrt{\frac{2}{\sigma^2\nu}+\frac{\theta^2}{\sigma^4}}-\frac{\theta}{\sigma^2}.$$

The density for the variance gamma process can display both skewness and excess kurtosis. The density is symmetric when $\theta = 0$ and $G = M$, and the kurtosis is $s(1 + \nu)$ in this case. The parameter θ generates skewness and we have a negatively skewed density for $\theta < 0$ and a positively skewed one when $\theta > 0$.

We may accommodate a separate mean by considering the process

$$H(t) = \mu t + \theta\big(G(t) - t\big) + \sigma W\big(G(t)\big) = (\mu - \theta)t + X_{VG}(t)$$

with the characteristic function

$$\phi_{H(t)}(u) = E\big[e^{iuH(t)}\big] = e^{iu(\mu-\theta)t}\phi_{X_{VG}(t)}(u).$$

This gives us a four parameter process capturing the first four moments of the density.

A particularly instructive economic interpretation of the parameters, valid for all three parameter Lévy processes, is obtained by a reparameterization in terms of realized quadratic variation or volatility, a directional premium, and a size premium. One may view the height of the Lévy measure at negative 2% to its height at 2% as a measure of the premium for negative moves over positive ones, either in likelihood or price, depending on whether one is considering the statistical or risk neutral measure. For the variance gamma this is given essentially by $D = (G - M)$. The size premium is the premium of a 2% move over a 4% move and this is given here by $S = (G + M)$. Finally the quadratic variation V is captured by the parameter C, given G and M and is

$$V = C\left(\frac{1}{G^2} + \frac{1}{M^2}\right).$$

One may therefore work equivalently with D, S, and V the direction and size premia and the quadratic variation.

We note furthermore that the Gaussian model is a special case that results on letting the variance of the gamma process approach zero or equivalently by letting the level of kurtosis approach 3.

4 Exponential utility and investment in zero cost VG cash flows

We present in two subsections, first the results for investment in a single risky asset and then the generalization to asset allocation across portfolios.

4.1 A single risky asset

Suppose we invest y dollars in a zero cost cash flow with a VG distribution for the investment horizon of length h with mean $(\mu - r)h$. We may write the zero cost cash flow accessed as X

$$X = (\mu - r)h + \theta(g - 1) + \sigma W(g), \tag{4.1}$$

where g is gamma distributed with unit mean and variance ν, and $W(g)$ is Gaussian with zero mean and variance g. We suppose the VG parameters are for the holding period h as the unit period. We also suppose that μ and r have been adjusted for the length of the period and take this to be unity in what follows.

The final period wealth is

$$W = yX.$$

We employ exponential utility and write

$$U(W) = 1 - \exp(-\eta W), \tag{4.2}$$

where η is the coefficient of risk aversion. The certainty equivalent CE solves

$$E\big(U(W)\big) = 1 - \exp(-\eta CE).$$

The goal of the investment design is the maximization of this expected utility function. The expected utility is

$$E\big(U(W)\big) = E\big(1 - \exp(-\eta W)\big) = 1 - E\big(\exp(-y\eta X)\big). \tag{4.3}$$

To determine the risky asset investment level y it is equivalent to minimize the following expression with respect to y:

$$E\big(\exp(-y\eta X)\big).$$

Theorem 4.1. *Suppose we invest y dollars in a zero cost cash flow with a VG distribution described in Equation (4.1) for the investment horizon of length h. And suppose that we employ the exponential utility function as in Equation (4.2). The optimal solution for the investment is*

$$\tilde{y} = \left(\frac{\theta}{\sigma^2} - \frac{1}{(\mu - r - \theta)\nu}\right) + \operatorname{sign}(\mu - r)\sqrt{\left(\frac{\theta}{\sigma^2} - \frac{1}{(\mu - r - \theta)\nu}\right)^2 + \frac{2(\mu - r)}{(\mu - r - \theta)\nu\sigma^2}},$$

where $\tilde{y} = \eta y$ and η is the risk aversion coefficient.

Proof. See Appendix B. □

When $\mu > r$, y is positive and we have a long position. Likewise for $\mu < r$, y is negative and we have a short position.

4.2 Asset allocation in returns driven by VG components

We take an investment horizon of length h and wish to study the construction of optimal portfolios for investment in a vector of assets whose zero cost

excess returns or financed returns over this period are $R - rh$. Once again we suppose all parameters are adjusted for the time horizon and take this to be unity in what follows.

Let the vector y denote the dollar investment in the collection of assets. We suppose the mean excess return is $\mu - r$ and hence that

$$R - r = \mu - r + x,$$

where x is the zero mean random asset return vector.

Our structural assumption is that there exist a vector of independent zero mean VG random variables s of the same dimension as x and a matrix A such that

$$x = As. \tag{4.4}$$

We noted in the Introduction the reasons for supposing that the number of independent components matches the number of assets. Essentially this is a simplification in keeping with multivariate Markowitz theory where for a full rank covariance matrix the number of independent random variables driving the asset returns matches the number of assets. A factor model specification with a few independent systematic factors and idiosyncratic components increases the number of random variables driving returns to one that is greater than the number of assets. When working with infinitely many assets one may attempt to reduce the exposure to the smaller number of factors as in the Ross (1976) arbitrage pricing theory, but here the focus is on a relatively small number of assets with exposure to all the random variables involved. We proceed, in the first instance here, with as many independent random variables as there are assets. For the interested reader we mention that the factor model approach is studied further for its equilibrium implications in Madan (2005). Procedures for identifying the matrix A will be discussed later when we introduce the methods of independent components analysis.

The probability law of the components s_i is that of

$$s_i = \theta_i(g_i - 1) + \sigma_i W_i(g_i),$$

where the W_is are independent Brownian motions, and the g_i are gamma variates with unit mean and variance ν_i.

Theorem 4.2 below identifies the optimal investment in all the assets for an investor with exponential utility.

Theorem 4.2. *Let the vector y denote the dollar investment in the collection of assets. We suppose the mean excess return is $\mu - r$ and the zero cost excess return is $R - r$, hence that*

$$R - r = \mu - r + x,$$

where x is the zero mean random asset return vector and assume that $E[xx'] = I$. Let

$$x = As$$

and assume the law of s_i is

$$s_i = \theta_i(g_i - 1) + \sigma_i W_i(g_i),$$

where A is the mixing matrix, the W_is are independent Brownian motions, and the g_i are gamma variates with unit mean and variance ν_i. Denote

$$\zeta = A^{-1}\frac{\mu - r}{\eta} - \frac{\theta}{\eta}$$

and

$$y = \frac{1}{\eta}A^{-1}\tilde{y},$$

where $y = (y_1, y_2, \ldots, y_n)'$, $\tilde{y} = (\tilde{y_1}, \tilde{y_2}, \ldots, \tilde{y_n})'$, and η is the risk aversion coefficient. Then the solution of $\tilde{y_i}$, for $i = 1, 2, \ldots, n$, is given by

$$\begin{aligned}\tilde{y_i} &= \frac{|\zeta_i|\theta_i\nu_i - \operatorname{sign}(\zeta_i)\frac{\sigma_i^2}{\eta}}{|\zeta_i|\sigma_i^2\nu_i} \\ &\quad \pm \frac{\sqrt{(|\zeta_i|\theta_i\nu_i - \operatorname{sign}(\zeta_i)\frac{\sigma_i^2}{\eta})^2 + 2(|\zeta_i| + \operatorname{sign}(\zeta_i)\frac{\theta_i}{\eta})|\zeta_i|\sigma_i^2\nu_i}}{|\zeta_i|\sigma_i^2\nu_i} \\ &= \frac{\theta_i}{\sigma_i^2} - \frac{1}{\eta\zeta_i\nu_i} \pm \sqrt{\left(\frac{\theta_i}{\sigma_i^2} - \frac{1}{\eta\zeta_i\nu_i}\right)^2 + 2\frac{\zeta_i + \frac{\theta_i}{\eta}}{\zeta_i\sigma_i^2\nu_i}},\end{aligned} \tag{4.5}$$

and we take the positive or the negative root depending on the sign of $(\zeta_i + \frac{\theta_i}{\eta})$ mean of the implied component exposure.

Proof. See Appendix C. □

5 Identifying the joint distribution of returns

The joint distribution of returns is identified on finding the mixing matrix A and the parameters of the distributions for the independent components. Assuming we have the matrix A we may obtain data on a time series of the independent factors by

$$s_t = A^{-1}x_t,$$

where x_t is the mean corrected series of asset returns. The probability law of the independent components may then be estimated by maximum likelihood applied to the unconditional distribution of the component data s_{it}. We construct a histogram of these observations and then apply the fast Fourier transform to the characteristic function of the variance gamma law to compute

the theoretical probability $f(s)$ of s by

$$f(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iux}\frac{e^{-iu\theta}}{(1 - iu\theta\nu + \frac{\sigma^2\nu}{2}u^2)^{\frac{1}{\nu}}}\,\mathrm{d}u.$$

The log likelihood of the binned data is then computed using n_k observations at the midpoint s_k of the kth interval of the histogram by

$$\mathcal{L}(\sigma, \nu, \theta) = \sum_k n_k \log\big(f(s_k); \sigma, \nu, \theta\big).$$

This likelihood is maximized to get estimates of the parameters $\sigma_i, \nu_i, \theta_i$ for each component on performing N separate univariate estimations.

The identification of the matrix A is done via an application of independent components analysis. One may first perform on the demeaned data for the asset returns a principal components analysis or prewhitening step and construct zero mean, unit variance and uncorrelated random variables $\tilde{x}_t$. For this we construct the covariance matrix C of the demeaned asset returns and write it as

$$C = VDV',$$

where V is the orthogonal matrix of the eigenvectors of C, and D is the diagonal matrix of the corresponding eigenvalues (see Anderson, 1958). We denote $D = \mathrm{diag}(d_1, \ldots, d_n)$, and let $D^{-\frac{1}{2}} = \mathrm{diag}(d_1^{-\frac{1}{2}}, \ldots, d_n^{-\frac{1}{2}})$. Whitening gives us

$$\tilde{x}_t = VD^{-\frac{1}{2}}V'x_t \tag{5.1}$$

$$= Ux_t. \tag{5.2}$$

It follows that

$$E(\tilde{x}\tilde{x}') = I.$$

From Equations (4.4) and (5.1), we then have

$$\tilde{x} = VD^{-\frac{1}{2}}V'x = VD^{-\frac{1}{2}}V'As = \tilde{A}s.$$

The literature on independent components analysis has observed that this whitening procedure is not a good way to recover the original signals of the vector s when there are true signals with non-Gaussian distributions displaying positive information theoretic content. It is observed that in the presence of signals, inherently non-Gaussian, the acting of mixing them via a mixing matrix A takes their distributions via central limit considerations towards the Gaussian density. It is further noted that the whitening step, useful as it is, only determines variables up to a rotation. Any orthonormal transformation of the $\tilde{x}$ is another set of zero mean, unit variance and uncorrelated random variables.

These two observations lead to the formulation of the *ICA* procedure. This is the identification of the right rotation matrix. This is done with a view to undoing the transition to Gaussianity by successively choosing the elements or columns of the rotation matrix W by maximizing a metric of non-Gaussianity and keeping the column on the N dimensional unit sphere and orthogonal to the previous columns. A variety of metrics are suggested and have been studied. These include, excess kurtosis, negentropy, and the expected log cosh of the transformed data. It is suggested that the maximization of expected log cosh is particularly robust and we implemented fast ICA that employs this metric.

We then obtain

$$s = W\tilde{x} = WUx.$$

The matrix A is then constructed as

$$A = (WU)^{-1}.$$

The procedure is described in greater detail in Amari et al. (1996) and Cardoso (1998). For a textbook presentation of independent components analysis the reader is referred to Hyvärinen et al. (2001).

6 Non-Gaussian and Gaussian investment compared

For a back test of the performance of non-Gaussian investment in a set of stocks with Gaussian investment we take daily closing prices on five stocks from January 1990 to May 2004. The five stocks chosen are 3M Company, Boeing Company, IBM, Johnson & Johnson, McDonald's Corp., and Merck & Co. From the time series of prices $p(t)$ for these stocks we construct daily returns

$$R_t = \frac{p_t - p_{t-1}}{p_{t-1}}.$$

We take the returns for the first 1000 days since January 1990 to determine our first positions to be taken in a portfolio of these five stocks. We then move forward one month at a time to get a set of rolling 1000 day time series data for our subsequent positioning that are unwound at month end. This investment is repeated for 125 monthly time periods from January 1990 to May 2004. Thus, we have 125 different 5 by 1000 matrices Y_m, $m = 1, 2, \ldots, 125$ of the set of 5 daily returns.

Performing an ICA analysis on the demeaned data of these matrices yields 125 sets of 5 non-Gaussian independent components on which we estimate the *VG* process by 5 univariate applications done 125 times. To appreciate the degree of non-Gaussianity attained by the *ICA* we present a table with the average level of kurtosis attained for each of the five independent components. We observe that the average kurtosis level for the first factor is five times the Gaussian level and even for the third factor it is double the Gaussian level. It

Table 1.
Performance measures

	VG	Gauss
Sharpe ratio	0.2548	0.2127
CE ($\eta = 0.0005$)	47.6883	0.0230
Gain–loss ratio	2.3909	1.4536

Table 2.
Summary of the kurtosis for the five ICs

	Mean	Minimum	Maximum
1st IC	15.3388	4.2466	54.1112
2nd IC	12.9027	3.9871	49.4759
3rd IC	8.6070	3.9973	41.8942
4th IC	6.3648	3.7159	18.5333
5th IC	5.4536	3.5134	12.0329

can get on occasion to well over 15 times the Gaussian level using just portfolios of five stocks. With more stocks we have obtained much higher levels. We also did such an *ICA* analysis on a Monte Carlo vector of truly Gaussian returns and found no ability to generate any excess kurtosis at all. We therefore conjecture that actual investment returns provide considerable access to informative or kurtotic return scenarios that would be of interest to preferences reflecting a concern for these higher moments.

We study investment design by using Equation (4.5) to compute the vector of dollars, y, invested in each stock under the hypothesis of returns being a linear mixture of independent *VG* processes. We also compute dollar amounts invested for the Gaussian process for comparison (see Elton and Gruber, 1991). At the end of each investment time period, we invest an amount of money y according to our analysis. We look forward in the time series by one month and calculate the cash flow *CF* at the end of the month for each time period. The formula is as follows:

$$CF = y \cdot \left(\frac{p_{t+21} - p_t}{p_t} - \frac{r_t}{12} \right),$$

where p_t is the initial price of the investment, p_{t+21} is the price at the month end unwind, and r_t is the interest rate on the 3-month treasury bill. Note that we use p_{t+21} as the month end, as there are 21 trading days in a month on average. Table 1 presents the three performance measures, the Sharpe ratio, the certainty equivalent (CE), and the gain–loss ratio of both the *VG* and the Gaussian processes (Farrell, 1997). Table 2 displays the summary of the kurtosis of the five independent components.

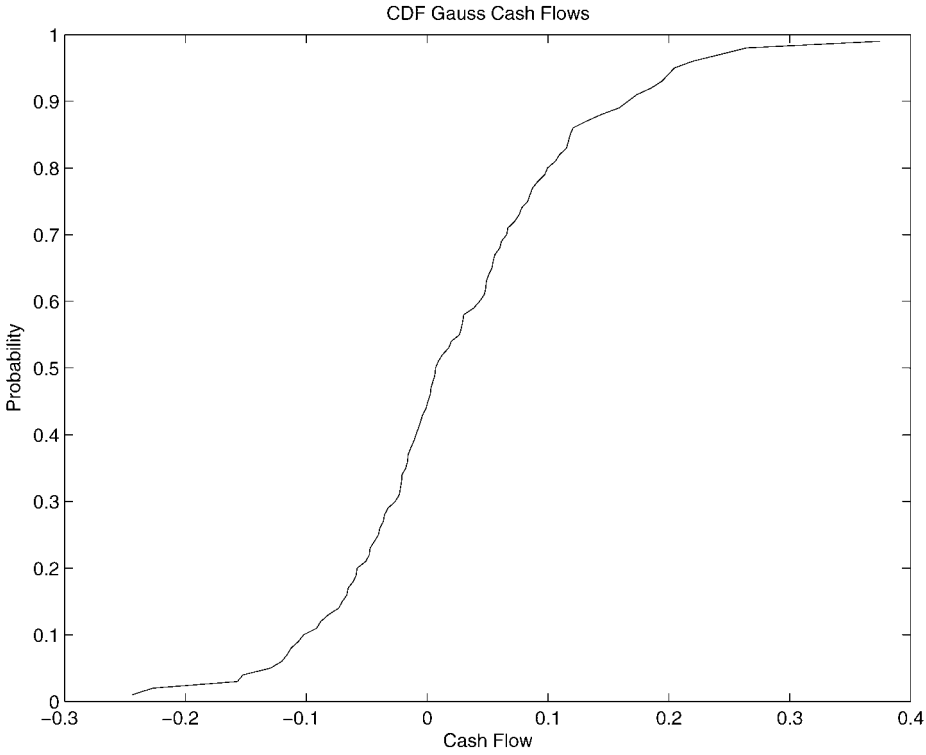


Fig. 1. Cumulative distribution function for Gaussian investment returns.

Figures 1 and 2 plot the cumulated cash flows through the 125 investment time periods of our analysis for the VG and the Gaussian processes.

7 Conclusion

We present and back test an asset allocation procedure that accounts for higher moments in investment returns. The allocation procedure is made computationally efficient by employing independent components analysis and in particular the fast ICA algorithm to identify long-tailed independent components in the vector of asset returns. Univariate methods based on the fast Fourier transform then analyze these components using models popularized in the literature on derivative pricing. The multivariate portfolio allocation problem is then reduced to univariate problems of component investment and the latter are solved for in closed form for exponential utility.

The back test shows that the resulting allocations are substantially different from the Gaussian approach with an associated cumulated cash flow that can outperform Gaussian investment. The combination of fast ICA, the fast Fourier transform and the wide class of Lévy process models now available

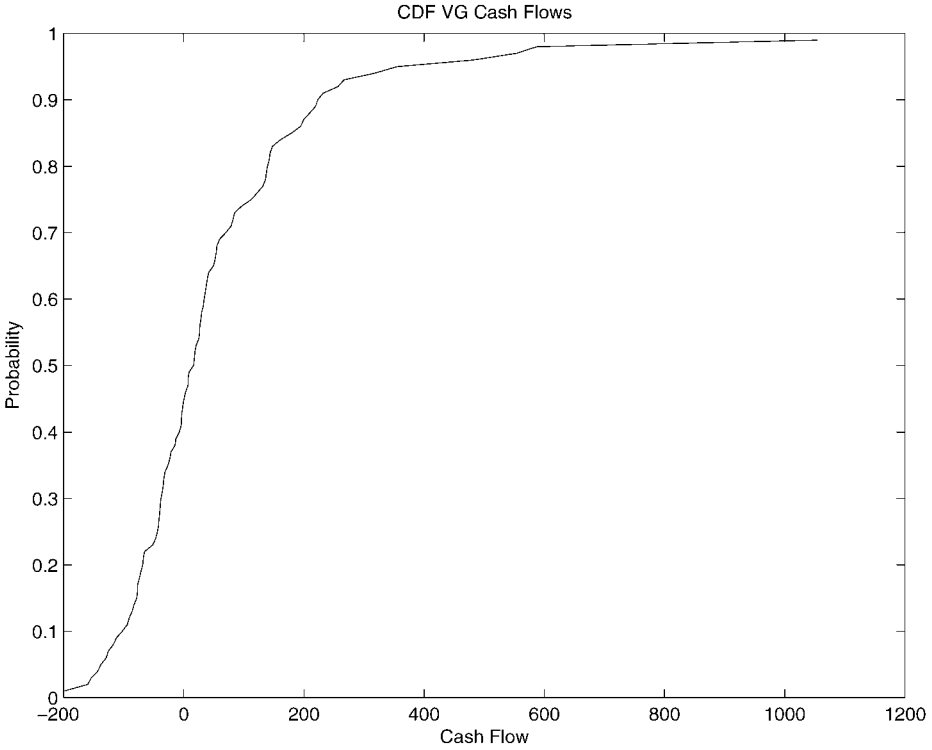


Fig. 2. Cumulative distribution function for non-Gaussian investment returns.

make higher moment asset allocation a particularly attractive area of investment design and future research.

Appendix A. Formal analysis of skewness preference and kurtosis aversion

For a formal analysis of the comparative statistics, we evaluate the differential of the first-order condition with respect to y^*, s, and k as our particular interest. This yields the following equation:

$$\left(\frac{\mu - r}{(y^*)^2} - \frac{\eta^2}{2}s\sigma^3 + \frac{\eta^3}{3}k\sigma^4 y^*\right)\mathrm{d}y^* = \frac{\eta^2}{2}\sigma^3 y^*\,\mathrm{d}s - \frac{\eta^3}{6}\sigma^4(y^*)^2\,\mathrm{d}k.$$

We then have

$$\frac{\mathrm{d}y^*}{\mathrm{d}s} = \frac{\eta^2\sigma^3 y^*}{2}\left(\frac{\mu - r}{(y^*)^2} - \frac{\eta^2}{2}s\sigma^3 + \frac{\eta^3}{3}k\sigma^4 y^*\right)^{-1},$$

$$\frac{\mathrm{d}y^*}{\mathrm{d}k} = -\frac{\eta^3\sigma^4(y^*)^2}{6}\left(\frac{\mu - r}{(y^*)^2} - \frac{\eta^2}{2}s\sigma^3 + \frac{\eta^3}{3}k\sigma^4 y^*\right)^{-1}.$$

The effects of skewness and kurtosis on investment are, respectively, positive and negative, provided the term in the denominator is positive. We may also write that

$$\frac{\mathrm{d}y^*}{\mathrm{d}s} = \frac{\eta^2\sigma^3(y^*)^2}{2}\left(\frac{\mu - r}{y^*} - \frac{\eta^2}{2}s\sigma^3 y^* + \frac{\eta^3}{3}k\sigma^4(y^*)^2\right)^{-1}, \tag{A.1}$$

$$\frac{\mathrm{d}y^*}{\mathrm{d}k} = -\frac{\eta^3\sigma^4(y^*)^3}{6}\left(\frac{\mu - r}{y^*} - \frac{\eta^2}{2}s\sigma^3 y^* + \frac{\eta^3}{3}k\sigma^4(y^*)^2\right)^{-1}. \tag{A.2}$$

Substituting Equation (2.2) into Equations (A.1) and (A.2), we obtain:

$$\frac{\mathrm{d}y^*}{\mathrm{d}s} = \frac{\eta^2\sigma^3(y^*)^2}{2}\left(\eta\sigma^2 - \eta^2 s\sigma^3 y^* + \frac{\eta^3}{2}k\sigma^4(y^*)^2\right)^{-1}, \tag{A.3}$$

$$\frac{\mathrm{d}y^*}{\mathrm{d}k} = -\frac{\eta^3\sigma^4(y^*)^3}{6}\left(\eta\sigma^2 - \eta^2 s\sigma^3 y^* + \frac{\eta^3}{2}k\sigma^4(y^*)^2\right)^{-1}. \tag{A.4}$$

Hence for the signs of Equations (A.3) and (A.4) to be positive and negative, respectively, we need that

$$1 - \eta s\sigma y^* + \frac{\eta^2}{2}k\sigma^2(y^*)^2 > 0.$$

The second derivative of expected utility evaluated at the optimum is

$$-\eta\sigma^2 + \eta^2 s\sigma^3 y^* - \frac{\eta^3}{2}k\sigma^4(y^*)^2.$$

For a maximum, the above expression must be negative. This gives us

$$\eta s\sigma y^* < 1 + \frac{\eta^2}{2}k\sigma^2(y^*)^2$$

or equivalently,

$$1 - \eta s\sigma y^* + \frac{\eta^2}{2}k\sigma^2(y^*)^2 > 0.$$

Hence we observe that investment is positively responsive to skewness and negatively responsive to kurtosis.

Appendix B. Proof of Theorem 4.1

Proof. To find the optimal solution for the investment, our goal is to maximize the expected utility function as in Equation (4.3). It is equivalent to minimizing

$$E\big(\exp(-y\eta X)\big)$$

over y.

$$
\begin{aligned}
&E\big(\exp(-y\eta X)\big)\\
&\quad= \exp\big(-y\eta(\mu - r - \theta)\big)E\left(\exp\left(-\left(y\eta\theta - \frac{y^2\eta^2\sigma^2}{2}\right)g\right)\right)\\
&\quad= \exp\left(-y\eta(\mu - r - \theta) - \frac{1}{\nu}\ln\left(1 + \nu\left(y\eta\theta - \frac{y^2\eta^2\sigma^2}{2}\right)\right)\right).
\end{aligned}
$$

Minimizing the above expression is equivalent to maximizing

$$
z(y) = y\eta(\mu - r - \theta) + \frac{1}{\nu}\ln\left(1 + \nu\left(y\eta\theta - \frac{y^2\eta^2\sigma^2}{2}\right)\right).
$$

Suppose $\alpha, \beta \in R$ and $\alpha < 0 < \beta$. Let

$$
q(y) = 1 + \nu\left(y\eta\theta - \frac{y^2\eta^2\sigma^2}{2}\right),
$$

and $q(\alpha) = q(\beta) = 0$. The function $q(y) > 0$ for $y \in (\alpha, \beta)$ and q is differentiable on (α, β) and continuous on $[\alpha, \beta]$. We have

$$
\begin{aligned}
&z(0) = 0,\\
&\lim_{y\to\alpha^+} z(y) = -\infty,\\
&\lim_{y\to\beta^-} z(y) = -\infty,
\end{aligned}
$$

so that a maximum of $z(y)$ exists on the interval (α, β). The first-order condition with respect to y leads to

$$
z'(y) = \eta(\mu - r - \theta) + \frac{\eta\theta - \eta^2\sigma^2 y}{1 + \nu\eta\theta y - \nu\eta^2\sigma^2 y^2/2}.
$$

Furthermore, assume y_1 and y_2 are two roots for $z'(y) = 0$, and $y_1 < 0$, $y_2 > 0$. That is, $z'(y_1) = z'(y_2) = 0$. Setting $z'(y) = 0$, we obtain

$$
\begin{aligned}
&(\mu - r - \theta)\left(1 + \nu\eta\theta y - \frac{\nu\eta^2\sigma^2}{2}y^2\right) + \theta - \eta\sigma^2 y\\
&\quad= \mu - r + \big((\mu - r - \theta)\nu\theta - \sigma^2\big)\eta y - (\mu - r - \theta)\frac{\nu\eta^2\sigma^2}{2}y^2. \quad \text{(B.1)}
\end{aligned}
$$

Observe that $z'(0) > 0$ if $\mu > r$. We have $z(y_1) < 0$ and $z(y_2) > 0$. According to the mean value theorem, y_2 is the root which gives the optimal solution. Similarly, if $\mu < r$, then $z'(0) < 0$. We have $z(y_1) > 0$ and $z(y_2) < 0$ so that y_1 gives the optimal solution in this condition. Let $\tilde{y} = y\eta$ and solve for this magnitude, noting that y is then $\tilde{y}/\eta$. Hence we rewrite Equation (B.1) as

$$\begin{aligned}
&\tilde{y}^2 - 2\frac{(\mu - r - \theta)\nu\theta - \sigma^2}{(\mu - r - \theta)\nu\sigma^2}\tilde{y} - \frac{2(\mu - r)}{(\mu - r - \theta)\nu\sigma^2} \\
&\quad = \tilde{y}^2 - 2\left(\frac{\theta}{\sigma^2} - \frac{1}{(\mu - r - \theta)\nu}\right)\tilde{y} - \frac{2(\mu - r)}{(\mu - r - \theta)\nu\sigma^2} \\
&\quad = 0.
\end{aligned}$$

Hence we have

$$\begin{aligned}
\tilde{y} = &\left(\frac{\theta}{\sigma^2} - \frac{1}{(\mu - r - \theta)\nu}\right) \\
&+ \operatorname{sign}(\mu - r)\sqrt{\left(\frac{\theta}{\sigma^2} - \frac{1}{(\mu - r - \theta)\nu}\right)^2 + \frac{2(\mu - r)}{(\mu - r - \theta)\nu\sigma^2}}. \qquad \square
\end{aligned}$$

Appendix C. Proof of Theorem 4.2

Proof. We choose the investment vector y to maximize expected exponential utility for a risk aversion coefficient η. The objective is therefore that of maximizing

$$1 - e^{-\eta y'(\mu - r)} E\left[e^{-\eta y' x}\right] = 1 - e^{-\eta y'(\mu - r)} E\left[e^{-\eta y' A s}\right].$$

The expectation is then given by

$$\begin{aligned}
E\left[e^{-\eta y' A s}\right] = \exp\Bigg(&\sum_{i=1}^{n} \eta (y'A)_i \theta_i \\
&- \frac{1}{\nu_i}\ln\left(1 + \theta_i \nu_i \eta (y'A)_i - \frac{\sigma_i^2 \nu_i}{2}\eta^2 (y'A)_i^2\right)\Bigg).
\end{aligned}$$

It follows that the certainty equivalent is

$$\begin{aligned}
CE = &\, y'(\mu - r) + \sum_{i=1}^{n} (-y'A)_i \theta_i \\
&+ \frac{1}{\eta\nu_i}\ln\left(1 + \theta_i \nu_i \eta (y'A)_i - \frac{\sigma_i^2 \nu_i}{2}\eta^2 (y'A)_i^2\right).
\end{aligned}$$

We may write equivalently

$$\begin{aligned}
CE = &\, \eta (y'A)\left(A^{-1}\frac{\mu - r}{\eta} - \frac{\theta}{\eta}\right) \\
&+ \sum_{i=1}^{n} \frac{1}{\eta\nu_i}\ln\left(1 + \theta_i \nu_i \eta (y'A)_i - \frac{\sigma_i^2 \nu_i}{2}\eta^2 (y'A)_i^2\right).
\end{aligned}$$

Now define

$$\tilde{y}' = \eta y' A,$$
$$\zeta = A^{-1}\frac{\mu - r}{\eta} - \frac{\theta}{\eta},$$

and write

$$CE = \sum_{i=1}^{n}\left[\zeta_i\tilde{y}_i + \frac{1}{\eta\nu_i}\ln\left(1 + \theta_i\nu_i\tilde{y}_i - \frac{\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2\right)\right] = \sum_{i=1}^{n}\psi(\tilde{y}_i).$$

We have additive functions in the vector $\tilde{y}_i$ and these may be solved for using univariate methods in closed form. We then determine

$$y = \frac{1}{\eta}A^{-1}\tilde{y}.$$

First observe that the argument of the logarithm is positive only in a finite interval for $\tilde{y}_i$. Hence the *CE* maximization problem has an interior solution for $\tilde{y}_i$.

The first-order condition yields

$$\psi'(\tilde{y}_i) = \zeta_i + \frac{\frac{\theta_i}{\eta} - \frac{\sigma_i^2}{\eta}\tilde{y}_i}{1 + \theta_i\nu_i\tilde{y}_i - \frac{\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2} = 0.$$

It is clear that

$$\psi'(0) = \zeta_i + \frac{\theta_i}{\eta}$$

and the optimal value for $\tilde{y}_i$ is positive when $\psi'(0) > 0$ and negative otherwise.

We may write the condition as

$$|\zeta_i| + \frac{\operatorname{sign}(\zeta_i)(\frac{\theta_i}{\eta} - \frac{\sigma_i^2}{\eta}\tilde{y}_i)}{1 + \theta_i\nu_i\tilde{y}_i - \frac{\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2} = 0.$$

The argument of the logarithm must be positive and so we write

$$|\zeta_i|\left(1 + \theta_i\nu_i\tilde{y}_i - \frac{\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2\right) + \operatorname{sign}(\zeta_i)\left(\frac{\theta_i}{\eta} - \frac{\sigma_i^2}{\eta}\tilde{y}_i\right) = 0.$$

We may rewrite this expression as the quadratic

$$\left(|\zeta_i| + \operatorname{sign}(\zeta_i)\frac{\theta_i}{\eta}\right) + \left(|\zeta_i|\theta_i\nu_i - \operatorname{sign}(\zeta_i)\frac{\sigma_i^2}{\eta}\right)\tilde{y}_i - \frac{|\zeta_i|\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2 = 0,$$

or equivalently that

$$\frac{|\zeta_i|\sigma_i^2\nu_i}{2}\tilde{y}_i{}^2 - \left(|\zeta_i|\theta_i\nu_i - \operatorname{sign}(\zeta_i)\frac{\sigma_i^2}{\eta}\right)\tilde{y}_i - \left(|\zeta_i| + \operatorname{sign}(\zeta_i)\frac{\theta_i}{\eta}\right) = 0.$$

The solution for $\tilde{y}_i$ is given by

$$\tilde{y}_i = \frac{|\zeta_i|\theta_i\nu_i - \text{sign}(\zeta_i)\frac{\sigma_i^2}{\eta}}{|\zeta_i|\sigma_i^2\nu_i} \pm \frac{\sqrt{(|\zeta_i|\theta_i\nu_i - \text{sign}(\zeta_i)\frac{\sigma_i^2}{\eta})^2 + 2(|\zeta_i| + \text{sign}(\zeta_i)\frac{\theta_i}{\eta})|\zeta_i|\sigma_i^2\nu_i}}{|\zeta_i|\sigma_i^2\nu_i}$$

$$= \frac{\theta_i}{\sigma_i^2} - \frac{1}{\eta\zeta_i\nu_i} \pm \sqrt{\left(\frac{\theta_i}{\sigma_i^2} - \frac{1}{\eta\zeta_i\nu_i}\right)^2 + 2\frac{\zeta_i + \frac{\theta_i}{\eta}}{\zeta_i\sigma_i^2\nu_i}}. \qquad \square$$

References

Amari, S.-I., Cichocki, A., Yang, H.H. (1996). A new learning algorithm for blind source separation. *Advances in Neural Information Processing* 8, 757–763.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc., New York.

Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type. *Finance and Stochastics* 2, 41–68.

Billingsley, P. (1955). *Probability and Measure*, third ed. John Wiley & Sons Inc., New York.

Cardoso, J.-F. (1998). Blind signal separation: Statistical principles. *Proceedings of the IEEE* 9, 2009–2025.

Carr, P., Geman, H., Madan, D., Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75 (2), 305–332.

Cover, T.M., Thomas, J.A. (1991). *Elements of Information Theory*. John Wiley & Sons Inc., New York.

Eberlein, E., Keller, U., Prause, K. (1998). New insights into smile, mispricing and value at risk. *Journal of Business* 71, 371–406.

Eberlein, E., Prause, K. (1998). The generalized hyperbolic model: Financial derivatives and risk measures. *FDM Reprint 56*.

Elton, E., Gruber, M. (1991). *Modern Portfolio Theory and Investment Analysis*, fourth ed. John Wiley & Sons Inc., New York.

Farrell Jr., J.L. (1997). *Portfolio Management: Theory and Application*, second ed. McGraw-Hill, New York.

Harvey, C., Siddique, A. (2000). Conditional skewness in asset pricing tests. *Journal of Finance* 55, 1263–1295.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transaction on Neural Networks* 10, 626–634.

Hyvärinen, A., Karhunen, J., Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons Inc., New York.

Karr, A.F. (1993). *Probability*. Springer-Verlag, Berlin.

Krauss, A., Litzenberger, R. (1976). Skewness preference and the valuation of risk assets. *Journal of Finance* 31, 1085–1100.

Madan D. (2005). Equilibrium asset pricing: With non-Gaussian factors and exponential utility. *Working paper*, Robert H. Smith School of Business, University of Maryland.

Madan, D., Seneta, E. (1990). The variance gamma (VG) model for share market returns. *Journal of Business* 63, 511–524.

Madan, D., Carr, P., Chang, E. (1998). The variance gamma process and option pricing. *European Finance Review* 2, 79–105.

Markowitz, H.M. (1952). Portfolio selection. *Journal of Finance* 7, 77–91.
Rohatgi, V.K. (2003). *Statistical Inference*. Dover Publications Inc., New York.
Ross, S.A. (1976). The arbitrage theory of capital asset pricing. *The Journal of Economic Theory* 13, 341–360.
Rubinstein, M. (1973). The fundamental theorem of parameter-preference security valuation. *Journal of Financial and Quantitative Analysis* 8, 61–69.
Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

This page intentionally left blank

J.R. Birge and V. Linetsky (Eds.), *Handbooks in OR & MS, Vol. 15*

DOI: 10.1016/S0927-0507(07)15024-6

Chapter 24

Large Deviation Techniques and Financial Applications

Phelim Boyle

School of Accountancy, University of Waterloo, Ontario, Canada N2L 3G1
E-mail: pboyle@uwaterloo.ca

Shui Feng

Department of Mathematics and Statistics, McMaster University, Ontario, Canada L8S 4K1
E-mail: shuifeng@mcmaster.ca

Weidong Tian

Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada N2L 3G1
E-mail: wdtian@uwaterloo.ca

Abstract

This chapter introduces large deviation techniques and surveys recent applications in finance. Large deviations deal with the theory of rare events and can be used to establish exponential bounds on the probability of such events. If we can establish a so-called large deviation principle for a family of random variables this provides information not only on convergence but also on the speed of convergence. We begin with an introduction to large deviations and outline some of the major results. We discuss a number of applications in finance. These include applications in portfolio management, risk management and Monte Carlo simulations. We also describe some recent work which uses concepts from large deviations to analyze incomplete markets and we illustrate this application with stochastic volatility models.

1 Introduction

The early foundations of mathematical finance were laid down over one hundred years ago by Louis Bachelier in his doctoral thesis. Bachelier created an intuitive model of stochastic processes to represent stock price movements. Over sixty years later Robert Merton used the more formal models of continuous time stochastic processes in his seminal papers on portfolio selection and option pricing. Since then probability theory and stochastic processes have

emerged as essential tools in the study of mathematical finance. Many of the most beautiful and powerful ideas in probability have found natural applications in the field of finance. Large deviation techniques provide a recent example. These techniques have been used in portfolio optimization, large risk estimation and Monte Carlo simulation. In this chapter, we give a brief summary of the large deviation theory and survey some recent applications to the finance discipline. The survey is intended to be representative rather than exhaustive.

The remainder of this chapter is organized as follows. Section 2 introduces the basic ideas of large deviation theory and discusses large deviation principles (LDP). Section 3 examines the application of LDP to portfolio selection and portfolio measurement. Section 4 discusses the application to the tail risk of portfolios. Section 5 deals with applications to Monte Carlo simulation. Section 6 summarizes a recent application of large deviation techniques to pricing in an incomplete market. Section 7 concludes this chapter, and suggests some possible directions where large deviation techniques may prove useful.

2 Large deviation techniques

This section introduces some of the main concepts of large deviation theory. We start with some well-known results in probability and then use these as a reference point. Many probability distributions are characterized by a few parameters that are related to moments of different orders. If the values of these parameters are fixed, then the distribution is completely determined. Thus the study of a probability distribution can sometimes be reduced to the estimation of several parameters. Various approximation mechanisms have been devised to develop estimations of unknown parameters. The law of large numbers (LLN), the central limit theorem (CLT), and the large deviation principle (LDP) are the classical trio of limiting theorems in probability theory that provide the theoretical framework for statistical estimation. The LLN describes the average or mean behavior of a random population. The fluctuation around the average is characterized by the CLT. The theory of large deviations is concerned with the likelihood of unlikely[1] deviations from the average.

We recall that many financial models often involve the solutions of certain stochastic differential equations. Looking at a stochastic differential equation, the drift term corresponds to the average while the random term describes the fluctuation. If the stochastic differential equation represents the dynamics of an asset price, the mean rate of return is related to the LLN while the volatility term is connected to the CLT.

What is LDP then? How can it be used in mathematical finance? Before we answer this it is useful to discuss a simple example.

[1] Hence the name large deviations.

Example 1. Let Z be a standard normal random variable and for $n \geqslant 1$, we define $Z_n = \frac{1}{\sqrt{n}} Z$. Then it is clear that for any $\delta > 0$

$$\lim_{n\to\infty} P\{|Z_n| \geqslant \delta\} = 0, \tag{2.1}$$

and $\sqrt{n} Z_n$ converges (actually they are equal in this case) to Z. These correspond to the weak LLN and the CLT, respectively. By l'Hospital's rule, we get

$$\begin{aligned}\lim_{n\to\infty} P\{|Z_n| \geqslant \delta\} e^{\frac{\delta^2}{2}n} &= 2 \lim_{n\to\infty} e^{\frac{\delta^2}{2}n} \int_{\sqrt{n}\delta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= 2 \lim_{n\to\infty} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{\delta^2}{2}n}}{\frac{\delta^2}{2} e^{-\frac{\delta^2}{2}n}} = \sqrt{\frac{2}{\pi}} \frac{2}{\delta^2}.\end{aligned}$$

Equivalently, we have shown that for large n, $P\{|Z_n| \geqslant \delta\} \approx \sqrt{\frac{2}{\pi}} \frac{2}{\delta^2} e^{-\frac{\delta^2}{2}n}$, i.e., $P\{|Z_n| \geqslant \delta\}$ approaches zero at an exponential decay rate. Clearly this provides more information than the weak LLN and leads to the establishment of the strong LLN. This type of result dealing with the rate of convergence belongs to the subject of large deviations.

We will introduce several general LDP results and illustrate their usefulness through examples. All results will be stated in a form that will be sufficient for our purposes. For more general versions of these and other results on LDP, we recommend (Varadhan, 1984; Freidlin and Wentzell, 1998; Dembo and Zeitouni, 1998) and the references therein.

We now give the formal definition of a large deviation principle. Let E be a complete, separable metric space with metric ρ.

Definition 2.1. A family of probability measures $\{P_\varepsilon\colon\ \varepsilon > 0\}$ defined on the Borel σ-algebra $\mathcal{B}$ of E is said to satisfy a large deviation principle (LDP) with speed $1/\varepsilon$ and rate function $I(\cdot)$ if

$$\text{for any closed set } F,\ \limsup_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{F\} \leqslant -\inf_{x\in F} I(x), \tag{2.2}$$

$$\text{for any open set } G,\ \liminf_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{G\} \geqslant -\inf_{x\in G} I(x), \tag{2.3}$$

$$\text{for any } c \geqslant 0,\ \Phi(c) = \{x \in E\colon\ I(x) \leqslant c\} \text{ is compact}, \tag{2.4}$$

where $\Phi(c)$ is called the level set at level c.

The first two conditions are equivalent to the following statement: for all $A \in \mathcal{B}$,

$$-\inf_{x\in A^0} I(x) \leqslant \liminf_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{A\} \leqslant \limsup_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{A\} \leqslant -\inf_{x\in \overline{A}} I(x). \tag{2.5}$$

In many applications, $A \in \mathcal{B}$ *satisfies* $\inf_{x\in A^0} I(x) = \inf_{x\in \overline{A}} I(x)$. Such an event A is called an I-continuity set. Thus for an I-continuity set A, we have that $\lim_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{A\} = -\inf_{x\in \overline{A}} I(x)$.

The definition will be the same if the set for ε is $\{1/n\colon\ n \geqslant 1\}$. The only difference is that we will write P_n instead of $P_{1/n}$ for $\varepsilon = 1/n$. Going back to Example 1, we have $E = (-\infty, +\infty) = R$, P_n is the law of Z_n, and $\{P_n\colon\ n \geqslant 1\}$ satisfies a LDP with rate function $I(x) = \frac{x^2}{2}$.

If the closed set in (2.2) is replaced by a compact set, then we say the family $\{P_\varepsilon\colon\ \varepsilon > 0\}$ satisfies a weak LDP. To establish a LDP from a weak LDP, one needs to check the following condition which is known as *exponential tightness*: For any $M > 0$, there is a compact set K such that on the complement K^c of K we have

$$\limsup_{\varepsilon\to 0} \varepsilon \log P_\varepsilon\{K^c\} \leqslant -M. \tag{2.6}$$

The relation between weak convergence of probability measures and tightness, is described in Pukhalskii (1991) where a similar relation between LDP and exponential tightness is investigated.

It is clear from the definition that establishing a LDP directly might not be straightforward. Fortunately several general principles (results) have been established, and in many cases the establishment of a LDP is equivalent to verifying the assumptions which lead to these results.

At this stage it is convenient to introduce two important concepts. Given a probability law μ over $\mathcal{B}$ and a random variable Y with distribution law μ, $E = R^d$, then the *logarithmic moment generating function* of law μ is defined as

$$\Lambda(\theta) = \log E\big[e^{\langle\theta, Y\rangle}\big] \quad \text{for all } \theta \in E, \tag{2.7}$$

where $\langle,\rangle$ denotes the usual scalar product in R^d. $\Lambda(.)$ is also called the *cumulant generating function*. The Fenchel–Legendre transformation of $\Lambda(\theta)$ is

$$\Lambda^*(x) := \sup_{\theta\in E}\big\{\langle\theta, x\rangle - \Lambda(\theta)\big\}. \tag{2.8}$$

The first principle is Cramér's theorem. We now describe it in the context of Euclidean space.

Theorem 2.1 (Cramér). *Let $E = R^d$ be the d-dimensional Euclidean space, and $\{Y_n\colon\ n \geqslant 1\}$ be a sequence of i.i.d. random variables. Denote the law of $\frac{1}{n}\sum_{k=1}^n Y_k$ by P_n. Assume that*

$$\Lambda(\theta) = \log E\big[e^{\langle\theta, Y_1\rangle}\big] < \infty \quad \textit{for all } \theta \in R^d.$$

Then the family $\{P_n\colon\ n \geqslant 1\}$ *satisfies a LDP with speed* n *and rate function* $I(x) = \Lambda^*(x)$.

Thus, when dealing with i.i.d. random variables the calculation of the logarithmic moment generating functions is the key in establishing LDP, and its Fenchel–Legendre transformation is the rate function.

Example 2. Let $\{X_n\colon\ n \geqslant 1\}$ be i.i.d. Bernoulli trials with parameter $p \in (0, 1)$. Set $Y_k = \frac{X_k - p}{\sqrt{p(1-p)}}$

$$S_n = \frac{1}{n}\sum_{k=1}^{n} Y_k.$$

Then LLN and CLT imply that S_n converges to zero, $\sqrt{n}S_n$ converges to a standard normal random variable when n becomes large. Let $E = R$, P_n be the law of S_n. The conditions of Cramér's theorem are clearly satisfied. By direct calculation, we have

$$\Lambda(\theta) = \log E\big[e^{\theta Y_1}\big] = \log\Big[(1-p) + pe^{\frac{\theta}{\sqrt{p(1-p)}}}\Big] - \theta\sqrt{\frac{p}{1-p}}.$$

Hence $\{P_n\colon\ n \geqslant 1\}$ satisfies a LDP with rate function

$$I(x) = \begin{cases} \sqrt{p(1-p)}\big[\big(\sqrt{\frac{p}{1-p}} + y\big)\log\big(1 + \sqrt{\frac{1-p}{p}}y\big) + \big(\sqrt{\frac{1-p}{p}} - y\big) \\ \quad \times \log\big(1 - \sqrt{\frac{p}{1-p}}y\big)\big], \quad x \in \big[-\sqrt{\frac{p}{1-p}}, \sqrt{\frac{1-p}{p}}\,\big] \\ \infty, \qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

An infinite dimensional generalization of Cramér's theorem is also available. Here we only mention one particular case – Sanov's theorem.

Let $\{X_k\colon\ k \geqslant 1\}$ be a sequence of i.i.d. random variables in R^d with common distribution μ. For any $n \geqslant 1$, define

$$\eta_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k},$$

where δ_X is the Dirac measure concentrated at X. The sequence of η_n is in the space $M_1(R^d)$ of all probability measures on R^d equipped with the weak topology. It is the empirical distribution of a random sample $X_1, \ldots, X_n$. A well-known result from statistics says that when n becomes large one will recover the true distribution μ from η_n. Clearly $M_1(R^d)$ is an infinite dimensional space. Denote the law of η_n by P_n. Then we have

Theorem 2.2 (Sanov). *The family* $\{P_n\colon\ n \geqslant 1\}$ *satisfies a LDP with speed n and rate function*

$$H(\nu|\mu) = \begin{cases} \int_{R^d} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\, \mathrm{d}\nu, & \text{if } \nu \ll \mu, \\ \infty, & \text{otherwise}. \end{cases}$$

The rate function in Sanov's theorem is also called the relative entropy of ν with respect to μ.

Example 3. Let $\{X_k\colon\ k \geqslant 1\}$ be a sequence of i.i.d. Poisson random variables with parameter 1. Then η_n converges to the Poisson law as n approaches infinity. From Sanov's theorem η_n will stay close only to those probabilities that are supported on nonnegative integers. If ν is supported on nonnegative integers, then

$$H(\nu|\mu) = \sum_{i=0}^{\infty} \nu(i)\big[\log \nu(i) + \log(i!)\big] + 1.$$

The assumption of i.i.d. plays a crucial role in Cramér's theorem. For more general situations one has the following Gärtner–Ellis theorem.

Theorem 2.3 (Gärtner–Ellis). *Let* $E = R^d$, *and* $\{Y_n\colon\ n \geqslant 1\}$ *be a sequence of random variables. Denoting the law of* $\frac{1}{n}\sum_{k=1}^{n} Y_k$ *by* P_n. *Define*

$$\begin{aligned} &\Lambda_n(\theta) = \log E\big[e^{\langle \theta, Y_n\rangle}\big], \\ &\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\theta), \\ &\mathcal{D} = \big\{\theta \in R^d\colon\ \Lambda(\theta) < \infty\big\}, \end{aligned}$$

where in the second equation we assume that the limit exists. Assume further that $\mathcal{D}$ *has a nonempty interior on which* Λ *is differentiable. Also the gradient of* Λ *approaches infinity when* θ *approaches the boundary from the interior.* (Λ *satisfying these conditions is said to be* essentially smooth.) *Then the family* $\{P_n\colon\ n \geqslant 1\}$ *satisfies a LDP with speed n and rate function*

$$I(x) = \sup_{\theta\in R^d} \big\{\langle \theta, x\rangle - \Lambda(\theta)\big\}.$$

Example 4. For any $n \geqslant 1$, let Y_n be a Beta random variable with parameters 1 and n. Then by direct calculation, we have

$$\Lambda_n(\theta) = \log \int_0^1 e^{\theta x} n(1-x)^{n-1}\, \mathrm{d}x,$$

and

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\theta) = \begin{cases} \theta - 1 - \log\theta, & \text{if } \theta > 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Clearly $\mathcal{D} = R$ and Λ is differentiable. Hence by the Gärtner–Ellis theorem, the law of Y_n satisfies a LDP with speed n and rate function

$$I(x) = \begin{cases} -\log(1-x), & \text{if } x \in [0,1], \\ \infty, & \text{else.} \end{cases}$$

From all these results, we can see the importance of the form of the rate function. If one knows the candidate for the rate function, then what is required is to verify the related inequalities. A well-known result of Varadhan provides a method to guess the rate function.

Theorem 2.4 (Varadhan's lemma). *Let $E = R^d$, and $\{P_n\colon\ n \geqslant 1\}$ be a family of probabilities satisfying a LDP with speed n and rate function I. Then for any bounded continuous function f on R^d,*

$$\lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\left[e^{nf(x)}\right] = \sup_{x\in R^d} \left\{f(x) - I(x)\right\}.$$

Hence by calculating the left-hand side in the above equation, we will get an idea of the form of the rate function.

Another important large deviation technique is the following so-called contraction principle.

Theorem 2.5 (Contraction principle). *Let E, F be complete, separable spaces, and h be a continuous function from E to F. For a family of probability measures $\{P_n\colon\ n \geqslant 1\}$ on E, we denote $\{Q_n\colon\ n \geqslant 1\}$ the family of probability measures on F such that $Q_n = P_n \circ h^{-1}$. If $\{P_n\colon\ n \geqslant 1\}$ satisfies a LDP with speed n and rate function I, then $\{Q_n\colon\ n \geqslant 1\}$ also satisfies a LDP with speed n and rate function*

$$I'(y) = \inf\left\{I(x)\colon\ y = h(x)\right\}.$$

As one important implication of this "Contraction Principle," a LDP can be transformed by a continuous function from one space to another.

If the random variables are replaced by stochastic processes, we are then dealing with large deviations at the path level. The stochastic processes could be an independent system, a weakly interacting system (McKean–Vlasov limit), a strongly interacting system (hydrodynamic limit), measure-valued process, or a random perturbation of a deterministic dynamical system. In this section we focus on the last class: random perturbation of a deterministic dynamical system. These types of large deviations are considered in the so-called Freidlin–Wentzell theory (Freidlin and Wentzell, 1998).

For any fixed $T > 0$, let $E = C([0, T], R^d)$ be the collection of all R^d valued continuous functions on $[0, T]$. The topology on E generated by the following metric:

$$\rho\big(\phi(\cdot), \varphi(\cdot)\big) = \sup_{t\in[0,T]} \big|\phi(t) - \varphi(t)\big|$$

is complete and separable.

Consider the ordinary differential equation (ODE)

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = b\big(X(t)\big), \qquad X(0) = x \in R^d. \tag{2.9}$$

For any $\varepsilon > 0$, a random perturbation of the ODE (2.9) is the following stochastic differential equation:

$$\mathrm{d}X^\varepsilon(t) = b\big(X^\varepsilon(t)\big)\,\mathrm{d}t + \sqrt{\varepsilon}\sigma\big(X^\varepsilon(t)\big)\,\mathrm{d}B(t), \quad X^\varepsilon(0) = x, \tag{2.10}$$

where $B(t)$ is a d-dimensional Brownian motion, $\sigma(X)$ is a nonnegative definite $d \times d$ matrix.

Let P_ε denote the law of the process $X^\varepsilon(\cdot)$ on E. Then we have

Theorem 2.6 (Freidlin–Wentzell). *Assume that all elements of $b(X)$ and $\sigma(X)$ are bounded, Lipschitz continuous, and $\sigma(X)$ is positive definite. Then the family $\{P_\varepsilon\colon\ \varepsilon > 0\}$ satisfies a LDP with speed $1/\varepsilon$ and rate function (or action functional)*

$$I\big(\phi(\cdot)\big) = \begin{cases} \frac{1}{2}\int_0^T \langle(\dot\phi(t) - b(\phi(t))), D^{-1}(\phi(t))(\dot\phi(t) - b(\phi(t)))\rangle\,\mathrm{d}t, \\ \quad \textit{if } \ \phi \in \mathcal{H}_x, \\ \infty, \quad \textit{elsewhere}, \end{cases}$$

where $D(X) = \sigma(X)\sigma^*(X)$, and H_x is a subset of $C([0, T], R^d)$ containing all absolutely continuous functions starting at x.

Example 5. For any $\varepsilon > 0$, let $d = 1$ and $X^\varepsilon(t)$ be the solution of the SDE

$$\mathrm{d}X^\varepsilon(t) = \sqrt{\varepsilon}\,\mathrm{d}B(t), \quad X^\varepsilon(0) = x.$$

Thus we have $b(X) = 1$, $\sigma(X) = 1$, and all conditions in Theorem 2.6 are satisfied. Therefore, the law of $X^\varepsilon(\cdot)$ satisfies a LDP with speed ε and rate function

$$I\big(\phi(\cdot)\big) = \begin{cases} \frac{1}{2}\int_0^T |\dot\phi(t)|^2\,\mathrm{d}t, & \text{if } \phi \in \mathcal{H}_x, \\ \infty, & \text{elsewhere.} \end{cases}$$

The result in this example is called one-dimensional Schilder theorem (Schilder, 1966). Historically, the Schilder theorem was established earlier than the Freidlin–Wentzell theory even though we derive it as an application of the latter.

Next we present an example that is popular in mathematical finance.

Example 6. Let B_t be a standard one-dimensional Brownian motion, and $Y^\varepsilon(t) = e^{\sqrt{\varepsilon}B_t}$. In other words, $Y^\varepsilon(t)$ is a geometric Brownian motion. If we write $f(X) = e^X$, then clearly $Y^\varepsilon(t) = f(X^\varepsilon(t))$ where $X^\varepsilon(t)$ is the process in Example 5 with starting point 0. Then the contraction principle combined with Example 5 implies that the law of $Y^\varepsilon(t)$ satisfies a LDP with speed ε and rate function

$$I'\big(\varphi(\cdot)\big) = I\big(\log\big(\varphi(\cdot)\big)\big).$$

In the above Theorem 2.6, the positive definite requirement of $\sigma(X)$ can be relaxed (see Dembo and Zeitouni, 1998). But in some financial models such as the Cox–Ingersoll–Ross (CIR) model the Lipschitz condition does not hold.

Recently in the study of the large deviations for Fleming–Viot process, Dawson and Feng (1998, 2001), Feng and Xiong (2002) established LDPs for the following processes.

Let

$$\mathrm{d}x_t^\varepsilon = \big(a + bx_t^\varepsilon\big)\,\mathrm{d}t + \sqrt{\varepsilon x_t^\varepsilon}\,\mathrm{d}B_t, \quad x_0^\varepsilon = c \geqslant 0,\ a \geqslant 0, \tag{2.11}$$

$$\mathrm{d}y_t^\varepsilon = \theta\big(p - y_t^\varepsilon\big)\,\mathrm{d}t + \sqrt{\varepsilon y_t^\varepsilon\big(1 - y_t^\varepsilon\big)}\,\mathrm{d}B_t,$$
$$y_0^\varepsilon = d,\ \theta > 0,\ p, d \in [0, 1], \tag{2.12}$$

and P_ε, Q_ε be the laws of x_t^ε and y_t^ε, respectively. Here is their result.

Theorem 2.7. *The families* $\{P_\varepsilon\colon\ \varepsilon > 0\}$ *and* $\{Q_\varepsilon\colon\ \varepsilon > 0\}$ *satisfy LDPs with speed* $1/\varepsilon$ *and respective rate function* $I_1(\cdot)$ *and* $I_2(\cdot)$. *For any absolutely continuous path* $\phi(\cdot)$ *satisfying* $\inf\{\phi(t)\colon\ t \in [0, T]\} > 0$ *and* $\phi(0) = c$ *we have*

$$I_1\big(\phi(\cdot)\big) = \frac{1}{2}\int_0^T \frac{(\dot\phi(t) - a - b\phi(t))^2}{\phi(t)}\,\mathrm{d}t.$$

For any absolutely continuous path $\phi(\cdot)$ *satisfying*

$$0 < \inf\big\{\phi(t)\colon\ t \in [0, T]\big\} \leqslant \sup\big\{\phi(t)\colon\ t \in [0, T]\big\} < 1, \quad \phi(0) = d,$$

we have

$$I_2\big(\phi(\cdot)\big) = \frac{1}{2}\int_0^T \frac{(\dot\phi(t) - \theta(p - b\phi(t)))^2}{\phi(t)(1 - \phi(t))}\,\mathrm{d}t.$$

This completes our brief overview of large deviations theory. We are now ready to discuss applications in mathematical finance.

3 Applications to portfolio management

The first application deals with the problem of portfolio management. We start with some of the institutional background. Portfolio fund managers need

clear objectives and scientific procedures to manage their portfolios to attain them. Both managers and investors also need measures of portfolio performance. In investment history, the growth-maximum or Kelly investment strategy might be the first strategy. (It dates back to Bernoulli in 17th century.) In modern times, the Markowitz–Sharpe's approach[2] provides optimal portfolios in a mean–variance framework. Merton extended this static framework to a continuous-time dynamic framework. More recently large deviation techniques have been employed to provide an alternative approach both for portfolio selection and the measurement of portfolio performance.

3.1 Portfolio selection criterion

First we consider the portfolio selection problem. Let $(\Omega, \mathcal{F}, P)$ denote an underlying probability space with information structure $\mathcal{F} = (\mathcal{F}_t)_{t\geqslant 0}$. For simplicity we consider a market with just two assets: a risky asset S and a riskless asset B. We assume a constant interest rate r. [For a discussion on a more general framework we refer to Stutzer (2003), and Pham (2003).] The price of S follows a diffusion process

$$\frac{\mathrm{d}S}{S} = \mu(S, t)\,\mathrm{d}t + \sigma(S, t)\,\mathrm{d}W_t. \tag{3.1}$$

Let X_t^π be the fund manager's wealth at time t and π_t the proportion of wealth invested in the risky asset. From the self-financing condition, we have

$$\mathrm{d}X_t^\pi = X_t^\pi\big[\big(r + (\mu - r)\pi_t\big)\,\mathrm{d}t + \pi_t\sigma\,\mathrm{d}W_t\big].$$

We consider those admissible strategies (π_t) such that $X_t^\pi \geqslant 0$ to get rid of *doubling* arbitrage strategies. Given an admissible strategy (π_t), the log rate of return over the period $[0, T]$ is

$$R_T^\pi = \frac{\log(X_T^\pi/x)}{T}, \tag{3.2}$$

where $x = X_0$ is the initial amount.

Shortfall probability strategies have been proposed in Browne (1999) and Föllmer and Leukert (1999). In this case the criterion is to maximize the probability of beating some benchmark return, say c, over a fixed investment horizon T. To illustrate how LDP can be applied to this problem we consider first a very simple special case and then deal with more general situations.

We first assume that π_t is a *constant percentage strategy* for all t. It means that $\pi_t = \pi$, a constant percentage, for all t. The investment period $[0, T]$ is divided into discrete periods (such as annual, monthly or daily) and $R_{\pi,t}$ denotes the

[2] It can be formulated as maximizing the Sharpe ratio.

log gross rate of return between times $t-1$ and t. Then

$$R_T^\pi = \frac{\sum_{t=1}^T R_{\pi,t}}{T}, \quad R_{\pi,t} = \log\bigl(X_t^\pi / X_{t-1}^\pi\bigr).$$

To simplify the problem further we first assume that both μ, r are constants. In this case, the $R_{\pi,t}$ have an independent identical distribution (i.i.d.). Furthermore this distribution is normal distribution with

$$R_{\pi,t} \sim \mathcal{N}\bigl(E[R_{\pi,t}], \operatorname{Var}[R_{\pi,t}]\bigr),$$

where

$$E[R_{\pi,t}] = r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2},$$

and $\operatorname{Var}[R_{\pi,t}] = \pi^2\sigma^2$. Hence by the LLN,

$$R_T^\pi \to r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2}. \tag{3.3}$$

We see that

$$\Lambda(\theta) := \log E\bigl[e^{\theta R_{\pi,t}}\bigr] = \theta\left[r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2}\right] + \frac{1}{2}\theta^2\pi^2\sigma^2. \tag{3.4}$$

Thus by Cramér theorem,

$$\liminf_{T\to\infty} \frac{1}{T}\log P\bigl(X_T^\pi \geqslant xe^{cT}\bigr) = -\inf_{x\geqslant c} I(x;\pi), \tag{3.5}$$

where

$$I(x;\pi) = \frac{(x - (r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2}))^2}{2\theta^2\pi^2\sigma^2}.$$

Hence

$$\begin{aligned}\liminf_{T\to\infty} \frac{1}{T}\log P\bigl(X_T^\pi \geqslant xe^{cT}\bigr) &= -I(c)\\ &:= -\frac{(c - (r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2}))^2}{2\theta^2\pi^2\sigma^2}.\end{aligned} \tag{3.6}$$

Therefore

$$P\bigl(X_T^\pi \geqslant xe^{cT}\bigr) \simeq \exp\left(-\frac{(c - (r + \pi(\mu - r) - \frac{\pi^2\sigma^2}{2}))^2}{2\theta^2\pi^2\sigma^2}T\right) \tag{3.7}$$

where the decay rate of the probability $P(X_T^\pi \geqslant xe^{cT})$ is $I(c,\pi)$.

The *criterion* proposed by LDP is to find the admissible strategy (π_t) to maximize (*the probability of beating the benchmark in an asymptotic sense*)

$$\liminf_{T\to\infty} \frac{1}{T} \log P\big(X_T^{\pi} \geqslant xe^{cT}\big)$$

or equivalently, to minimize the decay rate. One justification for this selection criterion is that the investment horizon can be uncertain and quite long for mutual funds and pension funds.

If we focus on all constant percentage strategies when dealing with this criterion, we can verify that the optimal constant percentage strategy π_{LDP} with minimal decay rate is independent of T, by using last formula. This strategy provides an *optimal constant percentage strategy* with an optimal asymptotic growth rate. Moreover

$$\pi_{\mathrm{LDP}} = \sqrt{\frac{2(c-r)}{\sigma^2}}. \tag{3.8}$$

Let us consider *all* possible admissible strategies. We will see very shortly how LDP is a powerful tool to deal with some technical problems. Suppose the investment horizon T is fixed and we want to maximize (*shortfall strategy*)

$$\mathrm{Max}\, P\big(X_T^{\pi} \geqslant xe^{cT}\big)$$

over all admissible strategies. The optimal percentage $\pi_{t,T}$ is (see Browne, 1999 for details)

$$\pi_{t,T} = \frac{1}{\sigma\sqrt{T-t}} \frac{n(v_t)}{N(v_t)}, \tag{3.9}$$

where

$$v_t = \frac{\log(S_t/K) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}$$

and K is given implicitly by

$$\frac{\log(S_0/K) + (r - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} = N^{-1}\big(e^{-(c-r)T}\big),$$

where $N(.)$ is the cumulative normal probability function. However, for general asset price processes it is difficult to obtain explicit expressions for the shortfall probability. But LDP, especially the Gärtner–Ellis theorem is powerful enough to handle the maximum asymptotic probability strategy.

Actually, a remarkable result, due to Pham (2003), implies that, the above constant percentage strategy π_{LDP} is the *optimal* strategy based on the following criterion:

$$\mathrm{Max} \liminf_{T\to\infty} \frac{1}{T} \log P\big(X_T^{\pi} \geqslant xe^{cT}\big)$$

over *all admissible strategies* (π_t).

We now state a result in a more general continuous-time framework.

Define

$$\Lambda(\lambda, \pi) = \liminf_{T\to\infty} \frac{1}{T} \log E\big[e^{\lambda T R_T^{\pi}}\big] \tag{3.10}$$

and

$$\Lambda(\lambda) = \mathrm{Max}_{\pi}\, \Lambda(\lambda, \pi) \tag{3.11}$$

over all possible admissible strategies. Define

$$J(c) = \mathrm{Sup}_{\pi} \liminf_{T\to\infty} \frac{1}{T} \log P\big[R_T^{\pi} \geqslant c\big]. \tag{3.12}$$

The next theorem of Pham (2003) express the dual-relationship between the maximum asymptotic strategy and the ergodic risk-sensitive control problem for $\Lambda(\lambda)$ for any fixed λ.

Theorem 3.1. *Suppose that there exists* $\bar{\lambda} \geqslant 0$ *such that for all* $\lambda \in [0, \bar{\lambda})$*, there exists a solution* $\overline{\pi}(\lambda)$ *to the optimal problem* $\Lambda(\lambda)$ *where the appropriate limit exists, i.e.,*

$$\Lambda(\lambda) = \liminf_{T\to\infty} \frac{1}{T} \log E\big[e^{\lambda T R_T^{\overline{\pi}(\lambda)}}\big]. \tag{3.13}$$

Suppose also that $\Lambda(\lambda)$ *is continuously differential on* $[0, \bar{\lambda})$ *with limit* $\lim_{\lambda\to\bar{\lambda}} \Lambda'(\lambda) = \infty$*. Then we have*

$$J(c) = -\mathrm{Sup}_{\lambda\in[0,\bar{\lambda})}\big[\lambda c - \Lambda(\lambda)\big], \quad \forall c \in R. \tag{3.14}$$

The optimal logarithmic moment generating function $\Lambda(\lambda)$ is easy to obtain explicitly. For instance, assume that $\log S$ satisfies a mean reversion process as follows:

$$\frac{\mathrm{d}S}{S} = [a - \kappa \log S]\,\mathrm{d}t + \sigma\,\mathrm{d}W. \tag{3.15}$$

Then

$$\Lambda(\lambda) = \frac{\kappa}{2}\big[1 - \sqrt{1-\lambda}\,\big] + \frac{\lambda}{2}\left(\frac{a-r}{\sigma}\right)^2. \tag{3.16}$$

Write $\bar{c} = \frac{\lambda}{2}(\frac{a-r}{\sigma})^2 + \frac{\kappa}{4}$. By straightforward calculation,

$$J(c) = \begin{cases} -\frac{(c-\bar{c})^2}{c-\bar{c}+\frac{\kappa}{4}}, & \text{if } c \geqslant \bar{c}, \\ 0, & \text{if } c < \bar{c}. \end{cases}$$

The sequence of *nearly* optimal portfolios for $c \geqslant \bar{c}$ is

$$\pi_t^n = -\left[4\left(c + \frac{1}{n} - \bar{c}\right) + \kappa\right] \log S_t + \frac{a-r}{\sigma}. \tag{3.17}$$

The optimal portfolio for $c < \bar{c}$ is explicitly given by

$$\pi_t = -\kappa \log S_t + \frac{a-r}{\sigma}. \tag{3.18}$$

Remark. It is interesting to compare this portfolio selection criterion with other portfolio selection criteria. Given a constant percentage strategy π, $R_{\pi,t}$ is not i.i.d., in general. The log moment generating function of the time average of partial sums is

$$\begin{aligned}\Lambda(\lambda;\pi) &= \liminf_{T\to\infty} \frac{1}{T} \log E\big[e^{\lambda T R_T^\pi}\big] \\ &= \liminf_{T\to\infty} \frac{1}{T} \log E\left[\left(\frac{X_T^\pi}{x}\right)^\lambda\right].\end{aligned} \tag{3.19}$$

Assuming standard technical assumptions in the Gärtner–Ellis theorem, we have

$$P\big(X_T^\pi \geqslant x e^{cT}\big) \simeq \exp\big(-I(c;\pi)T\big), \tag{3.20}$$

where $I(x;\pi)$ is the Fenchel–Legendre transform of $\Lambda(\lambda;\pi)$. Therefore, the maximum asymptotic strategy is to find the constant percentage π to maximize $I(c;\pi)$ which is equivalent to maximizing

$$\begin{aligned}\mathrm{Max}_\pi\, I(c;\pi) &= \mathrm{Max}_\pi\, \mathrm{Max}_\lambda \left\{\lambda c - \lim_{T\to\infty} \frac{1}{T} \log E\left[\left(\frac{X_T^\pi}{x}\right)^\lambda\right]\right\} \\ &= \mathrm{Max}_\pi\, \mathrm{Max}_\lambda \lim_{T\to\infty} \frac{1}{T} \log E\left[-\left(\frac{X_T^\pi}{x e^{cT}}\right)^\lambda\right].\end{aligned}$$

There are both similarities and differences between this objective and a conventional power utility criterion. In the power utility case the relative risk aversion parameter is given. But in this asymptotic strategy, both the optimal strategy and the relative risk aversion parameter are determined simultaneously. Another difference is that the ratio of wealth to the benchmark xe^{cT} is involved. This feature appears in the shortfall probability strategy as well, because the probability measure $P(X_T^\pi \geqslant xe^{cT})$ is involved as well. Fortunately, unlike the maximum shortfall probability strategy, the optimal strategy is much easier to compute in this framework.

3.2 Portfolio performance index

We now turn to a discussion of the construction of performance indices. As Roll (1992) puts it, today's professional money manager is often judged by total return performance relative to a prespecified benchmark. Therefore, the design of a suitable portfolio performance index is important from a practical viewpoint. We now explain how Large Deviation theory can assist us in this task.

Let $R_p - R_b$ denote a portfolio p''s return over and above a benchmark, b. The natural generalization of mean–variance efficiency relative to a benchmark is Roll's *tracking error variance* (*TEV*)-*efficiency*, resulting from minimization of the tracking error variance $\mathrm{Var}[R_p - R_b]$ subject to a constraint on the desired size of $E[R_p - R_b] > 0$. Specifically, the most common scalar performance measure consistent with TEV efficiency is the *information ratio*, defined as

$$\frac{E[R_p - R_b]}{\sqrt{\mathrm{Var}[R_p - R_b]}}. \tag{3.21}$$

If R_b has the riskfree return, the information ratio becomes well-known Sharpe ratio (Sharpe, 1966). This information ratio is only defined in a single-period. Here we provide a natural generalization of this concept to the multi-period setting (or even the continuous-time framework).

Given a benchmark portfolio b, assume the time periods are $\{0, 1, \ldots, T\}$. Write

$$W_T^p = W_0 \prod_{t=1}^{T} R_{pt},$$

where R_{pt} denotes the random gross return from the strategy p between times $t-1$ and time t. Similarly we define

$$W_T^b = W_0 \prod_{t=1}^{T} R_{bt}.$$

The *outperformance event* is the event such that

$$\frac{\log W_T^p - \log W_T^b}{T} \equiv \frac{1}{T} \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt}) > 0. \tag{3.22}$$

A natural measure is to employ a *rank ordering* of the probabilities

$$\mathrm{Prob}\left[\frac{1}{T} \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt}) > 0 \right]. \tag{3.23}$$

In this ranking, the investment horizon T is involved. Because of the difficulty in determining the precise length of an investor's horizon (when one exists), and because short horizon investors may have different portfolio rankings than the long horizon investors, one can use the asymptotic approach that $T \to \infty$. That is rank the portfolio strategies p for which

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} (\log R_{pt} - \log R_{bt}) > 0. \tag{3.24}$$

As an example, let c denote the (constant) return for the benchmark R_b. Then it suffices to rank the probability $\lim_{T\to\infty} P(\frac{1}{T}\sum_{t=1}^{T}\log R_{pt} \geqslant c)$. By the discussion in the previous section, $\lim_{T\to\infty} P(\frac{1}{T}\sum_{t=1}^{T}\log R_{pt} \geqslant c)$ can be estimated via the rate function $I(c;\, p)$ of LDP. Thus, the under performance probability rate, $I(c;\, p)$, of decay to zero as $T \to \infty$ is the proposed *ranking index for portfolios*. A portfolio whose under performance probability decays to zero at a higher rate will be ranked higher than a portfolio with a lower decay rate. See Stutzer (2000) for details.

Similar ideas can be used to study diagnosis of asset pricing models. We refer to Stutzer (1995), Glasserman and Jin (1999), and Kitamura and Stutzer (2002) for details.

4 Tail risk of portfolios

Modern risk management is often concerned with low probability events in the tail of the distribution of the future profit and loss. Since LDP deals with the probability of rare events, it is not surprising that it has a natural application to risk management and the estimation of large losses which occur in the tail of the distribution. In this section we explain how LDP can be used in portfolio risk management. We confine ourselves to the risk management of a credit portfolio. [See Dembo et al. (2004), Glasserman (2005) and Gordy (2002) for more complete details.]

Assume we are dealing with a credit portfolio. There are m obligors to which the portfolio is exposed. Assume Y_k is default indicator for kth obligor; U_k is the loss resulting from the default of kth obligor. Hence the total loss exposure is

$$L = \sum_{i=1}^{m} Y_i U_i. \tag{4.1}$$

We assume the credit risk is modeled by some risk factors $Z_1, Z_2, \ldots, Z_d$, and $Y_1, \ldots, Y_m$ are independent conditional on $Z = (Z_1, Z_2, \ldots, Z_d)$. For simplicity $U_1, \ldots, U_m$ are constants. These risk factors Z could be some macro-economic factors. The default indicator Y_i can be chosen as

$$Y_i = 1_{\{X_i > x_i\}}, \quad i = 1, 2, \ldots, m, \tag{4.2}$$

where we assume X_i is a standard normal distribution and the threshold x_i denotes the default boundary as in the Merton model. Let p_i denote the marginal probability that ith obligor defaults. Then p_i and x_i are related to one another as follows:

$$x_i = -N^{-1}(p_i), \quad p_i = N(-x_i),\ i = 1, 2, \ldots, m. \tag{4.3}$$

For a given $x > 0$, our objective is to estimate the large loss probability $P(L \geqslant x)$. Conditional on Z, the cumulant generating function $\Lambda(\theta; z)$ is defined by

$$\Lambda(\theta; z) = \log E\big[e^{\theta L} | Z = z\big]; \quad z \in \mathcal{R}^d. \tag{4.4}$$

Clearly,

$$P(L > x|Z) \leqslant e^{\Lambda(\theta;Z)-\theta x}. \tag{4.5}$$

Hence $\Lambda(\theta; Z)$ contains information about the tail risk of the *conditional* loss distribution.

Write

$$F(x; z) :\equiv -\mathrm{Sup}_{\theta \geqslant 0}\big\{\theta x - \Lambda(\theta; z)\big\} \tag{4.6}$$

which is obtained at unique $\theta_x(z)$. Actually

$$\theta_x(z) = \begin{cases} \text{unique } \theta \text{ such that } \frac{\partial \Lambda(\theta;z)}{\partial \theta} = x, & \text{if } x > E[L|Z = z], \\ 0, & \text{if } x \leqslant E[L|Z = z]. \end{cases}$$

Thus one is able to estimate the large loss probability, as presented in the following result of Glasserman (2005).

Theorem 4.1. (i) *For every* $x > 0$,

$$P(L > x) \leqslant E\big[e^{F(x;Z)}\big].$$

(ii) $F(x; z) = 0$ *if and only if* $E[L|Z = z] \geqslant x$.

(iii) *If the function* $F(x, .)$ *is concave for* $x > 0$, *then*

$$P(L > x) \leqslant e^{-J(x)}$$

where

$$J(x) = -\max_z \left\{F(x; z) - \frac{1}{2} z^T z\right\}.$$

5 Application to simulation

In this section we explain how we can use large deviations to better simulate rare events in financial engineering. As we have seen in previous sections, rare events are of increasing interest in different areas of financial engineering e.g. credit risk applications. We first explain the use of importance sampling in simulating rare events. Then we describe applications of LDP in this context.

5.1 Importance sampling

Importance sampling is a useful technique for simulating rare events. The idea is to change the probability measure so that more weight is shifted to the region of interest. If we are interested in simulating a given (rare) event A under the P-probability, the idea in importance sampling is to sample from a different distribution, say Q, under which A has a larger chance of occurring. This is done by specifying the Radon–Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P}$. Given a Radon–Nikodym derivative, set $\frac{\mathrm{d}P}{\mathrm{d}Q} := (\frac{\mathrm{d}Q}{\mathrm{d}P})^{-1}$ and we have

$$P(A) = \int 1_A \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q, \tag{5.1}$$

where q_A denotes the indicator function of the event A. The importance sampling estimator of $P(A)$ is founded by drawing N independent samples $X_1, \ldots, X_N$ from Q:

$$P(A; Q, N) := \frac{1}{N} \sum_{i=1}^{N} 1_{\{X_i \in A\}} \frac{\mathrm{d}P}{\mathrm{d}Q}(X_i). \tag{5.2}$$

It is clear that $P(A; Q, N)$ is an unbiased estimator, i.e., $E^Q[P(A; Q, N)] = P(A)$. However, to reduce the effects of noise, one should choose an efficient distribution Q in the sense that the variance of the estimator $P(A; Q, N)$ is as small as possible. In particular, we wish to find Q to minimize (the variance) $\int 1_A (\frac{\mathrm{d}P}{\mathrm{d}Q})^2\,\mathrm{d}Q$. It is well known that the zero-variance estimator is possible by setting Q corresponding to the conditional distribution of P given A. However, since $P(A)$ is unknown, such a zero-variance estimator is not useful in practice.

An alternative criterion is to consider the so-called *relative error* as

$$\eta_N(A; Q/P) := \frac{E^Q[P(A; Q, N)^2]}{P(A)^2} - 1. \tag{5.3}$$

This "relative error" concept measures the variability of $P(A; Q, N)$ because that the square root of the relative error is proportional to the width of a confidence interval relative to the expected estimate itself. The idea is choose the smallest sample required to obtain a fixed confidence level. In other words, choose a fixed maximum relative error $0 < \eta_{\max} < \infty$, and define

$$N(Q/P) := \inf\bigl\{N\colon\ \eta_N(A; Q/P) \leqslant \eta_{\max}\bigr\}. \tag{5.4}$$

Large deviation principles can be used to find the following *asymptotic efficiency* estimator. Consider a sequence of measure $\{P_\varepsilon\}$ and $P = P_{\varepsilon_0}$ for some $\varepsilon_0 > 0$. Given a continuous linear functional $\eta : \Omega \to R$, and assume that

$$\lim_{M\to 0} \limsup_{\varepsilon\to 0} \varepsilon \log \int\limits_{\{\eta^{-1}([M,\infty))\}} \exp\bigl[\eta(\omega)/\varepsilon\bigr] P_\varepsilon\,(\mathrm{d}\omega) = -\infty \tag{5.5}$$

and similarly for ζ replaced by $-\zeta$. Such a ζ is said to be *exponentially twisting*.

Define a new family of probability measures $\{Q_\varepsilon^\zeta\}$ as follows:

$$\frac{\mathrm{d}Q_\varepsilon^\zeta}{\mathrm{d}P_\varepsilon}(x) := \exp\left\{\frac{\zeta(x)}{\varepsilon} - \log \int_{\exp} \left[\frac{\zeta(y)}{\varepsilon}\right] P_\varepsilon(\mathrm{d}y)\right\}. \tag{5.6}$$

The measures $\{Q_\varepsilon^\zeta\}$ are said to be *exponentially twisted with twist* ζ. ζ is *asymptotically efficient* if

$$\limsup_{\varepsilon\to 0} \varepsilon \log N(Q_\varepsilon/P_\varepsilon) = 0. \tag{5.7}$$

So the criterion is to find an asymptotically efficient exponential twist η such that the family Q_ε can be used to estimate $P(A)$ in the importance sampling. Recall A is an I-continuity set, where $I:\Omega \to R$ is a functional, if $\lim_{\varepsilon\to 0} \varepsilon \log P_\varepsilon(A) = -\inf_{x\in A} I(x)$. We have the following see Sadowsky (1996) and Dieker and Mandjes (2005).

Theorem 5.1. *Assume that $\{P_\varepsilon\}$ satisfies the LDP with a rate function I, and A is both an I-continuity set and an $(I+\zeta)$-continuity set. Then an exponentially twisting ζ is asymptotically efficient if and only if*

$$\inf_{x\in\Omega}\big[I(x) - \zeta(x)\big] + \inf_{x\in\overline{A}}\big[I(x) + \zeta(x)\big] = 2 \inf_{x\in\overline{A}} I(x).$$

The asymptotically efficient twist is useful in practice. For example, under some technical conditions, there exists a unique asymptotically efficient twist. So the best possible choice for simulation is to check whether it is asymptotically efficient. However, this is not always the best choice. See Glasserman and Wang (1997) for a counterexample.

5.2 Sharp large deviations

Monte Carlo simulation is a very powerful tool for pricing complicated path-dependent options [such as in Boyle (1977) and Boyle et al. (1997)]. For example under discrete monitoring, there is no analytical expression for the price of a *knock-and-out* call option. The payoff resembles that of the corresponding standard call, provided that the underlying asset price does not hit the barrier prior to option maturity. Otherwise its payoff is equal to zero. Thus, in the simulation of the underlying, we have to simulate the underlying path and the first hitting time.

The knock-and-out call option formula is

$$C(0) = e^{-rT} E_0\big[\max(S_T - K, 0) 1_{\{\tau \geqslant T\}}\big], \tag{5.8}$$

where τ denotes the first time the asset price S_t hits the barrier. To simulate the path of the underlying asset price, consider a partition $t_0 = 0 < t_1 < \cdots <$

$t_n = T$ of the time period $[0, T]$ with $t_{i+1} - t_i = \varepsilon := \frac{T}{n}$ for each $i = 0, 1, \ldots, n-1$. At each step the asset price S_{t_i} is simulated and the standard procedure sets the hitting time equal to the first time t_i in which S_{t_i} crosses a boundary. However, this procedure provides an *overestimate* of the first hitting time since S_{t_i} and $S_{t_{i+1}}$ might not have breached the barriers while S_t, $t \in (t_i, t_{i+1})$ had. To account for this, we can estimate the probability p_i^ε that S_t hits the barrier during the time interval $[t_i, t_{i+1})$, given the observations S_{t_i} and $S_{t_{i+1}}$.

In this section we provide an approximation, using LDP and its refined version (called sharp LDP), of p_i^ε by studying its asymptotic behavior as $\varepsilon \to 0$. Suppose we have two barriers an upper barrier denoted by $U(.)$ and a lower barrier defined by $L(.)$. We fix a time period $[T_0, T_0+\varepsilon]$, and let $p^\varepsilon_{U,L}(T_0, \zeta, y)$ the probability that the process S hits either barrier during the time interval, given the observations $\log S_{T_0} = \zeta$ and $\log S_{T_0+\varepsilon} = y$. The following result was derived by Baldi et al. (1999). To simplify notations we only state the result for the upper single-barrier case ($L = -\infty$).

Theorem 5.2. *Suppose U is continuous with Lipschitz continuous derivatives. Then for every $\zeta, y < U(T_0)$,*

$$p^\varepsilon_U(T_0, \zeta, y) = \exp\left\{-\frac{Q_U(T_0, \zeta, y)}{\varepsilon} - R_U(T_0, \zeta, y)\right\}\big(1 + 0(\varepsilon)\big), \quad (5.9)$$

where

$$Q_U(T_0, \zeta, y) = \frac{2}{\sigma^2}\big(U(T_0) - \zeta\big)\big(U(T_0) - y\big),$$
$$R_U(T_0, \zeta, y) = \frac{2}{\sigma^2}\big(U(T_0) - \zeta\big)U'(T_0). \quad (5.10)$$

Therefore, the *Corrected Monte Carlo Simulation Procedure* is as follows: During the simulation from t_i to $t_i + \varepsilon$, with probability

$$p_i^\varepsilon = \exp\left\{-\frac{Q_{U,L}(t_i, \log S_{t_i}, \log S_{t_{i+1}})}{\varepsilon} - R_{U,L}(t_i, \log S_{t_i}, \log S_{t_{i+1}})\right\} \quad (5.11)$$

we stop the simulation and set t_i as the hitting time τ. With probability $1 - p_i^\varepsilon$ we carry on the simulation.

In its implementation, this Corrected Monte Carlo Simulation Procedure does not add to the complexity of the work thanks to the simple analytical expression for p_i^ε. Thus this correction procedure can be used in any situation where first hitting time probability is involved, such as barrier options, and credit products in a structural credit model.

We now briefly explain how this result is derived using LDP and sharp LDP. Write

$$W_t = \zeta + \rho(t - T_0) + \sigma(B_t - B_{T_0}), \quad t \in [T_0, T_0 + \varepsilon]. \quad (5.12)$$

During the time period $[T_0, T_0+\varepsilon]$ we study the probability of $(W_\cdot)$ of hitting the upper barrier $U(.)$, namely

$$p_U^\varepsilon(T_0, \zeta, y) = P(\tau_U \leqslant \varepsilon \mid W_{T_0} = \zeta,\ W_{T_0+\varepsilon} = y) \tag{5.13}$$

where τ_U is the stopping time

$$\tau_U = \inf\bigl\{t > 0\colon\ W_{T_0+t} \geqslant U(T_0 + t)\bigr\}.$$

Note that the law of (W_{T_0+t}) over $t \in [0, \varepsilon]$ conditional on $\{W_{T_0} = \zeta,\ W_{T_0+\varepsilon} = y\}$, coincides with the probability on $C([0, \varepsilon], R)$ induced by the Brownian bridge:

$$\overline{W}_t^\varepsilon = \zeta + \frac{t}{\varepsilon}(y - \zeta) + \sigma\left(B_t - \frac{t}{\varepsilon}B_\varepsilon\right), \quad t \in [0, \varepsilon]. \tag{5.14}$$

Introduce the time change $t \to \frac{t}{\varepsilon}$. Then the process $Z_t^\varepsilon = \overline{W}_{t/\varepsilon}^\varepsilon$ is given by

$$Z_t^\varepsilon = \zeta + t(y - \zeta) + \sigma\sqrt{\varepsilon}(B_t - tB_1). \tag{5.15}$$

Z_t^ε can be seen to be a diffusion process with a small parameter $\sqrt{\varepsilon}$, a typically setting of the large deviation theory in Freidlin–Wentzell theory. For any $s < 1$ consider the following stochastic differential equation

$$\mathrm{d}\overline{Z}_t^\varepsilon = -\frac{\overline{Z}_t^\varepsilon - y}{1 - t}\,\mathrm{d}t + \sigma\sqrt{\varepsilon}\,\mathrm{d}B_t, \quad \overline{Z}_0^\varepsilon = \zeta, \tag{5.16}$$

then up to time s, $\overline{Z}_t^\varepsilon$ coincides with the law of Z_t^ε. Define $X_t^\varepsilon = \overline{Z}_t^\varepsilon - U(T_0 + \varepsilon t)$, then

$$\begin{aligned}\mathrm{d}X_t^\varepsilon = &-\left[\varepsilon U'(T_0 + \varepsilon t) + \frac{X_t^\varepsilon - y + U(T_0 + \varepsilon t)}{1 - t}\right]\mathrm{d}t \\ &+ \sigma\sqrt{\varepsilon}\,\mathrm{d}B_t, \quad X_0^\varepsilon = \zeta - U(T_0).\end{aligned} \tag{5.17}$$

Let $P_{x,s}^\varepsilon$ denotes the law of Z^ε with the initial condition $X_s^\varepsilon = x$, and τ_0 denote the hitting time of 0 for X^ε. Then it can be proved that the family $\{X_\varepsilon^\varepsilon\}$ satisfies a LDP (*the LDP of Brownian bridges*) on the space $C([s, 1], R)$ with rate function [see Baldi (1995) for details]

$$J(h) = \begin{cases} \frac{1}{2\sigma^2}\bigl[\int_s^1 h_r'^2\,\mathrm{d}r - \frac{(y-x-U(T_0))^2}{1-s}\bigr], & \text{if } h \in \Delta_{x,s}, \\ \infty, & \text{otherwise,}\end{cases}$$

where $\Delta_{x,s}$ is the set of absolutely continuous paths on $[s, 1]$, starting at x at time s and reaching $y - U(T_0)$ at time 1. Therefore,

$$\lim_{\varepsilon\to 0} \varepsilon \log P_{x,s}^\varepsilon(\tau_0 < 1) = -u(x, s) = \inf_{h\in\Delta_{x,s}} J(h). \tag{5.18}$$

By standard variational calculus, it can be proved that, if $x < 0$, then

$$u(x, s) = \frac{2}{\sigma^2(1 - s)}x\bigl(y - U(T_0)\bigr) \tag{5.19}$$

and otherwise $u(x, s) = 0$. Hence, since $\zeta < U(T_0)$, we see that

$$p_U^{\varepsilon}(T_0, \zeta, y) \sim \exp\left\{-\frac{2(\zeta - U(T_0))(y - U(T_0))}{\varepsilon\sigma^2}\right\}. \tag{5.20}$$

A refined LDP, namely sharp LDP which is proved in Fleming and James (1992), states that, under some fairly general technical assumptions

$$P_{x,s}^{\varepsilon}(\tau_0 \leqslant 1) \sim \exp\left\{-\frac{u(x, s)}{\varepsilon} - w(x, s)\right\}, \tag{5.21}$$

where $w(x, s)$ only depends on x and s. Thus the above theorem can be derived by straightforward verification for X^{ε}. This sharp LDP provides a refined estimation of the probability $P_U^{\varepsilon}(T_0, \zeta, y)$.

6 Incomplete markets

Asset pricing in incomplete financial markets is an important topic in finance. Market incompleteness means that there are infinitely many equivalent martingale measures. One important issue concerns the indeterminacy of the equivalent martingale measures (or market price of risk), and another concerns the effect of misspecification errors. In this section we briefly discuss a recent application of LDP to address these issues, and refer to Boyle et al. (2007) for further discussions and details.

In the incomplete markets literature, different approaches have been proposed to either select a particular equivalent martingale measure or restrict the range of equivalent martingale measures. For example one approach is to postulate a benchmark investor and value the security from the perspective of this investor. We use this investor's preferences to identify a particular equivalent martingale measure. Another approach is to construct a distance metric to facilitate a comparison between two equivalent martingale measures or stochastic discount factors. Then the range of stochastic discount factors can be reduced by imposing some suitable criterion. These criteria can be derived from economic considerations and several such criteria have been proposed in the literature.

In much of the literature on incomplete model asset pricing, model specification is often given exogenously. There is usually not much discussion on stochastic discount factor (SDF) misspecification and its effect on asset pricing apart from a few notable exceptions. These include Hensen and Jagannathan (1997), Hodrick and Zhou (2001) who studied SDF misspecification errors. Given a "reference SDF," the agent under consideration might suspect it to be misspecified and she behaves *conservatively*. She therefore wishes to consider the *worst-case misspecification* of alternative choices of models or pricing kernels. In dual terms, they defined a "quadratic distance" between two SDFs, and this quadratic distance captures the "maximum pricing errors" across all

contingent claims. Cochrane and Saa-Requejo (2000) discussed a similar concept. Bernardo and Ledoit (2000) presented a (relative) distance between two SDFs in terms of the extreme values of the ratios of these SDFs across different states. According to their duality result, this distance concept also characterizes the worse-case error across the contingent claims in some sense. While the worst-case misspecification is robust, these distance concepts cannot capture the misspecification error for one *particular* contingent claim.

Boyle et al. (2007) use LDP ideas to propose a framework to address the model misspecification and SDF misspecification together. In this analysis the agent has model misspecification concerns when she estimates the SDF. Her decision rule regarding the SDF depends on the *sensitivity*, of asset pricing with respect to some form of model misspecification. The robust SDF has the *least* impact with respect to some forms of model misspecification in asset pricing. In this analysis, the effect of model misspecification is explained by *pricing* under SDF. Therefore, this analysis works for any given contingent claim, not for the worst-case error on a large class of contingent claim.

As an example to illustrate this approach, we consider a the Stein and Stein stochastic volatility model see Stein and Stein (1991). The risky asset dynamics are given by

$$\frac{\mathrm{d}S(t)}{S(t)} = \sqrt{|v_t|}\,\mathrm{d}W_1(t), \tag{6.1}$$

where v_t satisfies

$$\mathrm{d}v_t = (\kappa_1 - \kappa_2 v_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_2(t). \tag{6.2}$$

Here κ_1, κ_2 and $\sigma > 0$ are model parameters, $v_0 = \sigma_0^2$, and $(W_1(t), W_2(t))$ is a two-dimensional Brownian motion. The instantaneous variance and the instantaneous volatility are $|v_t|$ and $\sqrt{|v_t|}$, respectively. In this model the risk-free asset is used as a numéraire. When $\sigma \equiv 0$, this model reverts to a standard Black–Scholes complete market model since in this case $v_t > 0$ is deterministic function of time. One might wonder if σ small enough, will this model be close to a complete model. However, for any small positive σ, the model is still incomplete, and its no-arbitrage bound (for a vanilla call option as above) is Merton's bound. This was proved by Frey and Sin. (1999). Hence making σ smaller does not reduce the "incompleteness" of the model. On the other hand, if we fix an equivalent martingale measure, the smaller σ, the smaller the price (expectation) under this measure. When $\sigma \to 0$, the expectation under this measure approaches to Black–Scholes value. Hence, making σ smaller does *reduce the incompleteness* of the model in some sense. This seems to be something of a paradox.

The paradox can be resolved by a robust approach which we will describe shortly. It turns out that the *rate function* from the appropriate large deviation principle provides a *measurement of distance* between the case when $\sigma > 0$ and the limit case when $\sigma = 0$ under each equivalent martingale measure. It is

convenient to restrict ourselves to linear-type market prices of risk of the form

$$\lambda_{(a,b)} \equiv \lambda = a + bv_t,$$

where both a, b are real numbers. The corresponding pricing kernel (or SDF) is determined by

$$\eta_T^\lambda = \exp\left[-\int_0^T (a + bv_t)\,\mathrm{d}W_2(t) - \frac{1}{2}\int_0^T (a + bv_t)^2\,\mathrm{d}t \right].$$

The corresponding equivalent martingale measure P^λ is determined by $\frac{\mathrm{d}P^\lambda}{\mathrm{d}P} = \eta_T^\lambda$. There exists a Brownian motion $W^\lambda(t) \equiv (W_1^\lambda(t), W_2^\lambda(t))$ such that the risky asset's dynamics under this measure P^λ satisfies the following stochastic differential equation:

$$\begin{aligned} &\frac{\mathrm{d}S(t)}{S(t)} = \sqrt{|v_t|}\,\mathrm{d}W_1^\lambda(t), \\ &\mathrm{d}v_t = \big[\kappa_1 - \kappa_2 v_t - \sigma(a + bv_t)\big]\,\mathrm{d}t + \sigma\,\mathrm{d}W_2^\lambda(t). \end{aligned} \tag{6.3}$$

The relationship between the two Brownian motions comes from the Girsanov transformation

$$W_1^\lambda = W_1(t); \qquad W_2^\lambda(t) = W_2(t) + \int_0^t (a + bv_s)\,\mathrm{d}s. \tag{6.4}$$

Consider a sequence of models of this risky asset, under the fixed measure P^λ, for every $\varepsilon \in [0, 1]$, as follows:

$$\begin{aligned} &\frac{\mathrm{d}S(t)}{S(t)} = \sqrt{|v_t^{\lambda,\varepsilon}|}\,\mathrm{d}W_1^\lambda(t), \\ &\mathrm{d}v_t^{\lambda,\varepsilon} = \big[\kappa_1 - \kappa_2 v_t^{\lambda,\varepsilon} - \sigma\big(a + bv_t^{\lambda,\varepsilon}\big)\big]\,\mathrm{d}t + \varepsilon\sigma\,\mathrm{d}W_2^\lambda(t). \end{aligned} \tag{6.5}$$

For every $\varepsilon \in (0, 1]$, this is an incomplete model under measure P^λ. Each model gives a different process for the risky asset dynamics, under which the volatility of volatility is $\varepsilon\sigma$. In particular when $\varepsilon = 1$ we recover the original model with market price of risk λ. The difference between the models come from the variations in the process for the instantaneous variance $|v_t^\varepsilon|$. In the limit case $\varepsilon = 0$, $v_t^{\lambda,\varepsilon} = V_t^\lambda$, and the model becomes a complete Black–Scholes model.

The limit model, when $\varepsilon = 0$, is known as the "*benchmark complete model*" for the given market price of risk λ. We denote it by $\mathcal{B}_\lambda$, $\lambda = \lambda_{(a,b)}$, $(a, b) \in \mathcal{R}^2$. Specifically, in this model $\mathcal{B}_\lambda$, the risky asset satisfies

$$\frac{\mathrm{d}S(t)}{S(t)} = \sqrt{v_t^\lambda}\,\mathrm{d}W_1^\lambda(t),$$
$$\mathrm{d}v_t^\lambda = \big[\kappa_1 - \kappa_2 v_t^\lambda - \sigma\big(a + bv_t^\lambda\big)\big]\,\mathrm{d}t, \quad v_0^\lambda = \sigma_0^2. \tag{6.6}$$

The market price of risk is chosen so that the process v_t^λ is positive. Write

$$\kappa_1^\lambda = \kappa_1 - \sigma a; \qquad \kappa_2^\lambda = \kappa_2 + \sigma b. \tag{6.7}$$

Thus the market price of risk λ corresponds to an affine transformation of the parameters $\{\kappa_1, \kappa_2, \sigma\}$:

$$\kappa_1 \to \kappa_1^\lambda; \qquad \kappa_2 \to \kappa_2^\lambda. \tag{6.8}$$

To analyze the prices in both the incomplete model and the complete model it is convenient to define $\overline{V}^{\lambda,\varepsilon} = \overline{V}^{\lambda,\varepsilon}(\kappa_1, \kappa_2) = \int_0^T |v_t^{\lambda,\varepsilon}|\,\mathrm{d}t$, and $\overline{V}^\lambda = \overline{V}^\lambda(\kappa_1, \kappa_2) = \int_0^T v_t^\lambda\,\mathrm{d}t$. These two quantities correspond to the integrated variance terms which play a critical role in pricing options. Using this notation we let $H_0 = C_{\mathrm{BS}}(S_0, \sqrt{\overline{V}^\lambda})$ be the value of the call option in the benchmark (complete market) model. Similarly we let $C_{\mathrm{BS}}(S_0, \sqrt{\overline{V}^{\lambda,\varepsilon}})$ be the value of the option with random volatility $\sqrt{\overline{V}^{\lambda,\varepsilon}}$ in the incomplete market. Moreover, the option price in the original incomplete model, under the choice of market price of risk λ, equals to conditional expectation of $C_{\mathrm{BS}}(S_0, \sqrt{\overline{V}^{\lambda,\varepsilon}})$ under the path $v_t^{\lambda,\varepsilon}$. We now explain briefly how a large deviation principle can be used to study the distance between incomplete and complete markets.

For any $\delta > 0$, let $B_\delta(H_0)$ be the collection of all bounded continuous functions $\phi(t)$ on $[0, T]$ such that

$$\left| C_{\mathrm{BS}}\left(S_0, \sqrt{\int_0^T \big|\phi(t)\big|\,\mathrm{d}t}\right) - H_0 \right| \geqslant \delta. \tag{6.9}$$

The LDP for the process $(v_t^{\lambda,\varepsilon})$ implies that for a given $\delta > 0$, we can make the following probability statement:

$$P^\lambda\big(v_t^{\lambda,\varepsilon} \in B_\delta(H_0)\big) \preceq e^{-\frac{1}{\varepsilon^2}\inf_{B_\delta(H_0)} I(\phi)}. \tag{6.10}$$

The function $I(\cdot)$ is the *action functional* or *rate function*. For this case it can be shown that the action functional has an explicit functional expression given by

$$I(\phi) = \frac{1}{2}\int_0^T \frac{(\phi_t - \kappa_1^\lambda + \kappa_2^\lambda \phi)^2}{\sigma^2}\,\mathrm{d}t$$

when ϕ is absolutely continuous on $[0, T]$ and $\phi(0) = \sigma_0^2$. Given any $\delta > 0$, the quantity $\inf_{B_\delta(H_0)} I(\phi)$ is known as the decay rate for deviations of no less than δ from H_0.

Let A_δ be the set of all continuous functions $\phi(t)$ on $[0, T]$ satisfying

$$\left|\sqrt{\int_0^T |\phi(t)|\,\mathrm{d}t} - \sqrt{\overline{V}^\lambda}\right| \geqslant \delta.$$

By using the large deviation principle on A_δ, we have

$$P^\lambda\big(v_t^{\lambda,\varepsilon} \in B_{\delta_1}(H_0)\big) \preceq e^{-\frac{1}{\varepsilon^2}\inf_{A_\delta} I(\phi)}. \tag{6.11}$$

The infimum in the exponent is a critical component of the large deviations approach. The use of the infimum means that the probability of a deviation greater than δ satisfies the bound for all functions ϕ in the set A_δ. Write

$$R(\lambda, \delta) = \inf_{A_\delta} I(\phi). \tag{6.12}$$

In our case $R(\lambda, \delta)$ will depend on the model parameters $\kappa_1, \kappa_2, \sigma$ and the selected market price of risk λ. The $R(\lambda, \delta)$ term represents the convergence speed for this case. The higher $R(\lambda, \delta)$ is, the faster the convergence from $C_{\mathrm{BS}}(S_0, \sqrt{\int_0^T |v_t^{\lambda,\varepsilon}|\,\mathrm{d}t})$ to H_0. We define the function $R(\lambda, \delta)$ as measure of the *proximity*.

In our case $v_t^{\lambda,\varepsilon}$ follows an Ornstein–Uhlenbeck process and we are able to obtain an explicit formula for $R(\lambda, \delta)$. It turns out that the functional dependence of R on the underlying parameters, $(\kappa_1^\lambda, \kappa_2^\lambda)$ is different for three different ranges. For our purpose here, we just present the explicit expression of $R(\lambda, \delta)$ when δ is sufficiently *small*. In this case the functional from is

$$R(\lambda, \delta) = \frac{1}{\sigma^2} J(\lambda, \delta), \tag{6.13}$$

where

$$J(\lambda, \delta) = J_1\big(\kappa_2^\lambda\big)\Big(2\sqrt{\overline{V}\big(\kappa_1^\lambda, \kappa_2^\lambda\big)}\delta - \delta^2\Big)^2, \tag{6.14}$$

$$J_1(x) := \frac{x^3(e^{2xT} - 1)}{(e^{\frac{x}{2}T} - e^{-\frac{x}{2}T})^4}. \tag{6.15}$$

The corresponding expressions for the case when $\kappa_2^\lambda = 0$ is also available.

To summarize, the connection between the incomplete market and the complete market can be formalized as follows. We have established the probability bound (for small η):

$$P^\lambda\left[\left|C_{\mathrm{BS}}\left(S_0, \sqrt{\int_0^T |v_t^\varepsilon|\,\mathrm{d}t}\right) - H_0\right| \geqslant \eta H_0\right] \preceq e^{-\frac{1}{\sigma^2\varepsilon^2} J(\lambda, \frac{\eta H_0}{S_0})}. \tag{6.16}$$

The rate function $R(\lambda, \delta)$ (or $J(.,.)$) provides a rich setting to investigate the effects of the choice of market price of risk on a given contingent claim. For example, given a δ, the above estimate can be used to study the effects

of moneyness. It is shown that, for any given λ, the larger the moneyness, the larger $J(\lambda; \frac{\eta H_0}{S_0})$. If we choose a special $\delta = \delta^* = \sqrt{\overline{V}(\kappa_1^\lambda, \kappa_2^\lambda)}$, the maximum point of function $J(.)$, the rate function can be used to define

$$d(\lambda; x) \equiv \frac{1}{\sigma^2} J_1\big(\kappa_2^\lambda\big)\overline{V}\big(\kappa_1^\lambda, \kappa_2^\lambda\big)^2, \tag{6.17}$$

where x denotes the option used in the computations. The function $d(\lambda; x)$ captures the decay rate information, namely the *proximity* between the incomplete market and the complete market in our framework. Note that if the decay rate is high the convergence speed is fast so the function d defined here is inversely related to the usual notion of distance.

An examination of the dependence of $d(\lambda; x)$ on λ is important for understanding the effect on the decay rate function. For example, it can be shown that, $d(\lambda; x)$ is increasing with respect to both κ_1^λ and κ_2^λ. It turns out that higher the parameters κ_2, and κ_1 leads to the price in incomplete market being closer to the corresponding complete market price. If we assume that all possible market prices of risk $\lambda_{(a,b)}$ on the following rectangle:

$$0 < A_{\min} \leqslant \kappa_1^\lambda \leqslant A_{\max}, \qquad 0 < B_{\min} \leqslant \kappa_2^\lambda \leqslant B_{\max}.$$

Therefore $d(\kappa_1^\lambda, \kappa_2^\lambda)$ reaches its maximum at the top right-hand corner of the rectangle when $\kappa_1^\lambda = A_{\max}$ and $\kappa_2^\lambda = B_{\max}$. In this case the market price of risk

$$\lambda^* = \frac{\kappa_1 - A_{\max}}{\sigma} + \frac{B_{\max} - \kappa_2}{\sigma} v_t \tag{6.18}$$

makes the stochastic volatility model *most closely resemble* a complete model. In other words, λ^* is the most robust market price of risk for the call option in the sense that its pricing under the corresponding SDF is most resistant to model misspecification. Other choices of the market price of risk choices make the *incompleteness* of the SV model more pronounced. This result makes intuitive sense since it is not surprising that the *optimal* choice lies on the boundary. In addition the fact that it corresponds to the highest possible value of κ_2^λ is intuitive. The κ_2^λ parameter captures the speed of mean reversion and measures how quickly v returns to its mean value. A high value of mean reversion indicates that the price in the incomplete market is more likely to converge more quickly to its complete market benchmark.

7 Conclusions and potential topics for future research

Large deviation principles deal with the probability of rare events. We have shown that LDP has several applications in mathematical finance including asset pricing, risk management and financial engineering. We should stress that there are other applications are not covered here because of space limitations. Interested readers should consult the literatures at the end of this chapter. See

Callen et al. (2000), Foster and Stutzer (2002), Glasserman and Jin (1999), Nummelin (2000), Sornette (1998), and Williams (2004).

We suggest there are other potential further applications of LDP. First, LDP provides an alternative criterion for portfolio selection. One possible extension is to consider more realistic processes for asset price dynamics such as a stochastic volatility model or a jump-diffusion model. Another possible topic is to use LDP to analyze portfolio selection with constraints. Under some constraints such as the credit limit as in Grossman and Vila (1992), closed form solutions are impossible to obtain. But from an asymptotic perspective these constraints are equivalent to some simpler constraints. LDP can provide the quantitative framework to deal with problems of this nature.

To cite another possible application, we can consider risk measures. Risk managers are increasingly interested in risk measures. For example, VaR is just one popular risk measure for the maximum loss with some confidence level. LDP enables us to discuss other risk measures. For instance, given one level of possible loss, it is also interesting to estimate the probability of the portfolio value meeting this threshold in a short time period. This turns out to be an exit time probability problem as we have discussed before.

The LDP approach can be used in conjunction with Monte Carlo simulation to estimate large portfolio risk and pricing of credit derivatives such as credit default obligations.

In the previous section we outlined briefly how LDP can be used to study the relationship between the market price of risk and security prices in an incomplete market. We did this by setting up a benchmark complete market model and analyzing the behavior of a sequence of incomplete markets that converge in the limit to a complete model. We used a large deviations approach to analyze this convergence. Large deviations techniques provide a powerful analytical tool that is well suited for this application. This approach was applied to a stochastic volatility model. However it may prove useful in the analysis of other types of incomplete markets. To fully apply this theory we need an LDP for the relevant stochastic process. In some cases establishing the appropriate LDP is a challenging technical exercise. Hence mathematical finance gives a further impetus for theoretical research in this important area of probability theory.

Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada for support.

References

Baldi, P. (1995). Exact asymptotics for the probability of exit from a domain and applications to simulation. *Annals of Probability* 23, 1644–1670.

Baldi, P., Caramellino, L., Iovino, M.G. (1999). Pricing general barrier options: A numerical approach using sharp large deviations. *Mathematical Finance* 9, 293–322.
Bernardo, A., Ledoit, O. (2000). Gain, loss, and asset pricing. *Journal of Political Economy* 108, 144–172.
Boyle, P.P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics* 4, 323–338.
Boyle, P., Broadie, M., Glasserman, P. (1997). Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control* 21, 1267–1321.
Boyle, P., Feng, S., Tian, W., Wang, T. (2007). Robust stochastic discount factors. Review of Financial Studies, in press.
Browne, S. (1999). Reaching goals by a deadline: Digital options and continuous-time active portfolio management. *Advances in Applied Probability* 31, 551–577.
Callen, J., Govindaraj, S., Xiu, L. (2000). Large time and small noise asymptotic results for mean reverting diffusion processes with applications. *Economic Theory* 16, 401–419.
Cochrane, J., Saa-Requejo, J. (2000). Beyond arbitrage: Good-deal asset price bounds in incomplete Markets. *Journal of Political Economy* 108, 79–119.
Dawson, D., Feng, S. (1998). Large deviations for the Fleming–Viot process with neutral mutation and selection. *Stochastic Processes and Their Applications* 77, 207–232.
Dawson, D., Feng, S. (2001). Large deviations for the Fleming–Viot process with neutral mutation and selection, II. *Stochastic Processes and Their Applications* 92, 131–162.
Dembo, A., Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, second ed. *Applications of Mathematics*, vol. 38. Springer-Verlag, New York.
Dembo, A., Deuschel, J., Duffie, D. (2004). Large portfolio losses. *Finance and Stochastics* 8, 3–16.
Dieker, A., Mandjes, M. (2005). On asymptotically efficient simulation of large deviation probabilities. *Annals of Applied Probability* 37, 539–552.
Feng, S., Xiong, J. (2002). Large deviations and quasi-potential of a Fleming–Viot process. *Electronic Communications in Probability* 7, 13–25.
Fleming, W.H., James, M. (1992). Asymptotics series and exit time probabilities. *Annals of Probability* 20, 1369–1384.
Föllmer, H., Leukert, P. (1999). Quantile hedging. *Finance and Stochastics* 3, 251–273.
Foster, F.D., Stutzer, M. (2002). Performance and risk aversion of funds with benchmarks: A large deviation approach. *Working paper*.
Freidlin, M.I., Wentzell, A.D. (1998). *Random Perturbations of Dynamical Systems*, second ed. Springer-Verlag, New York.
Frey, R., Sin, C. (1999). Bounds on European option prices under stochastic volatility. *Mathematical Finance* 9, 97–116.
Glasserman, P. (2005). Tail approximations for portfolio credit risk. *Journal of Computational Finance* 9, 1–41.
Glasserman, P., Jin, Y. (1999). Comparing stochastic discount factors through their implied measures. *Working paper*, Columbia Business School.
Glasserman, P., Wang, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *Annals of Applied Probability* 7, 731–746.
Grossman, S., Vila, J. (1992). Optimal dynamic trading with leverage constraints. *Journal of Financial and Quantitative Analysis* 27, 151–168.
Gordy, M. (2002). Saddlepoint approximation of CreditRisk+. *Journal of Banking & Finance* 26, 1335–1353.
Hensen, L., Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52, 557–590.
Hodrick, R., Zhou, Z. (2001). Evaluating the specification errors of asset pricing model. *Journal of Financial Economics* 62, 327–376.
Kitamura, Y., Stutzer, M. (2002). Connections between entropic and linear projections in asset pricing estimation. *Journal of Econometrics* 107, 159–174.
Nummelin, E. (2000). On the existence and convergence of price equilibria for random economies. *Annals of Applied Probability* 10, 268–282.
Pham, H. (2003). A large deviations approach to optimal long term investment. *Finance and Stochastics* 7, 169–195.

Pukhalskii, A.A. (1991). On functional principle of large deviations. In: Sazonov, V., Shervashidze, T. (Eds.), *New Trends in Probability and Statistics*. VSP Moks'las, Moskva, pp. 198–218.

Roll, R. (1992). A mean/variance analysis of tracking errors. *Journal of Portfolio Management* 18 (4), 13–22.

Sadowsky, J. (1996). On Monte Carlo estimation of large deviations probabilities. *Annals of Applied Probability* 6, 399–422.

Schilder, M. (1966). Some asymptotic formulas for Wiener integrals. *Transactions of the American Mathematical Society* 125 (1), 63–85.

Sharpe, W.F. (1966). Mutual fund performance. *Journal of Business* 1966, 119–138.

Stein, E., Stein, J. (1991). Stock price distributions with stochastic volatility: An analytic approach. *Review of Financial Studies* 4, 727–752.

Sornette, D. (1998). Large deviations and portfolio optimization. *Physica A* 256, 251–283.

Stutzer, M. (1995). A Bayesian approach to diagnosis of asset pricing models. *Journal of Econometrics* 68, 367–397.

Stutzer, M. (2000). A portfolio performance index. *Financial Analyst Journal* May/June, 52–61.

Stutzer, M. (2003). Portfolio choice with endogenous utility: A large deviation approach. *Journal of Econometrics* 116, 365–386.

Varadhan, S.R. (1984). *Large Deviations and Applications*, second ed. SIAM, Philadelphia.

Williams, N.M. (2004). Small noise asymptotics for a stochastic growth model. *Journal of Economic Theory* 119 (2), 271–298.

Subject Index